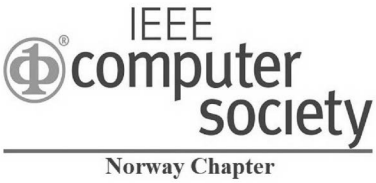
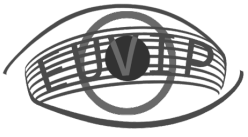


11^{th} EUROPEAN WORKSHOP ON VISUAL INFORMATION PROCESSING (EUVIP)

Gjøvik

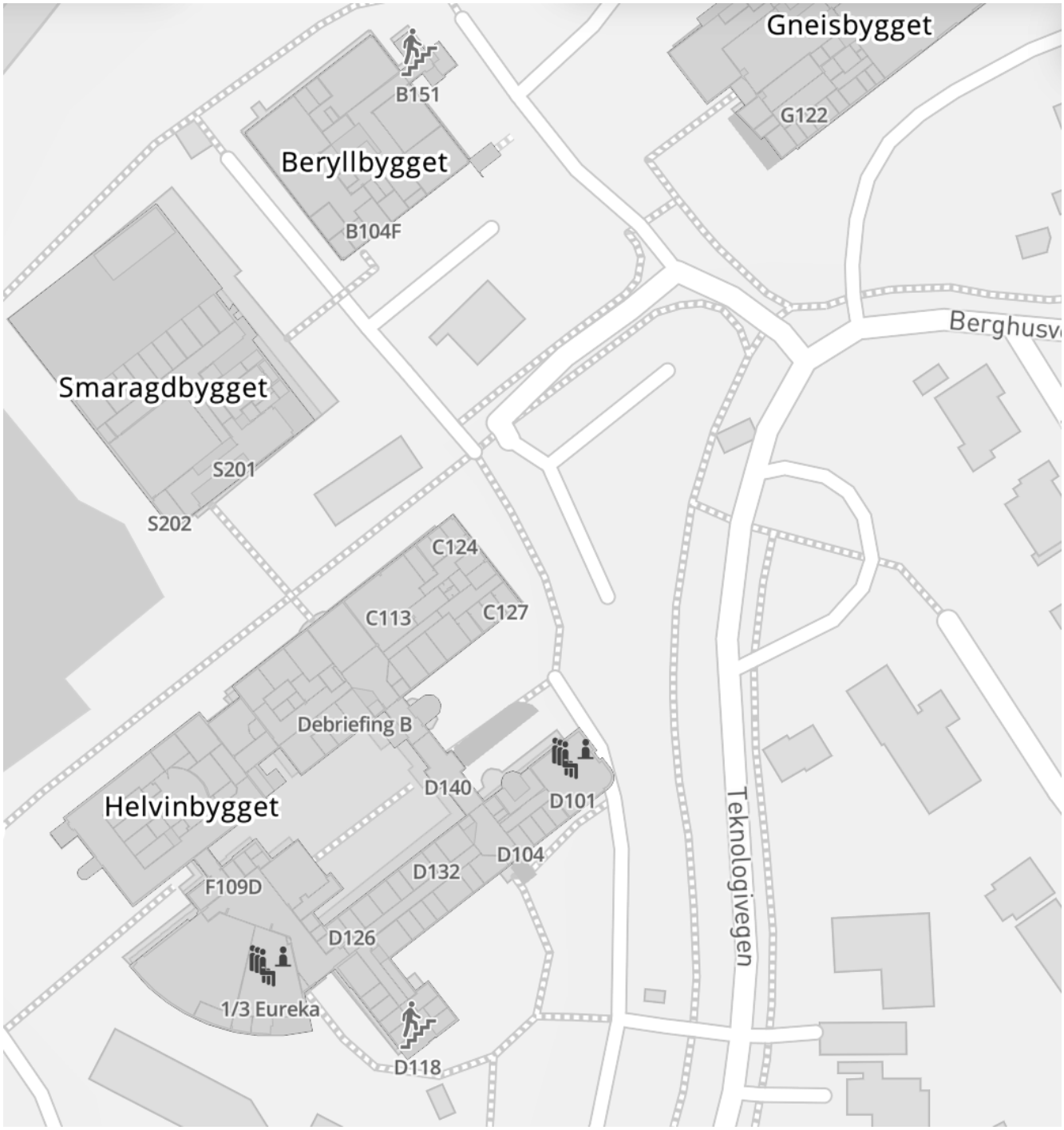
September 11 - 14, 2023



EUVIP 2023

Table of Content


K112
Koboltbygget
K101
A154
A156.2
A148
A149
A150
Ametystbygget
A121
A124.1
A117
A128
A102.3
A133.2
A114
A107
A110
G149
G161
G119
G153
G101
Gneisbygget
G103
G136
G104
G116
B151
get
Berghusvegen
10


Gneisbygget
B151
G122
Beryllbygget
B104F
Berghusv
Smaragdbygget
S201
S202
C124
C113
C127
Debriefing B
D140
D101
Helvinbygget
D104
D132
F109D
D126
1/3 Eureka
D118
Teknologivegen

Message – General Chairs EUVIP2023

Ladies and gentlemen, esteemed colleagues, and distinguished guests, Welcome to EUVIP 2023 – the 11th European Workshop on Visual Information Processing! As the General Chairs of this remarkable event, we are truly honored to have you join us for what promises to be an exceptional gathering, set in the picturesque city of Gjøvik. EUVIP has always been a nexus for the exchange of cutting-edge ideas, pioneering research, and transformative insights, and this year, our focus on "Visual Information Processing and its Applications" amplifies our commitment to advancing knowledge and innovation in this pivotal domain.

Over the next four days (11-14 September 2023), you will immerse yourselves in a diverse array of sessions, and discussions that span the breadth and depth of visual information processing. Our carefully curated program features some of the brightest minds from academia and industry who will share their expertise, discoveries, and visions that are reshaping the landscape of visual information processing.

Beyond the enlightening presentations and thought-provoking discussions, EUVIP 2023 provides a unique platform for forging connections and fostering collaborations. For the first time in EUVIP, we are thrilled to introduce a session for 3MT presentations, adding a dynamic dimension to our conference. We encourage you to make the most of this opportunity to network with peers, establish new partnerships, and pave the way for exciting collaborations that will shape the future of visual information processing and its myriad applications.

We extend our heartfelt gratitude to the leading contributors who have played pivotal roles in making this conference a success. Our appreciation goes out to the authors for their outstanding ideas, the reviewers for their constructive criticisms, the plenary and invited speakers for their generous knowledge sharing, the organizing committee for their unwavering commitment, and the technical program committee for their rigorous and fruitful efforts in shaping an engaging program.

We also express our deep appreciation to the academic institutions, research organizations, the European Association for Signal Processing (EURASIP), and the Institute of Electrical and Electronics Engineers (IEEE) for their invaluable support. Their dedication to promoting high-quality research encourages all of us to push the boundaries of knowledge further.

As we embark on this collective journey, let us remain open to the boundless possibilities that visual information processing offers, receptive to new ideas, and enthusiastic about the transformative potential of knowledge. We hope that EUVIP 2023 will be a wellspring of inspiration and motivation for all, sparking novel insights, fostering enduring connections, and propelling progress in visual information processing and its applications.

We trust that your time in Gjøvik will be delightful, affording you the opportunity to explore the region's natural beauty and immerse yourself in the rich tapestry of Norwegian culture.

Thank you for being a part of this extraordinary event, and we eagerly anticipate the rich exchanges and discoveries that await us at EUVIP 2023 – a workshop dedicated to charting new frontiers in visual information processing.

Warm regards,

Faouzi Alaya Cheikh, Stefania Colonnese, and Azeddine Beghdadi

General Chairs, EUVIP 2023

Message – Technical Chairs EUVIP2023

On behalf of the Technical Program Chairs of the 11th European Workshop on Visual Information Processing, we are glad to welcome you to EUVIP 2023. In the line of the previous successful EUVIP workshops, EUVIP 2023 focuses on visual information processing, modelling, and analysis methods inspired by the human and biological visual system. EUVIP 2023 provides a friendly and supportive environment for rich scientific exchange on perceptually-inspired image and video processing and communication methods, at a time in which Artificial Intelligence and Machine Learning methods are revolutionizing traditional approaches.

The quality of the submitted papers reflects an increasing interest in the workshop field. A total of 60 papers have been submitted to the technical program from countries from the five continents. After a double-blind review process involving two or three reviewers per paper, 33 papers have been accepted.

The technical program comprises four tutorials, as well as regular (oral and poster) technical sessions, three special sessions and four plenary talks. Similarly to previous editions, no difference is made between oral and poster presentations of regular papers.

Additionally, the workshop features a student paper award and a 3MT competition, to distinguish the best short presentations by students of research abstracts, thus providing young researchers with excellent opportunities to widen their experience.

Finally, the technical program includes a Project Dissemination Session and a Panel Discussion. We take the opportunity to thank all the technical program committee members, the reviewers, the thematic session chairs, the distinguished speakers and the authors for their invaluable contribution to the workshop success.

On behalf of the Technical Program Chairs, welcome to EUVIP2023! We hope that, besides discovering the scientific content of the workshop, you will enjoy the reception and the social program in the wonderful city of Gjøvik.

Organising committee – EUVIP2023

- **General Chairs**

  Faouzi Alaya Cheikh, NTNU, Norway
  Stefania Colonnese, Sapienza University of Rome, Italy
  Azeddine Beghdadi, USPN, France

- **Technical Program Chairs**

  Seyed Ali Amirshahi, NTNU, Norway
  Djamila Aouada, Luxembourg Univ., Luxembourg
  Nuno Rodrigues, Polytechnic of Leiria , Portugal

- **Special Sessions Chairs**

  Marius Pedersen, NTNU, Norway
  Claudio Guarnera, York University, UK

- **Tutorials Chairs**

  Giorgio Trumpy, NTNU, Norway
  Mounir Kaaniche, USPN, France
  Rahul Kumar, Oslo Univ. Hospital, Norway

- **Student Session Chairs**

  Lu Zhang, INSA, Rennes, France
  Tiziana Cattai, Sapienza University of Rome, Italy
  Mohamed Riad Yagoubi, NTNU, Norway

- **Awards Chairs**

  Ivar Farup, NTNU, Norway
  Habib Zaidi, HUGE, Switzerland
  Frederic Dufaux, Univ-Paris Saclay, France

- **Industry Liaison Chairs**

  Ahmad Iftikhar, TietoEvry, Finland
  Joseph Meehan, Huawei, France
  Mohammad Derawi, NTNU, Norway

- **Demos Exhibition Session Chairs**

  Wassim Hamidouche, TII, Abu Dhabi, UAE
  Mohib Ullah, NTNU, Norway
  Maja Krivokuca,InterDigital, Rennes, France

- **Plenary Chairs**

  Kjersti Engan, University of Stavanger, Norway
  Federica Battisti, University of Padova, Italy
  Ahmed Bouridane, University of Sharjah, UAE

- **Panel Discussion Chairs**

  Jon Yngve Hardeberg, NTNU, Norway
  Mohamed Deriche, Ajman Univ., the United Arab Emirates

- **Project Dissemination Chairs**

  Sony George, NTNU, Norway
  Rafael Palomar, Oslo Univ, Hospital / NTNU, Norway
  Joaquín Olivares, University of Cordoba, Spain

- **Publications Chairs**

  Aditya Sole, NTNU, Norway
  Steven Le Moan, NTNU, Norway
  Muhammad Ali Qureshi, IUB, Pakistan

- **Web Chairs**

  Anneli T. Østlien, NTNU, Norway
  Zuheng Ming, USPN, France
  Muhammad Muzzamil Luqman, La Rochelle University, France

- **Publicity Committee**

  Raju Shrestha, Oslo Met Univ., Norway
  Ridha Hamila, Qatar University, Qatar
  Zhaohui Wang, Hainan University, China
  Yubing Tong, University of Pennsylvania (UPenn), USA
  Madhu S. Nair, CUSAT, India
  Qiangfu Zhao, AIZU University, Japan
  Amina Serir, USTHB, Algiers, Algeria
  Mohammed El Hassouni, Mohammed V Univ., Rabat, Morocco
  Hamid Hassanpour, Shahrood University of Technology, Iran

- **Local Arrangements Committee**

  Adane Nega Tarekegn, NTNU, Norway
  Jana Blahova, NTNU, Norway

Technical program

Monday, September 11

8:30-14:00	**Registration Desk Open** [Location: Entrance of Ametyst building]
09:30-12:00	**Tutorials**
	Tutorial 1: Hard and soft metrology challenges in Material Appearance by Davit Gigilashvili and Adity Sole [Location: Ametyst building, room A154]
	Tutorial 2: Recent trends in MRI reconstruction by Joseph Suresh Paul [Location: Kobolt building, room K109]
12:00-13:00	**Lunch**
13:30-16:30	**Tutorials**
	Tutorial 3: Color in computer vision by Maria Vanrell [Location: Beryll building, room B211]
	Tutorial 4: From Volumetric Video to Interactive Virtual Humans by Peter Eisert [Location: Gneis building, room G303]

Tuesday, September 12

08:00 **Registration Desk Open**
[Location: Outside Eureka auditorium]

08:45 **Welcome Ceremony**
[Location: Helvin building, auditorium 1/3 Eureka]

09:00 **Plenary 1: *3D Computer Vision for Future Robots***
by Mohammed Bennamoun
Chair: Faouzi Alaya Cheikh

10:00 **Coffee Break**

10:15 **Special session 1: AI in the City: Efficient, Scalable and Privacy-Preserving Visual Scene Understanding in Man-Made Environments**
Chair: Amine Bourki
[Location: Helvin building, auditorium 1/3 Eureka]

10:20 **Invited talk**: Registration for Urban Modeling Based on Linear and Planar Features
by Pascal Monasse.

10:50 **Mono6D++: Learning Point Cloud Visibility for 3D Prior-based Vehicle 6D Pose Estimation**
by Yangxintong Lyu[1], Olivier Ducastel[1], Remco Royen[1], Adrian Munteanu[1].
[1]Vrije Universiteit Brussel.

11:10 **Attention-based Network for Image/Video Salient Object Detection**
by Omar Elharrouss[1], Soukaina ElIdrissi ElKaitouni[2], Younes Akbari[1], Somaya Al Maadeed[1], Ahmed Bouridane[3].
[1]Qatar University, [2]SIdi Mohamed Ben Abdellah Univerisity, [3]University of Sharjah.

11:30 **360-GAN: Cycle-Consistent GAN for Extrapolating 360-Degree Field-of-View**
by Jit Chatterjee[1], Maria Torres Vega[1].
[1]KU Leuven.

11:50 **All Predictions Matter: An Online Video Prediction Approach**
by Melan Vijayaratnam[1], Marco Cagnazzo[1], Giuseppe Valenzise[2], Enzo Tartaglione[3].
textsuperscript1LTCI, Télécom ParisTech, Institut Polytechnique de Paris, France, [2]CNRS, [3]Télécom Paris – Institut Polytechnique de Paris

12:10 **Lunch**

Oral Session 1: Quality and Performance Assessment
Chair: Ali Amirshahi
[Location: Helvin building, auditorium 1/3 Eureka]

13:40 **Evaluating the Vulnerability of Deep Learning-based Image Quality Assessment Methods to Adversarial Attacks**
by Hanene Fatima Zohra Brachemi Meftah[1], Sid Ahmed Fezza[2], Wassim Hamidouche[1], Olivier Deforges[3].
[1]INSA Rennes, [2]National Higher School of Telecommunications and ICT, [3]IETR, Rennes.

14:00 **Blind Video Stabilization Assessment based on convolutional LSTM**
by Mohamed Riad Yagoubi[1], Seyed Ali Amirshahi[1], Steven Le Moan[1], Azeddine Beghdadi[2], Erik Rodner[3].
[1]1Norwegian University of Science and Technology, [2]L2TI, University Sorbonne Paris Nord, [3]University of Applied Sciences, Berlin.

14:20 **VSTAB-QUAD: A New Video-stablization Quality Assessment Database**
by Borhen eddine Dakkar[1], Azeddine Beghdadi[1], Faouzi Alaya-Chekh[2], Amine Bourki[3].
[1]L2TI, University Sorbonne Paris Nord, [2]Norwegian University of Science and Technology. [3]VizioSense.

14:40 **Improving Viewer Training in Visual Assessment**
by Mathias Wien[1] Vittorio Baroncini[2].
[1]RWTH Aachen University, [2]VABtech.

15:00 **Poster Session: Detection, Classification, Data Protection and Scene Analysis**
Chair: Hantao Liu
[Location: Outside Eureka auditorium]
(Coffee will be available during poster session)

- **MAiVAR-T: Multimodal Audio-image and Video Action Recognizer using Transformers**, by Muhammad Bilal B Shaikh[1], Douglas Chai[1], Syed Islam[1], Naveed Akhtar[2]. [1]Edith Cowan University, [2]The University of Western Australia.
- **Semi-Supervised Anomaly Detection in Electronic-Exam Proctoring Based on Skeleton Similarity** by Habibollah Agh Atabay[1], Hamid Hassanpour[1]. [1]Shahrood University of Technology.
- **CD-COCO: A Versatile Complex Distorted COCO Database for Scene-Context-Aware Computer Vision** by Ayman Beghdadi[1], Azeddine Beghdadi[2], Malik Mallem[1], Faouzi Alaya-Chekh[3], Beji Lotfi[4] [1]Paris Saclay University, [2]L2TI, University Sorbonne Paris Nord, [3]Norwegian University of Science and Technology, [4]University of Evry.
- **Skeleton-based Hand Gesture Recognition using Geometric Features and Spatio-Temporal Deep Learning Approach** by Abu Saleh Musa Miah[1], Jungpil Shin[1], Md. Al. Mehedi Hasan[2], Yusuke Fujimoto1, Nobuyoshi Asai[1]. [1]The University of Aizu, [2]Rajshahi University of Engineering & Technology.
- **CTL-NET: Deep Learning Network for Cattle Teat Length Trait Analysis** by Hina Afridi[1], Mohib Ullah[1], Øyvind Nordbø[2], Anne Guro Larsgard[3], Faouzi Alaya-Chekh[1]. [1]Norwegian University of Science and Technology, [2]Norsvin SA, [3]Geno SA.

15:00 **Poster Session: Detection, Classification, Data Protection and Scene Analysis**
Chair: Hantao Liu
[Location: Outside Eureka auditorium]
(Coffee will be available during poster session)

- **Underwater Object Detection in AUVs using Image Enhancement and Deep Learning Models** by Adane N. Tarekegn[1], Faouzi Alaya-Chekh[1], Mohib Ullah[1], Erik Sollesnes[2], Cornelia Alexandru[3], Saeed Azar[4], Erdeniz Erol[5], George Suciu[3], [1]Norwegian University of Science and Technology, [2]USEA Ocean Data, [3]BEIA consult International, [4]OBSS Teknoloji A.Ş, [5]Elkon.
- **Wild Animal Species Classification from Camera Traps using Metadata Analysis** by Aslak Tøn[1], Ali Shariq Imran[1], Mohib Ullah[1]. [1]Norwegian University of Science and Technology.
- **A hitchhiker's guide to white-box neural network watermarking robustness** by Carl De Sousa Trias[1], Mihai Petru Mitrea[1], Enzo Tartaglione[1], Attilio Fiandrotti[2], Marco Cagnazzo[1], Sumanta Chaudhuri[1]. [1]Télécom ParisTech, Institut Polytechnique de Paris, France, [2]Università di Torino.
- **Classical approaches and new deep learning trends to assist in accurately and efficiently diagnosing ear disease from otoscopic images** by Dhruv C Jobanputra[1], Mohammed Bennamoun[1], Farid Boussaid[1], Lian Xu[1], Jafri Kuthubutheen[1].[1]The University of Western Australia.

16:00 **End of the day**

16:30 **Reception and Social Activity**
Location: Gjøvik Science Centre, Address: Brennerigata 1, 2815 Gjøvik.

Wednesday, September 13

09:00 **Plenary 2: *Quantitative imaging biomarkers in the era of precision medicine***
by Habib Zaidi
Chair: Djamila Aouada
[Location: Helvin building, auditorium 1/3 Eureka]

10:00 **Coffee Break**

10:15 **Oral Session 2 : Medical Image Processing, Analysis and Diagnosis**
Chair: Marius Pedersen
[Location: Helvin building, auditorium 1/3 Eureka]

10:20 **DCAN: DenseNet with Channel Attention Network for Super-resolution of Wireless Capsule Endoscopy**
by Hiren Vaghela[1], Anjali Sarvaiya[1], Pranav Premlani[1], Abhishek Agarwal[1], Kishor Upla[1], Kiran Raja[2], Marius Pedersen[2].
[1]Sardar Vallabhbhai National Institute of Technology, Surat, India, [2]Norwegian University of Science and Technology.

10:40 **Enhancement of Color Reproduction for Capsule Endoscopy Images**
by Léo Watine1, Pål Anders Floor[2], Marius Pedersen[1], Peter Nussbaum[1], Bilal Ahmad[1], Øistein Hovde[3].
[1]Université de Strasbourg, [2]Norwegian University of Science and Technology, [3]University of Oslo.

11:00 **A Quality-Oriented Database for Video Capsule Endoscopy**
by Tan-Sy Nguyen[1], Marie Luong[1], John Chaussard[1], Azeddine Beghdadi[1], Hatem Zaag[1], Thuong Le-Tien[2].
[1]Université Sorbonne Paris Nord, [2]HCMUT.

11:20 **FEES-IS: Real-time Instance Segmentation of Flexible Endoscopic Evaluation of Swallowing**
by Weihao Weng1, Xin Zhu[1], Mitsuyoshi Imaizumi[2], Shigeyuki Murono[2].
[1]University of Aizu, [2]Fukushima Medical University.

11:40 **Identification of Children with ADHD from EEG Signals Based on Entropy Measures and Support Vector Machine**
by Med Maniruzzaman[1], Med. Al Mehedi Hasan[2], Taro Suzuki[1], Jungpil Shin[1].
[1]The University of Aizu, [2]Rajshahi University of Engineering & Technology.

12:00 **Lunch**

Oral Session 3 : Visual information display, rendering and compression
Chair: Nuno Rodrigues
[Location: Helvin building, auditorium 1/3 Eureka]

13:40 **Evaluating the effect of sparse convolutions on point cloud compression**
by Davi Lazzarotto[1], Touradj Ebrahimi[1].
[1]École Polytechnique Fédérale de Lausanne.

14:00 **Study on Viewpoint-Dependent Time-Multiplexing for Weighted Optimization of 3D Layered Displays**
by Armand Losfeld[1], Daniele Bonatto[1], Gauthier Lafruit[1], Mehrdad Teratani[1].
[1]Université Libre de Bruxelles.

14:20 **Self-Supervised Super-Resolution Approach for Isotropic Reconstruction of 3D Electron Microscopy Images from Anisotropic Acquisition**
by Mohammad Khateri[1], Morteza Ghahremani[1], Alejandra Sierra[1], Jussi Tohka[1].
[1]University of Eastern Finland.

14:40 **Using Deep Generative Models for Glossy Appearance Synthesis and Exploration**
by Abhinav Reddy Nimma[1], Davit Gigilashvili[1].
[1]Norwegian University of Science and Technology.

15:00 **ConvNeXt-ChARM: ConvNeXt-based Transform for Efficient Neural Image Compression**
by Ahmed Ghorbel[1], Wassim Hamidouche[2], Luce Morin[1].
[1]INSA Rennes, [2]TII.

15:20 **Coffee Break**

15:40 **Plenary 3: Model-Based Optimization Meets Deep Learning in Image Analysis**
by Aleksandra Pizurica.
Chair: Stefania Colonnese.

16:40 **3MT Session**
Chairs: Tiziana Cattai

15:20 **End of the Day**

19:00 **Conference dinner**
Location: Samfundet, Address: Øvre Torvgate 24, 2815 Gjøvik

Thursday, September 14

09:00 **Plenary 4:** ***Plenary 4: Deep learning for inverse problems in imaging***
by Aleksandra Pizurica
Chair: Kiran Raja
[Location: Helvin building, auditorium 1/3 Eureka]

10:00 **Coffee Break**

Special Session 2: Image Quality Assessment and Enhancement in the Context of Medical Imaging and Diagnosis
Chair: Azeddine Beghdadi
[Location: Helvin building, auditorium 1/3 Eureka]

10:20 **Invited talk**: Image quality assessment for magnetic resonance imaging: Is it up to the task?
by Mohamed Seghier.

10:50 **Deep Learning Models for Low Dose CT Simulation**
by Lumi XIA[1], Meriem MO OUTTAS[1], Lu Zhang[1], Eric Frampas[2], Olivier Deforges[1].
[1]IETR, INSA Rennes, [2]Universitary Hospital.

11:10 **Medical Point Clouds Enhancement at the Network Edge**
by Paolo Giannitrapani[1], Tiziana Cattai[1], Stefania Colonnese[1].
[1]Sapienza Universit di Roma, Italy.

11:30 **Enhanced residue prediction for Lossless coding of multimodal image pairs based on image to image translation**
by Daniel S Nicolau[1], João Oliveira Parracho[2], Lucas Thomaz[2], Luis MN Tavora[1], Sergio M Faria[2].
[1]Instituto Politécnico de Leiria, [2]Instituto de Telecomunicacoes.

11:50 **Automatic lung nodule classification in CT images using Two-stage CNNs and Soft-voting of Multi-scale Classifiers**
by Lipeng Xie[1], Yubing Tong[2], Yuan Wan[3].
[1]Zhengzhou University, [2]University of Pennsylvania, [3]Binghamton University.

12:10 **Lunch**

Special Session 3: Spectral Imaging and its Applications
Chair: Sony George
[Location: Helvin building, auditorium 1/3 Eureka]

13:40 **Invited talk**: Spectral Imaging and its Application to Cultural Heritage
by Giorgio Trumpy.

14:10 **Bayesian Multispectral Videos Super Resolution**
by Hamid Fsian[1], Jean Baptiste Thomas[1], Pierre Gouton[2], Jon Yngve Hardeberg[3].
[1]University of Burgundy, France, [2]University of Burgundy, Franche-Comté, [3]Norwegian University of Science and Technology.

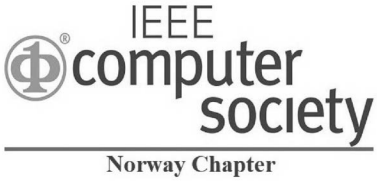
14:30 **Centralized Sample Expansion and Prior Correlation Evaluation for Hyperspectral Image Classification with Fully Convolutional Network**
by Ningyang Li[1], Zhaohui Wang[1].
[1]Hainan University.

14:50 **Coffee Break**

15:10 **Project dissemination session**
Chair: Joaquin Olivares.

16:10 **Panel discussion – Revolutionizing health care with AI-assisted medical imaging analysis**
Chair: Jon Yngve Hardeberg.

17:10 **Closing ceremony**



Proceedings Papers

Registration for Urban Modeling Based on Linear and Planar Features

Pascal Monasse
LIGM
École des Ponts, Univ. Gustave Eiffel, CNRS
Marne-la-Vallée, France
firstname.lastname@enpc.fr

Rahima Djahel
INRIA
Université Côte d'Azur
Sophia Antipolis, France
firstname.lastname@inria.fr

Bruno Vallet
LASTIG
Univ. Gustave Eiffel, ENSG, IGN
Saint-Mandé, France
firstname.lastname@ign.fr

***Abstract*—The production of a Building Information Model (BIM) from an existing asset is currently expensive and needs automation of the registration of the different acquisition data, including the registration of indoor and outdoor data. This kind of registration is considered a challenging problem, especially when both data sets are acquired separately and use different types of sensors. Besides, comparing a BIM to as-built data is an important factor to perform building progress monitoring and quality control. To carry out this comparison, the data sets must be registered. In order to solve both registration problems, we introduce two efficient algorithms. The first offers a potential solution for indoor/outdoor registration based on heterogeneous features (openings and planes). The second is based on linear features and proposes a potential solution for LiDAR data/BIM model registration. The common point between the approaches consists in the definition of a global robust distance between two segment sets and the minimization of this distance based on the RANSAC paradigm, finding the rigid geometric transformation that is the most consistent with all the information in the data sets.**

***Index Terms*—Planar region, 3D segments, Openings, RANSAC, BIM, Clustering, Registration, Global robust distance.**

I. Introduction

The indoor and outdoor modeling of buildings from images and dense point clouds is an important issue in building life cycle management. The objective is to achieve a complete, geometrically accurate, semantically annotated but nonetheless lean 3D CAD representation of buildings and objects they contain in the form of a Building Information Model (BIM). BIM helps to manage buildings in all their life cycle (renovation, simulation, deconstruction). The first challenge is to accommodate heterogeneous data as full building modeling calls for data acquisition inside and outside the building. BIM production is currently very expensive and needs automation of the registration of the different types of acquisition data. The indoor/outdoor registration is considered as a challenging problem for this kind of production, especially when both data sets are acquired separately and use different types of sensors. Comparing a BIM to as-built data of the same building is also necessary to perform building progress monitoring and quality control. To carry out this comparison, both data sets must be in the same coordinate system and a registration step is necessary.

A. State of the art

1) Indoor/outdoor registration: The registration of indoor and outdoor scans is a challenging problem for building modeling. The lack of overlap between indoor and outdoor data is the most prominent obstacle, especially when both data sets are acquired separately and use different types of sensors. Though indoor/outdoor registration is a very difficult problem, there have been several attempts to solve it. State-of-the-art approaches have used two types of features separately or together: geometric and semantic features. The key points are special points that hold important information about the global structure of the point cloud. A key point integration with the ICP algorithm (Iterative Closest Point) has been proposed in [6] to register the point clouds. When overlap between indoor and outdoor scans is low, additional information provided by the data can help the registration algorithm. The authors of [17] have extracted line segments of windows to automatically align indoor and outdoor models. To register scans with a small overlap in arbitrary initial poses, the authors of [5] have proposed a plane/line-based descriptor dedicated to establishing structure-level correspondences between point clouds. A planar polygon detection and matching method has been introduced in [1] to address the challenging problem of indoor/outdoor registration. The choice of planar polygons as appropriate attributes is grounded on the fact that they have a spatial extent limited to the areas where they have supporting points in the input data, so they form a good abstraction of the LiDAR scans. A semantic feature-matching method has been proposed in [7] to align an indoor and an outdoor point cloud. The basic idea is to include both the objects' semantic information and spatial distribution pattern by designing a semantic geometric descriptor (SGD). An efficient method for merging disconnected indoor and outdoor models of the same building into a single 3D model has been proposed in [16]. This method took semantic information (window information) into consideration to obtain candidate matches from which an alignment hypothesis can be computed.

2) LIDAR data/BIM model registration: Building Information Modeling (BIM) is seen as an important technology for building life cycle management. It plays a fundamental role in several stages, such as building progress monitoring and

979-8-3503-4218-5/23/31.002023$European Union$ © 2023$IEEE$

quality control. The progress monitoring task is based on the comparison between the as-built (the scan model of the building) and the BIM. To carry out this comparison, both data sets must be in the same coordinate system, hence a registration step is necessary. A patch-based co-registration with several static laser scans and BIM has been introduced in [8]. The main objective of this approach is to avoid the need for ground control points. An efficient method to solve the registration problem for scan/BIM has been proposed in [9]. The proposed method uses the corner points of the building structure and finds their congruent pairs to compute the optimum transformation. The authors of [10] have studied an automated registration method that aligns the as-built point cloud of a building to its as-planned model using its planar features. The basic idea is to measure the correspondence between the plane segments through a matching cost algorithm. This matching step leads to the determination of the transformation parameters to correctly register the as-built point cloud to its as-planned model. In order to co-register videogrammetric point clouds with BIM, the authors of [18] have introduced an improved matching algorithm to match 3d lines (from images) and 3d planes (from BIM).

B. Overview and contributions

The work carried out has confirmed that the environment and the type of data drive the choice of the registration algorithm [12]. So, the objective of this work is to explore the fundamental properties of the data and the environment in order to propose potential solutions for two challenging registration problems: indoor/outdoor registration and BIM model/LiDAR data registration. The man-made environments are rich in planar and linear features because they are mostly composed of elementary geometric primitives (planar polygons, openings, ...) delimited by 3D segments. Given this property, we have chosen to introduce new registration algorithms based on the minimization of global robust distance defined between two segment sets.

Our first methodological contribution is a new global distance between two 3D segment sets. Its first quality is that it is robust to segments present in on set but having no counterpart in the other set, yielding a nominal penalty for such a case. Moreover, it takes into account the fact that a long segment in one set may be detected as several shorter segments in the other set by using a notion of overlap between 3D segments. Our proposition for registration is to minimize this global distance using a guided RANSAC paradigm. This is implemented in our two demonstrated applications: the registration of indoor and outdoor LiDAR scans and the registration of indoor LiDAR scan to a BIM.

In Section II, we detail the robust global distance between 3D segment sets. Its usage for indoor/outdoor registration is explained in Section III and in Section IV for LiDAR data/BIM registration. Some experimental evaluation is presented in Section V and Section VI concludes the article.

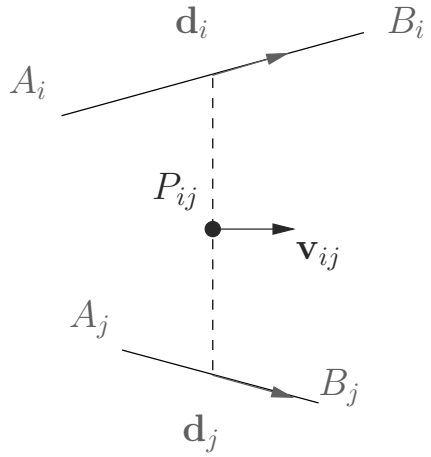


Fig. 1: Bisector line associated to two 3D segments $[A_iB_i]$ and $[A_jB_j]$. It goes through the barycenter P_{ij} of the segment endpoints and has direction $\mathbf{v}ij$.

II. Global robust distance between two segment sets

Inspired by [11], we propose to define a robust distance between two 3D line segment sets in a way that minimizing this distance will favor significant overlaps between segments and small distances over these overlaps while being robust to outliers, which is to be expected as it is possible that a segment extracted from one data set will have no counterpart in the other. We start first by defining the distance between two segments $s_i = [A_iB_i]$ and $s_j = [A_jB_j]$ by projecting orthogonally s_i and s_j on their bisector line $B(s_i, s_j)$, resulting in segments s'_i and s'_j. $B(s_i, s_j)$ is the line going through P_{ij}, the barycenter of the 4 endpoints (see Figure 1).

The direction of the bisector line is $\mathbf{v}_{ij} = (\mathbf{d}_i + \mathbf{d}_j)/2$. The projection $p_{ij}(P)$ on the bisector line of a point P is defined as

$$p_{ij}(P) = P_{ij} + ((P - P_{ij}) \cdot \mathbf{v}_{ij})\ \mathbf{v}_{ij}, \tag{1}$$

whose abscissa is $c_{ij}(P) = (P - P_{ij}) \cdot \mathbf{v}_{ij}$. The projected segment s'_k is defined by its endpoints $[p_{ij}(A_k)p_{ij}(B_k)]$ and its length is

$$|s'_k| = |c_{ij}(A_k) - c_{ij}(B_k)|, \tag{2}$$

and the overlap length is

$$\begin{aligned}|s'_i \cap s'_j| = &\min(\max(c_{ij}(A_i), c_{ij}(B_i)), \max(c_{ij}(A_j), c_{ij}(B_j))) - \\ &\max(\min(c_{(ij}A_i), c_{ij}(B_i)), \min(c_{ij}(A_j), c_{ij}(B_j))).\end{aligned} \tag{3}$$

The distance between the two segments is defined as

$$D(s_i, s_j) = \frac{1}{2}\left(D(G_i, s_j) + D(G_j, s_i)\right), \tag{4}$$

which involves the center points $G_k = (A_k + B_k)/2$, and the point-to-segment distance

$$D(G, [AB]) = \|G - p_{[AB]}(G)\|, \tag{5}$$

where the projection over the segment is

$$p_{[AB]}(G) = \begin{cases} A \text{ if } (G-A)\cdot(B-A) \leq 0 \\ B \text{ if } (G-B)\cdot(A-B) \leq 0 \\ A + \frac{(G-A)\cdot(B-A)}{\|B-A\|^2}(B-A) \text{ otherwise.} \end{cases} \quad (6)$$

This latter corresponds to a projection on the line joining A and B, then a clamping of this point to the segment $[AB]$. The distance between a segment s_1 and a set of segments S_2 is then defined as:

$$D_d(s_1, S_2) = d^2 - \sum_{s_2 \in S_2} \frac{|s_1^{'} \cap s_2^{'}|}{\min\left(|s_1^{'}|, |s_2^{'}|\right)} \max(0, d^2 - D(s_1, s_2)^2). \quad (7)$$

where parameter $d \in \mathbb{R}^+$ is a robustness parameter (all segments above this distance are considered unmatched and contribute equally to the distance). The first factor under the sum is a relative overlap of the 3D segments and is within the interval $[0, 1]$. Aggregating over all segments of S_1, we can write our global robust distance between two sets of 3D segments:

$$D_d(S_1, S_2) = \sum_{s_i \in S_1} D_d(s_i, S_2). \quad (8)$$

Let us justify the distance (7). First, if all segments of S_2 are too far from s_i ($D(s_i, s_j) \geq d^2$), all terms under the sum vanish and a nominal cost d^2 is paid. It means that "unmatched" segments in S_2 all contribute the maximum d^2. Second, if we have a segment s_j that covers s_i ($s_i' \subset s_j'$) the relative overlap is 1 and the associated cost is $D(s_i, s_j)^2$ provided it is smaller than d^2. The parameter d represents the maximum distance for which two 3D segments are considered as partially matching and a meaningful value related to the precision of the scan can be selected.

III. Indoor/outdoor registration

Openings are the most obvious common entity to link the inside and outside data. As such, they can help the registration of indoor and outdoor point clouds, so they must be automatically, accurately, and efficiently extracted. Therefore, in order to improve indoor/outdoor registration, we integrate the openings into our registration framework. The selection of opening correspondences is a crucial step for a successful registration because a bad choice can lead to a bad estimate of the optimal transformation. In our case, the features are two opening sets detected from indoor and outdoor scans. The openings are not characteristic enough to match them robustly independently. As an opening is defined by a rectangular shape composed of four segments, two of them horizontal and two vertical, and inspired by [3], we can write our registration problem as a minimization of the global robust distance (8) between two segment sets.

A. Feature extraction

1) Planar regions extraction: Due to its robustness to noise and outliers, Random Sample Consensus (RANSAC) has become the most popular method for LiDAR point cloud segmentation. Despite this success, it can generate false segments consisting of points from several nearly co-planar surfaces. Inspired by [1] we have exploited two methods depending on the nature of the data to overcome the RANSAC limitations.

a) RANSAC Based on Sensor Topology: For the outdoor scans, acquired by a Mobile Mapping System (MMS), we have access to the sensor topology (adjacency between successive pulses in the same line and between lines). Following [2], we exploit this property to extract compact planar patches.

b) MSAC: For the indoor scans, acquired in a static mode, we do not have access to the sensor topology. So, we could not use the RANSAC based on the sensor topology method for the extraction of planar regions. We follow [1] and use a straightforward adaptation of M-estimator Sample Consensus (MSAC) [13], a RANSAC extension that provides a potential solution to the spurious plane problem.

2) Openings detection: Inspired by [3], we have performed the detection of the openings in three main steps as shown in figures 2.

a) Segmentation and facade plane selection: Most openings have a rectangular vertical shape of a limited extent (a few meters) positioned within a vertical plane (a wall or the facade). So, to efficiently extract them, we need to detect the indoor and the outdoor planes and select the facades. The plane detection step has been done using the methods described in section III-A1 followed by polygons extraction using alpha shape technique [15]. Finally, the vertical large polygons have been selected as the facades.

b) Evidence of openings detection: As the LiDAR beams usually cross the facade through openings, we have started by detecting the evidence of openings as the intersection points of these beams and the detected facades using Ray Tracing:

1) For each point P_i, we trace a ray R_i from the LiDAR optical center O to this point.
2) we find P_i^j, the intersection points of R_i with the supporting planes $\mathcal{P}_j$ of each façade polygon F_j.
3) If one P_i^j lies inside the polygon and the distance between P_i and $\mathcal{P}_j > d_{min}$, we add P_i^j to the list E_j of evidences of openings on wall j. we use the threshold d_{min}, in order to exclude noisy points that may still represent points of the facade.

c) Outline openings extraction: For each wall plane $\mathcal{P}_j$ the evidences of openings E_j are grouped in vertical rectangles:

- Extract the connected components of E_j with a distance threshold.
- Estimate the minimum bounding rectangle of each component.
- For each rectangle, transform each 2D corner to a 3D point and create the four 3D line segments corresponding to its edges to get a 3D representation of our shape as shown in Figures (3.

The extraction of the connected components happens in a graph whose vertices are the points indicating an evidence of

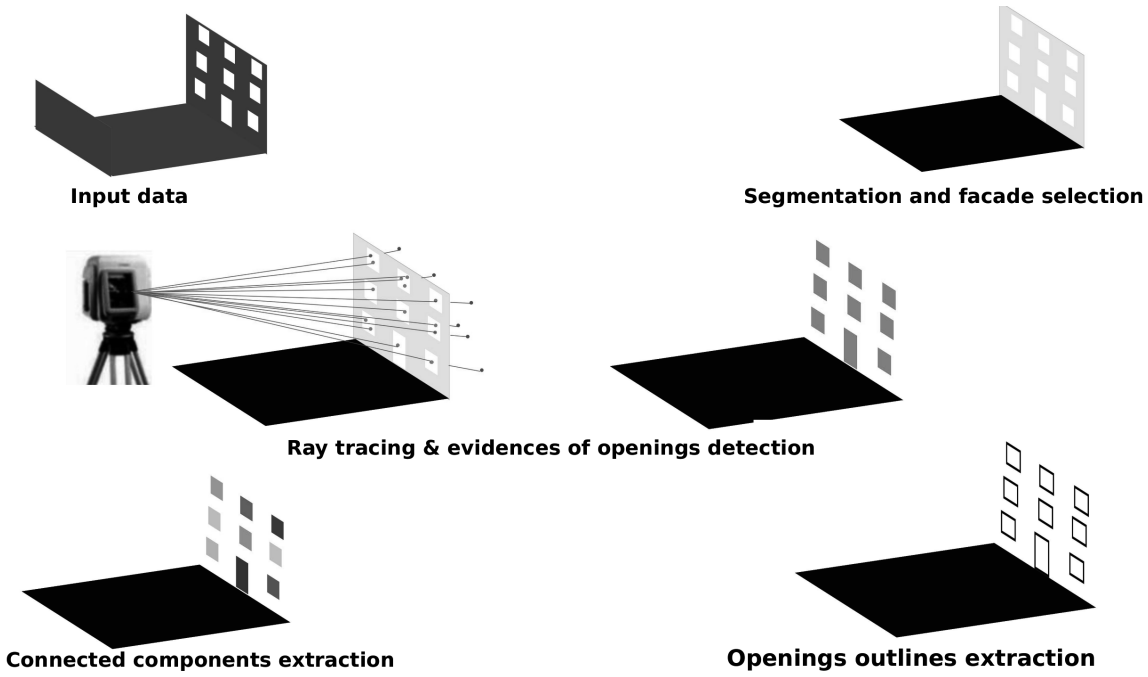


Fig. 2: Illustration of opening detection steps.

Fig. 3: Feature extraction in LiDAR scans. Left: planar regions detected in an indoor scan. Right: Openings detected in an outdoor scan.

openings and the edges link two points whenever their distance does not exceed a fixed threshold.

B. Feature matching and transformation estimation

1) Direction clustering: For each indoor scan, we greedily cluster planes P_i according to their normals $\vec{n}_i$ ordered by decreasing number of inliers n_i (number of points that have a distance to the plane less than a given threshold), using the algorithm proposed in [1], which produces for each scan three clusters: C^h, C^{v_1} and C^{v_2}.

2) RANSAC minimization: We proceed as in [2] and apply a RANSAC procedure based on two nested sampling strategies. The first one consists in matching the vertical planes representing the walls (facades) of the two scans, whereas the second one consists in matching the selected segments in the matched walls. The advantage of integrating these two matching strategies is to simplify the selection of the two pairs of segments and decrease the calculation time of the algorithm by reducing the number of iterations. Each RANSAC iteration consists in the following steps:

- Select a facade plane from each scan, p_1 for indoor plane and p_2 for outdoor plane.
- Make the two planes coincide by applying a rotation R on p_1 around their line of intersection, the result is p'_1
- Randomly selects a corner c_1 (the intersection point of

a horizontal segment and a vertical segment) from an opening of $p_1^{'}$ and a corner c_2 from an opening of p_2.
- Compute the translation component in the plane between the two selected corners, which fixes only two degrees of freedom.
- Randomly select two planes p_3 and p_4, each one from an indoor cluster perpendicular to the facade, and transform them.
- Compute A: the intersection point of $p_1^{'}$, p_3 and p_4; B: the intersection point of p_2 , p_3 and p_4.
- Find the missing degree of freedom of the translation as the vector joining A and B.

IV. LiDAR data/BIM model registration

A. 3D line segment detection

In this work, we have chosen the algorithm proposed in [4] to extract 3D line segments from LiDAR data. For the BIM model, we use a Poisson disk sampling method [14] to sample points over the whole model and extract the edges from the sampled point cloud with the same method. It is a simple and efficient algorithm that starts by segmenting the point cloud into planar 3D regions via region growing and merging. All the points belonging to the same planar region are projected into the supporting plane of this region to form a 2D image. Then, 2D contour extraction and least square fitting are performed to detect 2D line segments. Finally, these 2D line segments are transformed back into the 3D frame to get the 3D segments.

B. Feature matching and transformation estimation

1) 3D segments directional clustering: The first step of our pipeline is the clustering of the 3D segments of each dataset according to their direction. This is done using our proposed greedy algorithm described in Algorithm 1. For each input data, we cluster 3D segments L_i according to their direction d_i in decreasing order of length, which produces for each dataset three clusters: C_1^1, C_1^2 and C_1^3 for the first input data, C_2^1, C_2^2 and C_2^3 for the second input data. Notice that at step 5, for computing the average direction of a cluster, care must be taken to be invariant to the possibly opposite orientations of the different segments. That is why the sign function is used, so as to orient each direction as much along $\mathbf{v}_1$ as possible.

2) Direction cluster association: We associate each cluster C_1^i with the cluster C_2^j with the smallest angle between the mean direction:

$$A_k = \{C_1^i, C_2^j\}, |\mathbf{d}(C_1^i) \cdot \mathbf{d}(C_2^j)| \geq 1 - \epsilon \tag{9}$$

C. RANSAC minimization

RANSAC has proven its robustness and efficiency as an optimization algorithm in several applications. In this section, we describe a new version of RANSAC based on the selection of double pairs of segments at each iteration. The selected segment pairs define a unique transform. The clusters are used to ensure that these pairs of segments have compatible angles. At each RANSAC iteration:

- We randomly select two cluster associations.

Algorithm 1 Greedy direction clustering

1: Input: Set of segments L, each segment $L_i = [A_iB_i]$ has a director vector $\mathbf{v}_i = \overrightarrow{A_iB_i}$, a length $\|\mathbf{v}_i\|$ and a unit direction $\mathbf{d}_i = \mathbf{v}_i/\|\mathbf{v}_i\|$.
2: Clusters initialization:
- $C_1 = \{L_1\}$ where L_1 is the longest segment.
- $C_2 = \{L_2\}$ where L_2 is the longest one among segments for which $|d_i \cdot d_1| < cos(\epsilon)$.
- $C_3 = \{L_3\}$ where L_3 is the longest one among segments for which $\max(|d_i \cdot d_1|, |d_i.d_2|) < cos(\epsilon)$.

3: Mark L_1, L_2 and L_3 as processed and all other L_i as unprocessed
4: Let L_{cur} be the longest unprocessed segment. If there is no unprocessed segment, stop the algorithm, else mark L_{cur} as processed.
5: Each cluster C_k has a direction $\mathbf{d}(C_k)$ computed as mean of the directions of the 3D segments in the cluster:

$$\mathbf{d}(C) = \frac{\sum_{L_i \in C} \text{sign}(\mathbf{v}_i \cdot \mathbf{v}_1)\, \mathbf{v}_i}{\| \sum_{L_i \in C} sign(\mathbf{v}_i \cdot \mathbf{v}_1)\, \mathbf{v}_i \|}$$

6: Compute $k_{min} = \arg\min_k 1 - |d_{cur} \cdot d(C_k)|$
7: Compute $n_i = 1 - |d_{cur} \cdot d(C_{k_{min}})|$
8: If $n_i < \epsilon$, add L_{cur} to the cluster $C_{k_{min}}$.
9: Go back to step 4.

- We randomly select one segment from each associated cluster.
- We compute the transform (rotation and translation) that best aligns the matched 3D segments using the method of Section IV-D.
- For this transform, we estimate the global robust distance between all segments of the two sets using (8).

The final registration is given by the transformation that minimizes the global robust distance.

D. Transform estimation

Once two pairs of segments $\{v_i, v_j\}$ (first selected cluster association) and $\{h_i, h_j\}$ (second selected cluster association) are associated, we estimate the rotation that best aligns the corresponding two 3D lines. Let us call d_v^i, d_v^j, d_h^i and d_h^j the unit director vectors of v_i, v_j, h_i and h_j. We start by creating the orthonormal basis $\mathcal{O}^i = (\mathbf{x}^i, \mathbf{y}^i, \mathbf{z}^i)$, where:

$$\mathbf{x}^i = \mathbf{d_v^i} \qquad \mathbf{y}^i = \frac{\mathbf{d_h^i} - (\mathbf{d_h^i} \cdot \mathbf{x}^i)\mathbf{x}^i}{\|\mathbf{d_h^i} - (\mathbf{d_h^i} \cdot \mathbf{x}^i)\mathbf{x}^i\|} \qquad \mathbf{z}^i = \mathbf{x}^i \times \mathbf{y}^i$$

We then compute the rotation R that aligns the associated clusters as the base change matrix between $\mathcal{O}^i$ and $\mathcal{O}^j$:

$$R = \mathcal{O}^j \mathcal{O}^{i^{-1}} \tag{10}$$

To estimate the translation, we start by defining the point-to-line distance:

$$dist(\mathbf{p}, L = \mathbf{a} + \mathbf{d}t) = \frac{\|\mathbf{d} \wedge (\mathbf{a} - \mathbf{p})\|}{\|\mathbf{d}\|} = \|[\mathbf{d}]_\times (\mathbf{a} - \mathbf{p})\| \tag{11}$$

assuming that $\mathbf{d}$ is normalized, and again calling $[\mathbf{d}]_\times$ the matrix of the cross product with $\mathbf{d}$. We look for the translation $\mathbf{t}$ that minimizes:

$$\epsilon = \sum_i \|[\mathbf{d}]_\times(\mathbf{a}_i - (\mathbf{p}_i + \mathbf{t}))\|^2 \tag{12}$$

The minimum is reached when the gradient is null:

$$\nabla_\mathbf{t}\epsilon(\mathbf{t}) = -2\sum_i [\mathbf{d}_i]_\times^t [\mathbf{d}_i]_\times(\mathbf{a}_i - (\mathbf{p}_i + \mathbf{t})) = 0 \tag{13}$$

Noting:

$$\mathbf{w} = \sum_i [\mathbf{d}_i]_\times^t [\mathbf{d}_i]_\times(\mathbf{p}_i - \mathbf{a}_i) \quad M = -\sum_i [\mathbf{d}_i]_\times^t [\mathbf{d}_i]_\times,$$

we get a closed form solution for $\mathbf{t}$:

$$\mathbf{t} = M^{-1}\mathbf{w} \tag{14}$$

V. Evaluation and discussion

In this work, we were interested in two challenging problems for BIM production and BIM comparison with the as-built data in order to perform progress monitoring.

A. Indoor/outdoor registration

The first part of our contribution consists in proposing a heterogeneous features-based registration algorithm to address the challenging problem of indoor/outdoor registration. Our proposed method takes as attributes a set of openings and planar features. Using the openings we can find the rotation and the translation in the facade plane (two degrees of freedom). We recover the missing degree of freedom for the translation by adding the planes. The best transformation has been selected based on the minimization of the global robust distance between two segment sets. We tested our algorithm on real data, and the obtained results have proved the performance of our algorithm to register the indoor and outdoor scans whatever the initial position as shown in figure 4. As we do not know the ground truth we only considered the visual results. The introduced approach has exceeded the limitations of some existing methods:

- Iterative methods require a good approximation of the initial transformation to be able to converge toward the correct solution.
- Opening-based methods such as the method proposed in [3] have an uncertainty in the direction orthogonal to the facade.

B. LiDAR data/BIM model registration

The second part of our contribution consists in proposing a linear features-based registration algorithm to address the problem of LiDAR data/BIM model registration. The optimal transformation was estimated by minimizing the global robust distance between two sets of 3D segments after extracting them using a state-of-the-art algorithm. We have tested our algorithm on real data corresponding to a construction site in Spain. The obtained results have proved the performance of our method to register the LiDAR data and the model BIM as shown in figure 5.

VI. Conclusion and future work

In this paper, we are interested in two registration problems that remain challenging problems for urban modeling. The first is the indoor/outdoor registration which represents a very important step for BIM production. In order to carry out this kind of registration we have proposed an efficient algorithm based on heterogeneous features(openings and planes). The second is the LiDAR data/BIM model registration witch considered a key step in performing progress monitoring and quality control. The common point of the two proposed solutions consists in the definition of the global robust distance between two segment sets and the minimization of this distance based on the RANSAC paradigm. We can propose some improvements in future works:

- LIDAR data/BIM model registration: extraction of sharp edges on the BIM model and use them as input 3D segments of this model for the registration
- Indoor/outdoor registration: if we can see pieces of the indoor walls parallel to the facade during the outdoor scan, we can detect them as points where the LiDAR beams cross the facade. The extraction of the planes from these points and their integration into our registration framework can give us additional information on the thickness of the facade which can increase the accuracy of the registration.

Acknowledgments

The authors would like to thank:

- L'Institut national de l'information géographique et forestière (https://www.ign.fr/) for providing access to the outdoor data (Mobile LiDAR Scans).
- INSA Strasbourg (https://www.insa-strasbourg.fr/fr/) for providing access to the indoor data (Static LiDAR Scans).
- The partners of EU Horizon 2020 BIM2TWIN: Optimal Construction Management and Production Control project under Agreement No. 958398 (https://bim2twin.eu/) for providing access to the BIM model and the associated LiDAR data.
- The anonymous reviewers of this paper, especially one who provided many missing references of related work.

References

[1] Djahel, R., Vallet, B., and Monasse, P. (2021). Towards Efficient Indoor/outdoor Registration Using Planar Polygons. ISPRS annals of the photogrammetry, remote sensing and spatial information sciences, 2, 51-58.

[2] Guinard, S. A., Mallé, Z., Ennafii, O., Monasse, P., and Vallet, B. (2020). Planar polygons detection in lidar scans based on sensor topology enhanced RANSAC. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2, 343-350.

[3] Djahel, R., Vallet, B., and Monasse, P. (2022). Detecting openings for indoor/outdoor registration. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 177-184.

[4] Lu, X., Liu, Y., and Li, K. (2019). Fast 3D line segment detection from unorganized point cloud. arXiv preprint arXiv:1901.02532.

[5] Chen, S., Nan, L., Xia, R., Zhao, J., qnd Wonka, P. (2019). PLADE: A plane-based descriptor for point cloud registration with small overlap. IEEE Transactions on Geoscience and Remote Sensing, 58(4), 2530-2540.

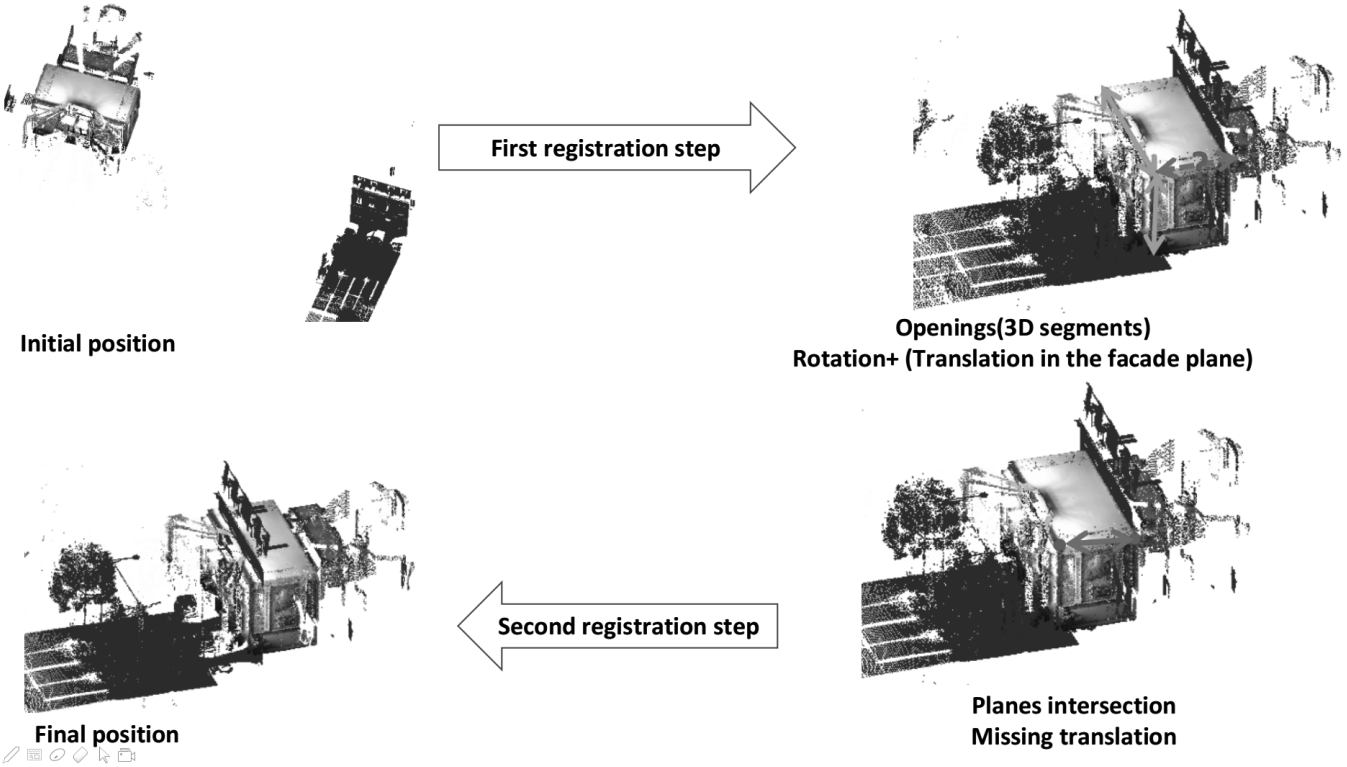


Fig. 4: Indoor/Outdoor registration between two LiDAR scans. Upper-left: the RGB-colored indoor scan is unregistered with the uncolored exterior scan. Upper-right: the RANSAC algorithm recovers the translation in the facade plane (green axes) but leaves the facade breadth undecided (magenta axis). Lower-right: Identifying the two corners from the scans allows registering the scans (lower-left) assuming no facade breadth.

[6] Assi, R., Landes, T., Murtiyoso, A., qnd Grussenmeyer, P. (2019). Assessment of a key points detector for the registration of indoor and outdoor heritage point clouds. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 133-138.

[7] Yang, Y., Fang, G., Miao, Z., and Xie, Y. (2022). Indoor–Outdoor Point Cloud Alignment Using Semantic–Geometric Descriptor. Remote Sensing, 14(20), 5119.

[8] Gruner, F., Romanschek, E., Wujanz, D., and Clemen, C. (2022). Co-Registration of Tls Point Clouds with Scan-Patches and Bim-Faces. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 46, 109-114.

[9] Sheik, N. A., Deruyter, G., and Velaerts, P. (2022). Automated Registration of Building Scan with BIM through Detection of Congruent Corner Points. In The 7th International Conference on Smart City Applications (Vol. 48, pp. 179-185). Copernicus GmbH.

[10] Sheik, N. A., Deruyter, G., and Veelaert, P. (2022). Plane-based robust registration of a building scan with its BIM. Remote Sensing, 14(9), 1979.

[11] Djahel, R., Monasse, P., and Vallet, B. (2022). A 3D segments based algorithm for heterogeneous data registration. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 129-136.

[12] Djahel, R. (2022). Registration of Heterogenous Data for Urban Modeling (Doctoral dissertation, Marne-la-vallée, ENPC).

[13] Torr, P. H., Zisserman, A., 2000. MLESAC: A new robust estimator with application to estimating image geometry. Computer vision and image understanding, 78(1), 138–156.

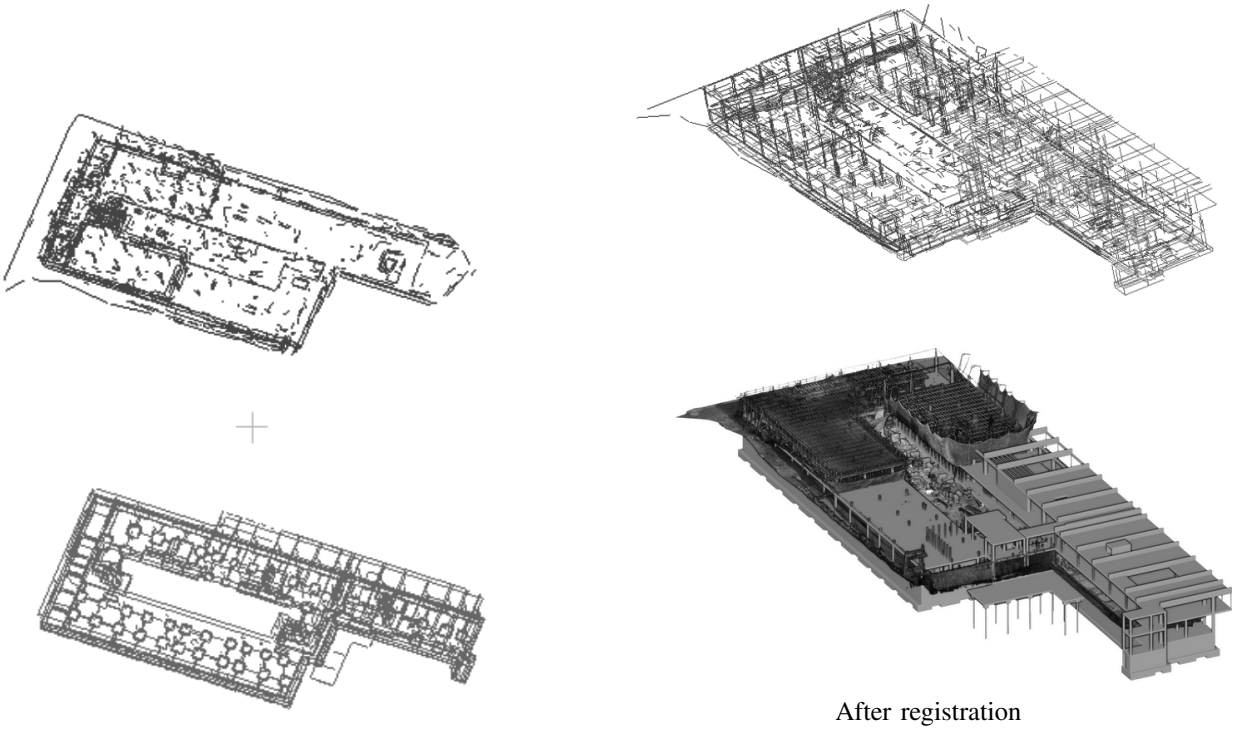
[14] Gamito, M. N., and Maddock, S. C. (2009). Accurate multidimensional Poisson-disk sampling. ACM Transactions on Graphics (TOG), 29(1), 1-19.

[15] Edelsbrunner, H., Kirkpatrick, D., and Seidel, R. (1983). On the shape of a set of points in the plane. IEEE Transactions on information theory, 29(4), 551-559.

[16] Cohen, A., Schönberger, J. L., Speciale, P., Sattler, T., Frahm, J. M., and Pollefeys, M. (2016). Indoor-outdoor 3d reconstruction alignment. In ECCV 2016: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, October 11-14, 2016, Part III 14 (pp. 285-300). Springer International Publishing.

[17] Koch, T., Korner, M., and Fraundorfer, F. (2016). Automatic alignment of indoor and outdoor building models using 3D line segments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 10-18).

[18] Kaiser, T., Clemen, C., and Block-Berlitz, M. (2022). Co-registration of video-grammetric point clouds with BIM–first conceptual results. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 46, 141-148.



Extracted segments before registration

After registration

Fig. 5: Lidar data/BIM model registration. Left: the LiDAR scan (red) and the point cloud obtained by sampling the BIM. Right: registered segments and registered LiDAR and BIM.

Mono6D++: Learning Point Cloud Visibility for 3D Prior-based Vehicle 6D Pose Estimation

Yangxintong Lyu, Olivier Ducastel, Remco Royen, and Adrian Munteanu
Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium
IMEC, Kapeldreef 75, 3001 Leuven, Belgium
Email: {yangxintong.lyu, olivier.ducastel, remco.royen, adrian.munteanu}@vub.be

***Abstract*—Point cloud visibility is a crucial attribute for 3D tasks as it links the visible object points to a given viewpoint. In this paper, we address the problem of point cloud visibility for monocular vehicle 6D pose estimation. To this end, a network, dubbed Mono6D++, is introduced which jointly predicts vehicle poses and the associated points visibility. Our method mainly consists of: 1) a multi-model feature extraction module and 2) a fusion unit for learning the pose- and visibility-specific representations. Consequently, the proposed method significantly outperforms the baseline approaches. Mono6D++ is capable of handling heavily occluded, truncated and/or appearance-ambiguous vehicles.**

***Index Terms*—Monocular vehicle pose estimation, $SE(3)$, multi-modal data processing, intelligent traffic systems, deep learning.**

I. Introduction

Recently, the estimation of a vehicle 6D pose from a single RGB camera has seen a surge of interest in both academic and industrial communities due to its low cost and applicability. It enables a spatial reasoning among vehicles, which provides a potential of reliable mobility applications for autonomous driving and intelligent traffic systems, such as trajectory planning and traffic monitoring. However, directly inferring the translation and rotation from a single view image is an ill-posed problem as it lacks scale information. Thanks to the advances in deep learning, the current state-of-the-art approaches achieve promising results.

The existing monocular-based techniques retrieve the 6D pose in either a 3D-generation or a 3D-utilisation manner. The former generates the coarse geometrical representations which are used in the subsequent steps. An energy function in terms of modeling vehicle 3D keypoints is proposed in Mono3D++ [1], while 3D-RCNN [2] encodes complex CAD models by using a low-dimensional shape space. It enables the network to jointly predict the 6D pose and shape. A 'render-and-compare' loss optimises the difference of the shape projection between the ground truth and predicted pose. However, the loss term is only applicable in the pre-training phase on synthetic data as it is impractical to obtain the real-world dense traffic depth. GSNet [3] further advances 3D-RCNN by learning from 2D specific vehicle features. Nevertheless, retrieving additional, highly accurate labeled keypoints from images is expensive and vulnerable to inaccuracies.

The 3D-utilisation methods regress the poses with the assistance of the given 3D information. In particular, vehicles are classified into different categories according to the appearance or body shape in the image, which can be linked with an appropriate shape corresponding to a specific vehicle category. By doing so, the queried 3D geometry, which can come in various forms such as 3D keypoints and point clouds, complements the single-view RGB image.

DeepManta [4] utilises specific vehicle 3D features and fits them with the corresponding colorful keypoints during pose refinement phase. Since geometrically fitting a 3D model is time-consuming, DeepManta is limited to offline applications. On the other hand, Mono6D [5] suggests the usage of point clouds, randomly sampled from CAD models, as 3D prior information. It retrieves the appropriate 3D prior from a database by employing a make and model recognition method [6]–[9]. The network estimates vehicle poses by fusing the RGB and point cloud channels without subsequent processing. By design, the time-consuming pose refinement step is omitted in Mono6D. Despite of its advantages, an important limitation of Mono6D is that it does not account for the point visibility information of the 3D prior.

The method proposed in this work follows a similar multi-modal paradigm to that of Mono6D. In addition, it solves its important limitation by determining the visibility information and by assigning each point a label indicating whether or not the point is visible from a given camera viewpoint. Learning the visibility information has proven to be efficient in many 3D tasks [10], [11] as it associates the visible object component with the corresponding viewpoint. To the best of our knowledge, there exists no 3D prior-based method employing point cloud visibility in the pose estimation process.

Our contributions can be summarised as follows:

- We are the first to tackle the point cloud visibility problem for the 3D prior-based monocular vehicle 6D pose estimation task.
- We formulate the visibility prediction as a binary classification task. The proposed method is able to jointly distinguish the 6D pose and visible points by fusing the representations of the vehicle image and complete point cloud. To do so, we extend an existing public real-world traffic dataset, Apollo3DCar, with the visibility annotations of the 3D prior.
- The proposed method significantly outperforms the baseline approaches on Apollo3DCar. Specifically, it substantially reduces the translation error and improves the rota-

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

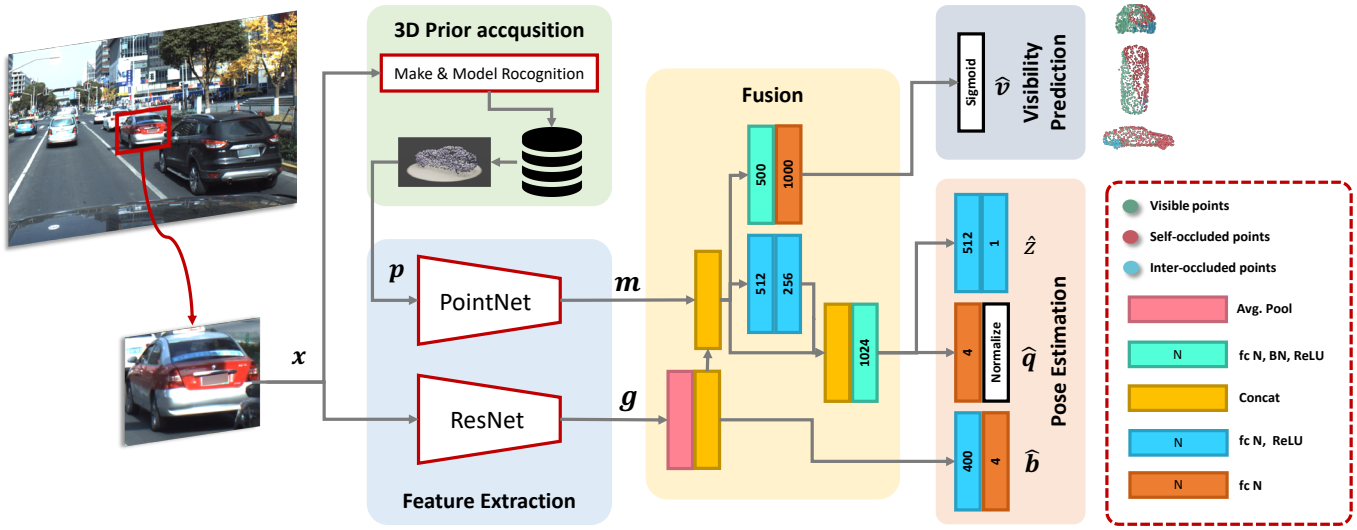


Fig. 1. The pipeline of Mono6D++.

tion accuracy. Consequently, Mono6D++ is able to handle challenging appearance-ambiguity, severe occlusion and truncation cases.

The reminder of this paper is organised as follows: Section II introduces the proposed Mono6D++. Section III presents the experimental results and the analysis of our approach. Section IV draws the conclusion of this work.

II. Methodology

A. Notation

In our 3D prior-based approach, the aim is to predict a detected vehicle pose in $SE(3)$ and the visibility information of the corresponding 3D prior from a single-view image I. We denote each sample $\boldsymbol{s}_i$, as below:

$$\boldsymbol{s}_i = (\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\varphi}_i, \boldsymbol{p}_i, \boldsymbol{v}_i), \tag{1}$$

where the dataset $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^{N}$ and N is the number of data samples. We represent each sample $\boldsymbol{s}_i$ by the amodal bounding box $\boldsymbol{b}_i \in \mathbb{R}^4$, which consists of the height H_i, the width W_i and the pixel coordinate (u_i, v_i) in the image plane by projecting the vehicle 3D centroid; the cropped RoI $\boldsymbol{x}_i \in \mathbb{R}^{H_i \times W_i \times 3}$ extracted from the entire image I; the pose $\boldsymbol{\varphi}_i$; the 3D prior $\boldsymbol{p}_i \in \mathbb{R}^{n \times 3}$ containing n points; and the visibility $\boldsymbol{v}_i \in \mathbb{R}^n$ of each point of the 3D prior. Each 3D prior $\boldsymbol{p}_i$ is a point cloud sampled from the corresponding vehicle CAD model.

Similar to previous 3D prior-based methods, our method predicts $\boldsymbol{b}_i$ in the normalization form [12], and denotes $\boldsymbol{\varphi}_i$ by $[z_i, \boldsymbol{q}_i]$, where z is the depth and $\boldsymbol{q}_i$ is the rotation in the form of quaternion. Given the camera intrinsic parameters,

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

the translation $[x, y, z]$ of a vehicle can be computed as following:

$$x = \frac{z(u - c_x)}{f_x}, y = \frac{z(v - c_y)}{f_y}, \tag{3}$$

once $[u, v]$ and z are known.

In the following part of this section, we present our visibility annotation method for vehicles in monocular images during the data generation phase. Then, we describe the proposed 3D prior-based method, dubbed Mono6D++, which jointly predicts vehicle 6D pose and the corresponding pointcloud visibility.

B. Point cloud visibility computation

To compute the point cloud visibility $\boldsymbol{v}_i$, we assume that a detected vehicle in an RGB image I is labeled with the associated bounding box $\boldsymbol{b}_i$, 6D pose $\boldsymbol{\varphi}_i$, 3D prior information $\boldsymbol{p}_i$ and shape model. Let $\tilde{\boldsymbol{p}}_i$ denote the point cloud which is transformed by the pose $\boldsymbol{\varphi}_i$.

Each vehicle point $P_i = [x_i, y_i, z_i] \in \tilde{\boldsymbol{p}}_i$ in the traffic scene can be categorized as either visible or invisible. We define a point as invisible if it satisfies any of the following conditions: 1) it is occluded by the vehicle itself, 2) it is occluded by other objects, or 3) it is truncated by the image. Otherwise, we consider the point as visible. Note that regarding condition 2, 'other objects' solely refers to other vehicles since the dataset lacks point cloud information for objects other than vehicles.

Inspired by the ray casting algorithm [13], we construct a virtual line segment OP_i between the camera center point O and P_i whose visibility must be determined. If OP_i intersects the object mesh at the point Q_i and $P_i \neq Q_i$, P_i is invisible. Otherwise, if $P_i = Q_i$, P_i is visible.

Among the points classified as invisible, a part of them are self-occluded when the corresponding intersection point is located on the vehicle itself. The remaining invisible points are

occluded by other vehicles present in the scene. On the other hand, within the subset of points classified as visible after ray casting, there exists a subset labeled as invisible due to their projection on the image plane being truncated

Consequentially, $\boldsymbol{v}_i$ is a vector consisting of zeros and ones which represent invisible and visible points, respectively.

C. Mono6D++ approach

As shown in Figure 1, given a detected vehicle in the image, Mono6D++ firstly queries the corresponding 3D prior from a vehicle database. Then, the features of the RGB and point cloud channels are learnt by a feature extraction module. Afterwards, a fusion unit is designed to fuse the multi-modal representations. Subsequently, the visible information of the prior and the pose are predicted by a visibility block and a pose estimation module, respectively. We detail each of these modules in the following.

3D prior acquisition. As stated in [5], a proper 3D geometry can be retrieved from a vehicle prior database in accordance with the make and model (M&M) information. Using the existing recognition techniques [6]–[9], M&M can be extracted from the RGB image. By doing so, the retrieved 3D data can be utilised in the subsequent modules as prior information and complementary modality data for the monocular 6D pose estimation task. In the proposed method, we follow this 3D prior query paradigm.

Feature extraction module. For the task of 6D pose estimation, it has been established by [14], [15] that using separate branches is an effective approach to extract efficient representations for multi-modal data. Therefore, we employ two distinct branches dedicated to the image and point cloud modalities. As shown in Figure 1, a ResNet-based backbone [16] denoted as $\mathcal{E}$ is designed to extract multi-resolution image embeddings, while a PointNet-like backbone [17], $\mathcal{M}$, learns geometry information from the 3D channel. Let $\boldsymbol{g}_i = \mathcal{E}(\boldsymbol{x}_i)$ and $\boldsymbol{m}_i = \mathcal{M}(\boldsymbol{p}_i)$.

Fusion for pointcloud visibility and pose estimation. The fusion module follows the design shown in Figure 1 such that the multi-modal features $\boldsymbol{g}_i$ and $\boldsymbol{m}_i$ can be fused for visibility- and pose-specific representative features. A novelty of our method is that we employ the visibility of each point of the queried 3D prior. To do so, we formulate the prediction as a binary classification task where the label of each point belongs to $\{visible, invisible\}$. To optimize the predictive visibility denoted as $\hat{\boldsymbol{v}}_i$, we employ the Binary Cross Entropy (BCE) loss [19], which is defined as follows:

$$l_i^{vis} = -(\boldsymbol{v}_i \log(\hat{\boldsymbol{v}}_i) + (1 - \boldsymbol{v}_i) \log(1 - \hat{\boldsymbol{v}}_i)). \tag{4}$$

The predicted rotation, bounding box, depth and prior visibility are denoted as $\hat{\boldsymbol{q}}_i$, $\hat{\boldsymbol{b}}_i$, $\hat{z}_i$ and $\hat{\boldsymbol{v}}_i$. As [3], [5], smoothL1 loss [20] and L1 loss are used to optimise the predictions of the 2D bounding box and 6D pose, respectively. Let us denote the loss functions following:

$$l_i^{bbox} = smoothL1(\hat{\boldsymbol{b}}_i - \boldsymbol{b}_i), \tag{5}$$

$$l_i^{rot} = |\frac{\hat{\boldsymbol{q}}_i}{\|\hat{\boldsymbol{q}}_i\|} - \boldsymbol{q}_i|, \tag{6}$$

$$l_i^{z} = |\hat{z}_i - z_i|, \tag{7}$$

Thus, we represent the final visibility loss $\mathcal{L}_{vis}$, bounding box loss $\mathcal{L}_{bbox}$, rotation loss $\mathcal{L}_{rot}$ and depth loss $\mathcal{L}_z$ respectively as:

$$\mathcal{L}_\theta = \frac{1}{N} \sum_{i=1}^{N} l_i^\theta, \tag{8}$$

where $\theta \in \{vis, bbox, rot, z\}$.

Joint loss optimization. The final loss function $\mathcal{L}_{total}$ is defined by:

$$\mathcal{L}_{total} = \lambda_{vis}\mathcal{L}_{vis} + \lambda_{rot}\mathcal{L}_{rot} + \lambda_z\mathcal{L}_z + \lambda_{bbox}\mathcal{L}_{bbox}. \tag{9}$$

Both the visibility and vehicle pose are optimised by minimising $\mathcal{L}_{total}$ when we set $\lambda_\theta = 1$, for all $\theta \in \{vis, rot, z, bbox\}$.

III. Experiments

A. Network details

We implement the proposed method based on Mono6D [5]. ResNet18 [16] and PointNet [17] are used as feature extractors for the RGB and pointcloud channels, respectively.

B. Dataset

We perform the experiments on the real-world dataset, Apollo3DCar [18], which is commonly used in the 6D pose estimation literature. It consists of 4036/200 color images for training/validation, respectively. All the vehicles are coarsely classified into 79 categories, each associated with the corresponding 3D CAD models. We use the same 3D priors as Mono6D [5] which are randomly sampled from each vehicle CAD model. To augment the images, we apply random cropping, keep-ratio resizing, and adjustments of brightness and hue saturation.

C. Training schema

We implement Mono6D++ in Pytorch [21]. The network is trained by an Adam optimizer [22] with initial Learning Rate (LR) of 10^{-4}. The LR decays following a cosine annealing policy [23] in which the LR is reduced to 5×10^{-5} after 150 epochs. We set 64 as batch size during training phase. All the experiments are tested on a machine equipped with a 2080 NVIDIA GTX GPU.

D. Evaluation metrics

Following [3], [5], [18], both the absolute and relative versions of the instance 3D average precision (A3DP-Abs, A3DP-Rel) are reported in the paper. We present the results under the loose and strict criteria [18], denoted as *c-l* and *c-s* respectively. Moreover, we measure the Average Relative Euclidean Distance (ARED) of the translation, as well as the accuracy with threshold δ and the median error in degrees (*Mederr*) of the rotation. We compute the accuracy of the point cloud visibility for the ith vehicle instance as follows:

$$acc_i = \frac{TP_i + TN_i}{N_i}, \tag{10}$$

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS.

Method	A3DP-Rel			A3DP-Abs			Method category
	mean ↑	c-l ↑	c-s ↑	mean ↑	c-l ↑	c-s ↑	
DeepMANTA [4]	16.04	23.76	19.8	20.10	30.69	23.76	3D-utilisation
Kpts-based [18]	16.53	24.75	19.8	20.40	31.68	24.75	3D-utilisation
3D-RCNN [2]	10.79	17.82	11.88	16.44	29.70	19.80	3D-generation
Direct-based [18]	11.49	17.82	11.88	15.15	28.71	17.82	3D-generation
GSNet [3]	20.21	40.50	19.85	18.91	37.42	18.36	3D-generation
GSNet [3]($IoU > 50\%$)	25.51	49.08	23.16	19.85	43.89	17.09	3D-generation
Mono6D [5] ($IoU > 50\%$)	30.54	59.76	27.25	18.63	40.43	17.78	3D-utilisation
Proposed method ($IoU > 50\%$)	**38.67**	**61.77**	**41.80**	**27.43**	**48.98**	**29.13**	3D-utilisation

TABLE II
DETAILED COMPARISON WITH GSNET [3] AND MONO6D [5].

Method	R		T
	$acc(\frac{\pi}{6}) \uparrow$	$Mederr \downarrow$	$ARED \downarrow$
GSNet	95.90%	3.17	5.20%
Mono6D	96.14%	2.66	4.28%
Proposed method	**97.18%**	**2.49**	**3.09%**

where TP and TN denote the number of True Positive and True Negative visibility classifications and N the number of points of each 3D prior.

E. Results and discussions

In Table I, we compare the proposed method with the state-of-the-art techniques. The results of the baseline methods are extracted from Mono6D [5]. As A3DP jointly estimates 6D pose and vehicle shape, we ensure a fair comparison by following the setup described in Mono6D. More specifically, we use the ground truth shapes as the estimated shapes for GSNet ($IoU > 50\%$), Mono6D ($IoU > 50\%$) and Ours ($IoU > 50\%$). One can note that by predicting the visibility information, Mono6D++ significantly outperforms the baseline methods providing the best absolute and relative A3DPs. As shown in Table II, compared to Mono6D, Mono6D++ achieves a 1.19% reduction in translation error and more than 1% improvement in rotation accuracy.

Furthermore, based on the depth of the vehicles, we divided the validation set into 5 subsets. As shown in Figure 2, both Mono6D and Mono6D++ exhibit lower translation errors when the distance between the vehicle and the reference camera decreases. The distant vehicles appear more blurry and indistinguishable in the image, which leads to inefficient RGB features and reduced accuracy. Additionally, though the accuracy of the predictive visibility steadily drops, as indicated in Figure 3, the proposed method is able to substantially alleviate the ARED errors of Mono6D in each group in Figure 2. Moreover, our method achieves similar performance as Mono6D but for the vehicles belonging to the further group. One notes that the median ARED error predicted by

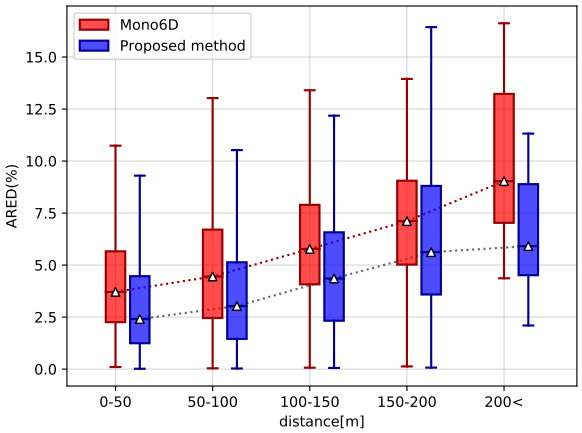


Fig. 2. The ARED of Mono6D and proposed method in function of distance. *Triangular* symbol represents the median ARED of each box.

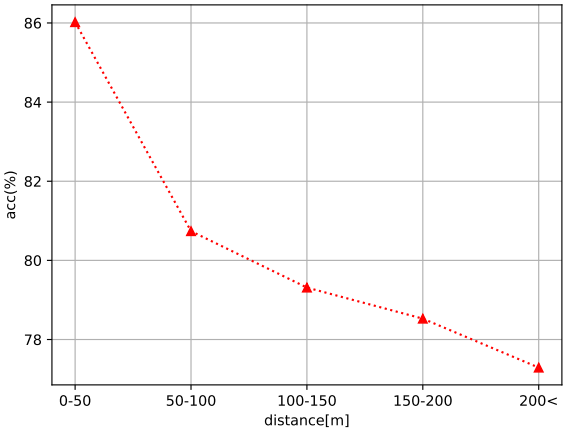


Fig. 3. The average accuracy of the predicted visibility in function of distance.

Mono6D++ for the '200 <' group is similar to that of the '150 – 200' group of Mono6D.

The results in Figure 4 show that Mono6D++ can es-

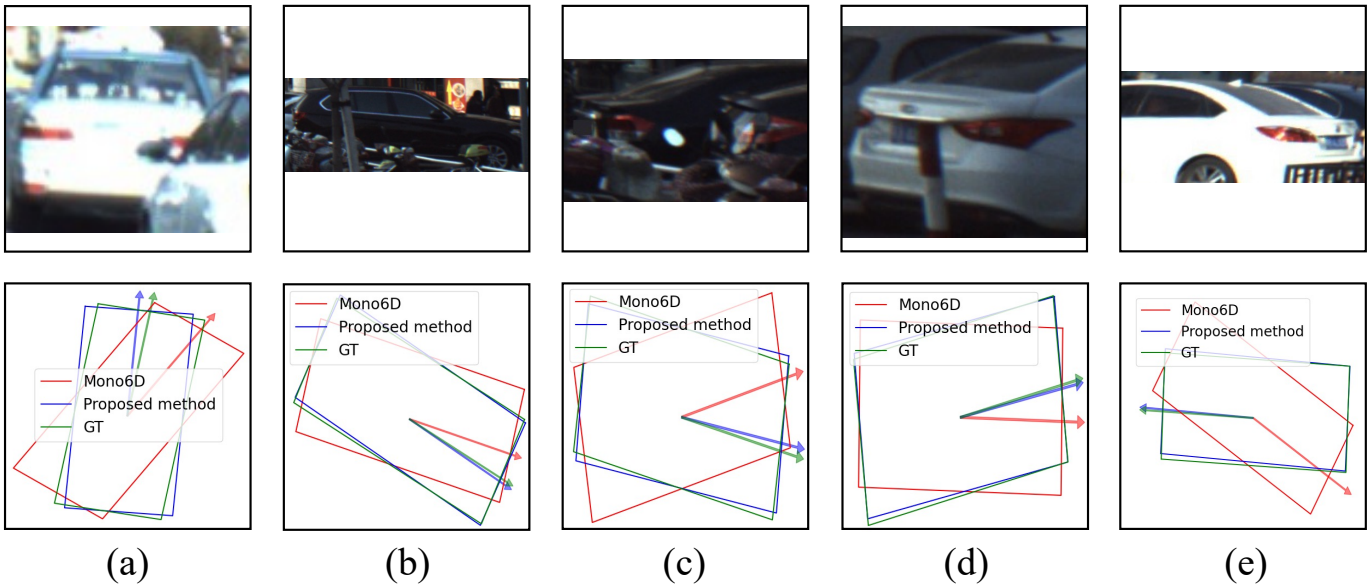


Fig. 4. The bird's eye view of the detected challenging vehicles. (a)-(b) Severe occlusion. (c)-(d) Truncation. (e) Vehicle appearance ambiguity. The *arrow* denotes the orientation of vehicle head.

timate highly accurate rotations for the challenging cases when Mono6D fails: 1) vehicles are severely occluded and/or truncated, leading to a limited visible component in the image (Figure 4 (a)-(d)), and 2) the frontal appearance of the vehicle is ambiguous with the back in special viewpoints, which introduces a confusion in determining the head orientation (Figure 4 (e)). Learning the visibility assists the network in identifying the visible vehicle 3D structure in the camera viewpoint. By doing so, Mono6D++ can extract more efficient multi-modal pose-specific representations than Mono6D as the latter uses complete vehicle point clouds.

IV. Conclusion

In this work, we tackle the point cloud visibility problem for monocular 3D prior-based vehicle 6D pose estimation. Our proposed method, Mono6D++, predicts a vehicle pose together with the visibility information of the shape prior. The experimental results demonstrate that Mono6D++ achieves more accurate rotation and translation with lower error compared to the baseline methods. Furthermore, our approach is able to handle severe occlusions, truncation and the challenging vehicle front-back ambiguity.

V. Acknowledgement

The authors would like to thank for the financial support provided by Innoviris (TORRES, SPECTRE) and by the Fonds Wetenschappelijk Onderzoek (FWO) - 1S89420N.

References

[1] Tong He and Stefano Soatto, "Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8409–8416.

[2] Abhijit Kundu, Yin Li, and James M Rehg, "3d-rcnn: Instance-level 3d object reconstruction via render-and-compare," in *CVPR*, 2018, pp. 3559–3568.

[3] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang, "Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision," in *ECCV*. Springer, 2020, pp. 515–532.

[4] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teuliere, and Thierry Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in *CVPR*, 2017, pp. 2040–2049.

[5] Yangxintong Lyu, Remco Royen, and Adrian Munteanu, "Mono6d: Monocular vehicle 6d pose estimation with 3d priors," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2187–2191.

[6] Adeel Ahmad Jamil, Fawad Hussain, Muhammad Haroon Yousaf, Ammar Mohsin Butt, and Sergio A Velastin, "Vehicle make and model recognition using bag of expressions," *Sensors*, vol. 20, no. 4, pp. 1033, 2020.

[7] Lei Lu, Ping Wang, and Hua Huang, "A large-scale frontal vehicle image dataset for fine-grained vehicle categorization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1818–1828, 2022.

[8] Bipul Neupane, Teerayut Horanont, and Jagannath Aryal, "Real-time vehicle classification and tracking using a transfer learning-improved deep learning network," *Sensors*, vol. 22, no. 10, 2022.

[9] Yangxintong Lyu, Ionut Schiopu, Bruno Cornelis, and Adrian Munteanu, "Framework for vehicle make and model recognition—a new large-scale dataset and an efficient two-branch–two-stage deep learning architecture," *Sensors*, vol. 22, no. 21, pp. 8439, 2022.

[10] Pierre Biasutti, Aurélie Bugeau, Jean-François Aujol, and Mathieu Brédif, "Visibility estimation in point clouds with variable density.," in *VISIGRAPP (4: VISAPP)*, 2019, pp. 27–35.

[11] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner, "Unsupervised point cloud pre-training via occlusion completion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9782–9792.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2016.

[13] Scott D Roth, "Ray casting for modeling solids," *Computer graphics and image processing*, vol. 18, no. 2, pp. 109–144, 1982.

[14] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *CVPR*, 2019, pp. 3343–3352.

[15] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet, "Pose from shape: Deep pose estimation for arbitrary 3D objects," in *BMVC*, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[17] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 652–660.
[18] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang, "Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving," in *CVPR*, 2019, pp. 5452–5462.
[19] Ma Yi-de, Liu Qing, and Qian Zhi-Bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 743–746.
[20] "Smoothl1loss," https://pytorch.org/docs/stable/generated/torch.nn.SmoothL1Loss.html, Accessed: 2021-11-01.
[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019.
[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015.
[23] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

Attention-based Network for Image/Video Salient Object Detection

1st Omar Elharrouss
Department of Computer Science
Qatar University
Doha, Qatar
elharrouss.omar@gmail.com

2nd Soukaina ElIdrissi ElKaitouni
Department of Computer Science
Sidi Mohamed Ben Abdellah Univerisity
Fez, Morocco
elidrissi1soukaina@gmail.com

3rd Younes Akbari
Department of Computer Science
Qatar University
Doha, Qatar
akbari_younes@yahoo.com

4th Somaya Al-Maadeed
Department of Computer Science
Qatar University
Doha, Qatar
s_alali@qu.edu.qa

5th Ahmed Bouridane
Centre for Data Analytics and Cybersecurity.
University of Sharjah
Sharjah, UAE
abouridane@sharjah.ac.ae

***Abstract*—The goal of video or image salient object detection is to identify the most important object in the scene, which can be helpful in many computer vision-based tasks. As the human vision framework has a successful capacity to effortlessly perceive locales of interest from complex scenes, salient object detection mimics a similar concept. However, the salient object detection (SOD) of complex video scenes is a challenging task. This paper mainly focuses on learning from channel and Spatio-temporal representations for image/video salient object detection. The proposed method consists of three levels, the frontend, the attention models, and the backend. While the frontend consists of VGG backbone which ultimately learns the representation of the common and the discrimination features. After that, both Attention, Channel-wise, and Spatiotemporal models are applied to highlight the significant object using a feature detector and to calculate the spatial attention. Then the output features are fused to obtain the final saliency result. Experimental investigation evaluations confirm that our proposed model has proved its validity and effectiveness compared with the state-of-the-art methods.**

***Index Terms*—Image Salient Object Detection, Video Salient Object Detection, Image segmentation, semantic segmentation.**

I. Introduction

Salient object detection is the detection of the most important object in an image/video. The human vision system has an effective ability to easily recognize regions of interest from complex scenes, even if the focused regions have similar colors or shapes as the background. But automatic detection can be difficult due to scale variation as well as the complexity of the scene. Therefore, salient object detection from images/videos plays an important role in many computer vision applications such as motion detection [1], semantic and instance segmentation [2], [3], object detection [3], [8], and many others. Many methods have been proposed for salient object detection on image and video using different techniques such as spatiotemporal analysis [5], [8], [29], [30], 3D analysis [9], [10], or using deep neural networks [11]. For example, the authors in [12] proposed an improved salient object detection using a hybrid Convolution Recurrent Neural Network. In another work, ConvLSTM has been used. While ConvLSTM efficiently captures the dynamics of saliency by learning the shift of human attention [37]. The proposed method optimizes functionality from multiple networks, typically consisting of spatial and temporal sub-networks. While in [38] the authors proposed a network by focusing on the mode of movement and the transmission of the continuity of the object. Encoder-decoder models based on fully convolutional networks (FCNs) have remarkably improved the performance of pixel-by-pixel image-to-frame learning operations [16]. Essentially, the tendency of the main SOD methods developed in recent years indicates that most of them operate in the encoder-decoder framework. Several researchers have developed encoder-decoder-based structures for the SOD task like in [46]. While in [47] the authors proposed a connectivity-based approach called bilateral connectivity network (BiconNet), which uses connectivity masks together with saliency masks as labels for effective modeling of inter-pixel relationships and object saliency. Techniques, such as Multi-scale Interactive Network for Salient Object Detection [48], are also developed and introduced into SOD models. In this paper, we proposed a CNN-based method consisting of analyzing the channel and spatial features in the image/video for detecting silent objects. Unlike, the proposed methods, the channel features are not used widely to detect the saliency. while it can give help in learning due to the information that can be extracted from the channel of objects. For that, the fusion of spatial and channel features can improve the learning of different aspects in the image/video to extract the saliency of objects. A description of the proposed method is presented in the following sections.

The rest of the paper is organized as follows. The second part briefly introduces related research work, the third part introduces the network model proposed in this paper in detail, the fourth part shows the experimental details of our paper, and finally, a summary of this paper is presented.

979-8-3503-4218-5/23/31.00 © 2023*IEEE*

TABLE I
IMAGE/VIDEO SALIENCY OBJECT DETECTION METHODS

Method	Technique	Datasets
Tang et al. [5]	Spatiotemporal attention neural networks	FBMS, DAVIS
STA-Net [6]	Spatiotemporal attention network for VSOD	ViSal, DAVIS
Shokri et al. [11]	VSDO using deep non-local neural networks	DAVIS, FBMS
Cong et al. [13]	Sparse reconstruction and CNN detection network	DAVIS, ViSal, SegTrackV1
Tu et al. [14]	Collaborative graph learning network on RGB and thermal images	VT821
Huang et al. [15]	Super-pixel segmentation and multi-scale learning network	MSRA10k, ECSSD, Pascal-1500
Chen et al. [25]	LSTM-based network	SegTrackV2, DAVIS
Tased-net [26]	Spatial encoder-decoder network	Hollywood2, UCFSports
XU et al. [27]	Motion Energy and Graph Clustering	UVSD, DAVIS
ConvLSTM [29]	Multi-scale spatiotemporal ConvLSTM model	FBMS, DAVIS, MCL
CAGNet [32]	Fusion of feature extraction network and the feature guidance network	DUTS, DUT-OMRON, HKU-IS
Wang et al. [33]	Deep evolution of the graph GCN structure to accurately predict VSOD	DAVIS, DAVIS,SegV2, FBMS, MCL

II. RELATED WORKS

Image/ Video salient object detection (VSOD) represents an important task in several real-world domains like video segmentation [39], [40], video compression [41], [42], video captioning [43], autonomous driving [44], [45]. However, some difficulties can affect the performance of any method such as the scale variation, image resolution, and scene complexity. In the following, we present some existing works for image and video salient object detection. while some of these methods are summarized in Table I in terms of techniques and datasets used in each method.

Image saliency detection: has been widely studied during the last decades. Therefore, We will consider some influential works. For example, in [14] the authors used a collaborative graph learning network. Huang et al . [15] proposed Image saliency detection via multi-scale iterative CNN by fusing each scale of the network to generate the final results. Zhang et al. [16] proposed Gradient-induced co-saliency detection. Jiang et al. [17] proposed Robust visual saliency optimization based on bidirectional Markov chains. Wang, X. [18] proposed a Region-based depth feature descriptor for saliency detection on a light field. Jian et al. [19] proposed Visual saliency detection by integrating spatial position prior to object with background cues. While Liu et al. [28] proposed The Single Stream Recurrent Convolution Neural Network (SSRCNN) borrows the VGG-16 network as the main backbone that first globally detects salient objects, and then applies the Depth Recurrent Convolution Neural Network (DRCNN) top-down side-output sub-network that hierarchically and progressively specifies the details of salient objects from depth to shallowness. The SSRCNN with four-channel RGBD inputs and the DRCNN sub-network is trained comprehensively by deep supervised learning. In [32] the authors proposed a content-based feature guidance network (CAGNet) containing three networks: the feature extraction network that extracts contextual information at multiple scales, the feature guidance network that guides the extracted features by exploiting the spatial details of the low-level features and the semantic information of the high-level features, and the feature fusion network that efficiently integrates the guided features to generate the saliency map [23]. Also, in [34] the authors proposed a saliency boundary detection stream and a saliency detection stream. While depth information is used to generate accurate saliency boundaries at four scales. Then, RGB images are used to combine multi-level multi-scale contextual feature maps and saliency boundary feature maps through an attention module to produce the four-scale saliency prediction maps. All proposed image salient methods succeed in salient object detection in simple scenarios but they suffer from many challenges including scale variation and the complexity of the scene.

Video saliency object detection: Besides the aforementioned saliency prediction methods for image, video saliency detection methods has some influential works including the detection of an object in a video stream that can offer an easy way to track the detected object. Also, it can make semantic and instance segmentation much easier. For that, object saliency detection from video become an interesting task in computer vision. Many methods have been proposed to overcome the challenges and using different techniques. For example, in [13] the authors proposed a salient object detection method via sparsity-based reconstruction and propagation. The authors in [20] proposed a Multi-level model for video saliency detection. The method used a multi-scale representation to take benefit from different scales of the network. While many methods worked on the type of network like in 3DCNN [9], [10], Adaptive diffusion [24] or the type of features used like in deep non-local neural network [11], Motion Quality Perception [21]. Also, some methods are based on the many inputs data to implement their network [31]. For example, Wang et al. [33] proposed a deep evolution of the graph convolutional network GCN structure to accurately predict salient objects in videos, where long-term structural dependencies between frames are explored. and adaptively evolved the structure of the clustering graph, which propagates the information flow hierarchically, reduces redundant edges, and adapts the graph to moving objects. The authors in [36] based on the intrinsic characteristics of light fields is to develop a multi-task collaborative network for salient object detection in light fields by leveraging collaborative learning of multiple tasks including edge detection, depth inference, and salient object detection.

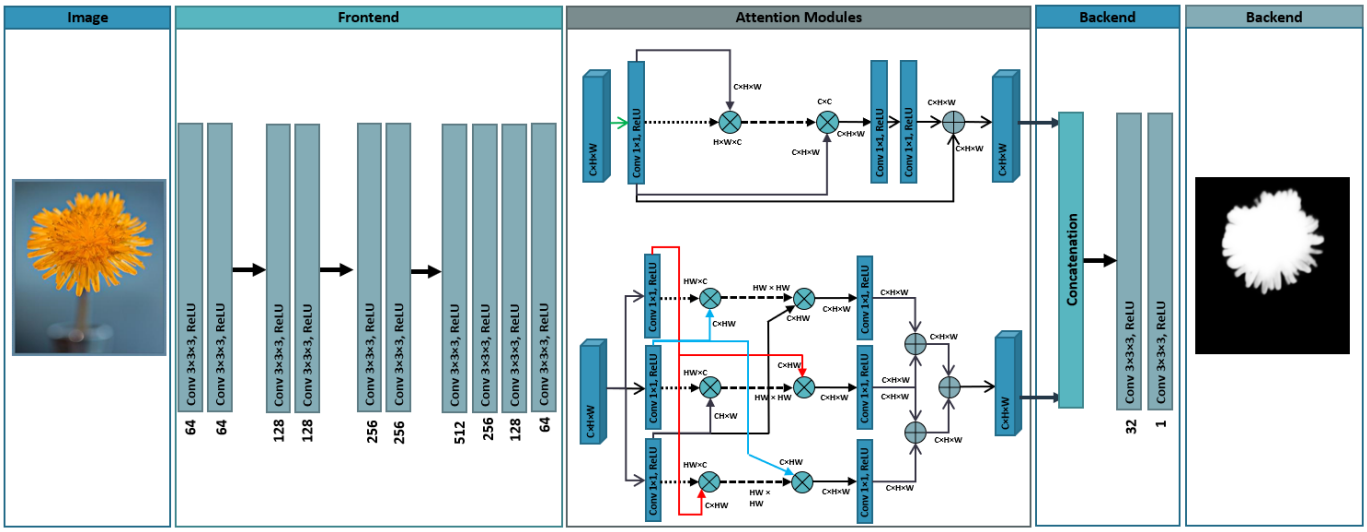


Fig. 1. Flowchart of the proposed network.

The spatiotemporal analysis is one of the key solutions for good object detection in video. many methods. While the spatial relationship between regions in a frame as well as the temporal coherence of the regions allows accurate learning of different characteristics and features. For that, the authors in [5] proposed Video salient object detection via spatiotemporal attention neural network. While in [6] the authors proposed a spatial-temporal attention network for video salient object detection. In the same context, Huang et al. [7] proposed Learning channel-wise spatiotemporal representations. Chen et al. [25] improved robust video saliency detection based on long-term spatial-temporal information. Min et al. [26] proposed a Temporally-aggregating spatial encoder-decoder network. Graph clustering with motion energy and spatiotemporal abjectness to detect the salient objects [27]. While in [29] proposed an LSTM-based network that introduces space-based and channel-based attention mechanisms and improves the network's ability to extract high-level semantic information and low-level spatial structure features. Fang et al. proposed a method for detecting salient objects in a video based on deep semantic and spatiotemporal cues, which consists of three components: the Conv2DNet for learning semantic features of objects, the Conv3DNet for learning spatiotemporal features, and the Deconv3DNet for learning sharing by merging semantic and spatiotemporal features [30]. The authors in [35] proposed a cross-attention-based encoding-decoding model in the Siamese framework (CASNet) for salient object detection in video. which consists of two parts: A cross-attention module is designed to capture the short-term spatiotemporal dependency between two adjacent video frames. In addition, a structure integrating the self-attention and cross-attention modules is integrated into a Siamese framework to preserve the spatiotemporal correlation of salience and increase the consistency of salience detection between two adjacent video frames.

III. Proposed Method

We proposed an encoder-decoder-based network for image/video salient object detection. The proposed model consists of an adapted VGG model with the introduction of a channel-wise attention module between VGG layers and then cascaded spatial-wise attention at the end of the network.

In order to implement the proposed model, we used an adapted version of the VGG model with Channel and spatial attention models. Unlike the other method, we used channel-wise attention modules to ensure the extraction of the important features at each phase of the network. Because in a complex scene, the foreground contains some regions that can be similar to the background in terms of texture. In order to differentiate between the foreground and the background, channel-wise attention is designed for that purpose. Also, we used a cascaded spatial attention module to focus on spatial information. Spatial-wise attention is designed to encode the consistent density change as well as the global and local density distribution regularity. Also to extract the contextual information and capture the change in density distribution. To enforce the learning of these features we adopted the spatial-wise attention to cascaded spatial-wise attention presented in Figure 1. The proposed architecture is light compared with the other architecture as well as easy to implement.

A. channel-wise attention

A channel-wise attention module is a channel-based attention module for fully convolutional neural networks. The purpose of the channel is to extract the important features of the input image with a feature detector that corresponds to each channel in the feature map. The spatial dimension of the input function map was compressed to measure the channel attention efficiently. As shown in Fig. 1, First, by using average pooling and max pooling operations, space information of a map was aggregated and this generated two different spatial

TABLE II
THE PERFORMANCE OF EACH METHOD ON THE EXISTING IMAGE AND VIDEO SALIENCY DETECTION DATASET. THE **BOLD** AND UNDERLINE FONTS RESPECTIVELY REPRESENT THE **FIRST** AND SECOND PLACE

Method	DAVIS		FBMS	
	F-measure	MAE	F-measure	MAE
Tang et al. [5]	0.834	0.041	0.812	0.087
STA-Net [6]	0.883	0.025	-	-
3DCNNX-shape [11]	0.815	0.050	0.823	0.085
Cong et al. [13]	0.683	0.094	-	-
ConvLSTM [29]	0.817	0.024	0.797	0.063
Conv2DNet [30]	0.830	0.029	-	-
DCB [33]	0.891	0.021	**0.873**	0.037
Ours	**0.901**	**0.020**	0.847	**0.032**

TABLE III
MAE RESULTS OF EACH METHOD ON DUT-OMTON, DUTS, AND ECSSD DATASETS. THE **BOLD** AND UNDERLINE FONTS RESPECTIVELY REPRESENT THE **FIRST** AND SECOND PLACE

Method	DUT-OMRON	DUTS	ECSSD
Zhao et al. [23]	0.0414	-	0.0405
Ours	**0.0341**	**0.0871**	**0.0391**

context descriptors: respectively that indicate average pooled features and max pooled characteristics. Both descriptors are then sent to a shared network to create our attention channel map.

B. Spatial-wise Attention

A Spatial-wise Attention Module is a spatial attention module for fully conventional neural networks. It produces a spatial care map by the use of the inter-space function relationship. Unlike the attention of the channel, the focus of spatial attention is where an information component complements the attention of the channel. We first apply the average pooling and max pooling operations along the channel axis and concatenate them to produce an effective characteristic descriptor for calculating spatial attention.

IV. EXPERIMENTAL RESULTS

In order to evaluate the obtained results using the proposed methods we used MAE and F-measure metrics. These results are compared with the existing works and presented in Table II and III which contains the results of four known datasets including DAVIS for image salient object detection and FBMS for video salient object detection.

A. Evaluation Metrics

To evaluate the performance of the proposed model we use the mean absolute error (MAE), which is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - z_i^{gt}| \quad (1)$$

Where N is the number of images used for testing, z_i^{gt} denotes the real scene, and z^i represents the obtained results. While MAE indicates the accuracy of the salient detection.

(a) DAVIS

(b) DUT-OMRON

Fig. 2. The obtained results using the proposed method on DAVIS and DUT-OMRON datasets. First row: Original image. Second row: ground truth. Third row: obtained result using the proposed method.

B. Quantitative Comparison

In order to evaluate the proposed method and compare it with the existing techniques we used MAE and F-measure as evaluation metrics. For that, II, III provide the obtained results on five datasets including image and video salient object detection datasets. From the table which represents the obtained results on image salient object detection datasets including DAVIS and FBMS, we can find that the proposed method, as well as the state-of-the-art methods, succeed to achieve good results. On the DAVIS dataset, which is the most used dataset for evaluating salient object detection from images, the proposed method achieved the best results in terms of MAE and F-measure metrics followed by DCB [33] with a difference of 1%. While the obtained results using the other

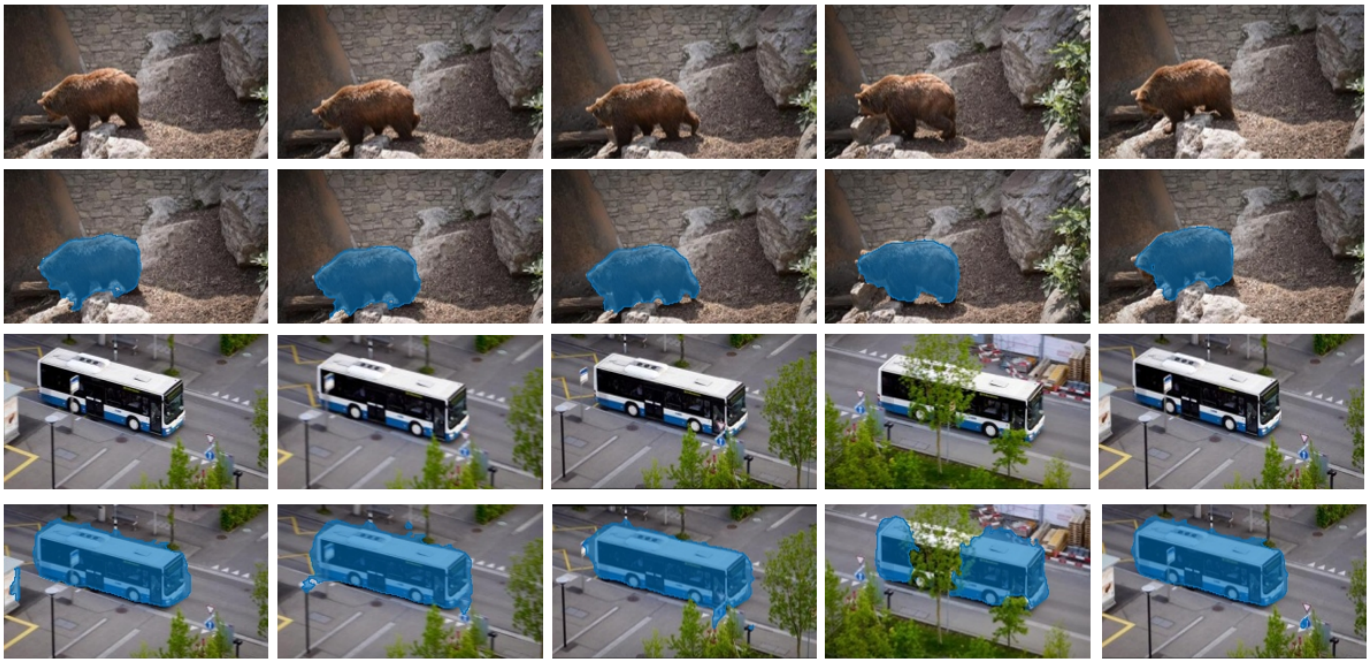


Fig. 3. video silent detection On FBMS dataset.

methods are close including Tang et al. [5], STA-Net [6], 3DCNNX-shape [11], ConvLSTM [29], and Conv2DNet [30].

For the FBMS dataset which is a video salient object detection dataset, the proposed method reached the best MAE metric value. The obtained results approve that the use of channel and spatial analysis can help in better saliency detection. Compared with the state-of-the-art methods, we can see that the proposed method and DCB [33] method provide better results in terms of F-measure and MAE metrics, with a difference of 3% for F-measure. While the other method as well obtains convincing results close to the proposed method results. Also, the reached values are better for images than the obtained results on videos. This is true regarding the presented results using MAE metric on other datasets presented in Table III. Because the proposed methods used different datasets to evaluate their methods for video saliency detection, we compare the proposed method with only one method [23].

C. Qualitative Comparison

In order to demonstrate the obtained results using MAE metric, we present the qualitative results using the visualization of some simple from each dataset. The visualization of the obtained results for some datasets is illustrated in Figures 2 and 3. From the figures, we can find that the generated images using the proposed method are of high quality and similar to the ground truth. Also, even for objects with different scales that makes the founding of a pattern difficult as well as the learning of the pattern. In addition, with the use of channel features, we obtained effective detection even while the object in a complex scene or the color of the object is similar to the background.

V. Conclusion

In this paper, a deep learning model on salient object detection was proposed. The paper focuses on learning channel-wise Spatiotemporal representations for video salient object detection. Local feature extraction was developed using VGG backbone. During the training process, the VGG backbone learns the mutual and discriminator feature representations. Both Spatial-wise attention and channel-wise attention modules were used to extract the important features of an image and to produce effective characteristics in order to calculate spatial attention. The results of the deep learning model successfully fulfilled the aim of this paper and proved that the model was able to detect the most important object and can be further used in complex cases.

Acknowledgment

This research work was made possible by research grant support (QUEX-CENG-SCDL-19/20-1) from Supreme Committee for Delivery and Legacy (SC) in Qatar.

References

[1] Elharrouss, O., Abbad, A., Moujahid, D., Riffi, J., & Tairi, H. (2016). A block-based background model for moving object detection. ELCVIA: electronic letters on computer vision and image analysis, 15(3), 0017-31.

[2] Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., & Himeur, Y. (2021). Panoptic segmentation: a review. arXiv preprint arXiv:2111.10250.

[3] Hassen Mohammed, H., Elharrouss, O., Ottakath, N., Al-Maadeed, S., Chowdhury, M. E., Bouridane, A., & Zughaier, S. M. (2023). Ultrasound Intima-Media Complex (IMC) Segmentation Using Deep Learning Models. Applied Sciences, 13(8), 4821.

[4] Akbari, Y., Elharrouss, O., & Al-Maadeed, S. (2022). Feature fusion based on joint sparse representations and wavelets for multiview classification. Pattern Analysis and Applications, 1-9.

[5] Tang, Y., Zou, W., Hua, Y., Jin, Z., & Li, X. (2020). Video salient object detection via spatiotemporal attention neural networks. Neurocomputing, 377, 27-37.

[6] Bi, H. B., Lu, D., Zhu, H. H., Yang, L. N., & Guan, H. P. (2021). STA-Net: spatial-temporal attention network for video salient object detection. Applied Intelligence, 51(6), 3450-3459.

[7] Huang, K., Li, G., & Liu, S. (2020). Learning channel-wise spatio-temporal representations for video salient object detection. Neurocomputing, 403, 325-336.

[8] Borji, A., Cheng, M. M., Hou, Q., Jiang, H., & Li, J. (2019). Salient object detection: A survey. Computational visual media, 5(2), 117-150.

[9] Dong, S., Gao, Z., Pirbhulal, S., Bian, G. B., Zhang, H., Wu, W., & Li, S. (2020). IoT-based 3D convolution for video salient object detection. Neural computing and applications, 32(3), 735-746.

[10] Wang, Q., Zhang, L., Li, Y., & Kpalma, K. (2020). Overview of deep-learning based methods for salient object detection in videos. Pattern Recognition, 104, 107340.

[11] Shokri, M., Harati, A., & Taba, K. (2020). Salient object detection in video using deep non-local neural networks. Journal of Visual Communication and Image Representation, 68, 102769.

[12] Kousik, N., Natarajan, Y., Raja, R. A., Kallam, S., Patan, R., & Gandomi, A. H. (2021). Improved salient object detection using hybrid Convolution Recurrent Neural Network. Expert Systems with Applications, 166, 114064.

[13] Cong, R., Lei, J., Fu, H., Porikli, F., Huang, Q., & Hou, C. (2019). Video saliency detection via sparsity-based reconstruction and propagation. IEEE Transactions on Image Processing, 28(10), 4819-4831.

[14] Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., & Tang, J. (2019). RGB-T image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia, 22(1), 160-173.

[15] Huang, K., & Gao, S. (2020). Image saliency detection via multi-scale iterative CNN. The Visual Computer, 36(7), 1355-1367.

[16] Zhang, Z., Jin, W., Xu, J., & Cheng, M. M. (2020, August). Gradient-induced co-saliency detection. In European Conference on Computer Vision (pp. 455-472). Springer, Cham.

[17] Jiang, F., Kong, B., Li, J., Dashtipour, K., & Gogate, M. (2021). Robust visual saliency optimization based on bidirectional Markov chains. Cognitive Computation, 13, 69-80.

[18] Wang, X., Dong, Y., Zhang, Q., & Wang, Q. (2021). Region-based depth feature descriptor for saliency detection on light field. Multimedia Tools and Applications, 80(11), 16329-16346.

[19] Jian, M., Wang, J., Yu, H., Wang, G., Meng, X., Yang, L., ... & Yin, Y. (2021). Visual saliency detection by integrating spatial position prior of object with background cues. Expert Systems with Applications, 168, 114219.

[20] Bi, H., Lu, D., Li, N., Yang, L., & Guan, H. (2019, September). Multi-level model for video saliency detection. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 4654-4658). IEEE.

[21] Chen, C., Song, J., Peng, C., Wang, G., & Fang, Y. (2020). A Novel Video Salient Object Detection Method via Semi-supervised Motion Quality Perception. arXiv preprint arXiv:2008.02966.

[22] Chen, C., Wang, G., & Peng, C. (2019). Structure-aware adaptive diffusion for video saliency detection. IEEE Access, 7, 79770-79782.

[23] Zhao, T., & Wu, X. (2019). Pyramid feature attention network for saliency detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3085-3094).

[24] Shang, J., Liu, Y., Zhou, H., & Wang, M. (2021). Moving object properties-based video saliency detection. Journal of Electronic Imaging, 30(2), 023005.

[25] Chen, C., Wang, G., Peng, C., Zhang, X., & Qin, H. (2019). Improved robust video saliency detection based on long-term spatial-temporal information. IEEE transactions on image processing, 29, 1090-1100.

[26] Min, K., & Corso, J. J. (2019). Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2394-2403).

[27] Xu, M., Liu, B., Fu, P., Li, J., & Hu, Y. H. (2019). Video saliency detection via graph clustering with motion energy and spatiotemporal objectness. IEEE Transactions on Multimedia, 21(11), 2790-2805.

[28] Liu, Z., Shi, S., Duan, Q., Zhang, W., & Zhao, P. (2019). Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing, 363, 46-57.

[29] Liu, B., Mu, K., Xu, M., Wang, F., & Feng, L. (2021). A novel spatiotemporal attention enhanced discriminative network for video salient object detection. Applied Intelligence, 1-16.

[30] Fang, Y., Ding, G., Wen, W., Yuan, F., Yang, Y., Fang, Z., & Lin, W. (2019). Salient object detection by spatiotemporal and semantic features in real-time video processing systems. IEEE Transactions on Industrial Electronics, 67(11), 9893-9903.

[31] Ji, Y., Zhang, H., Zhang, Z., & Liu, M. (2021). CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. Information Sciences, 546, 835-857.

[32] Mohammadi, S., Noori, M., Bahri, A., Majelan, S. G., & Havaei, M. (2020). CAGNet: Content-aware guidance for salient object detection. Pattern Recognition, 103, 107303.

[33] Wang, B., Liu, W., Han, G., & He, S. (2020). Learning long-term structural dependencies for video salient object detection. IEEE Transactions on Image Processing, 29, 9017-9031.

[34] Wu, J., Zhou, W., Luo, T., Yu, L., & Lei, J. (2021). Multiscale multilevel context and multimodal fusion for RGB-D salient object detection. Signal Processing, 178, 107766.

[35] Ji, Y., Zhang, H., Jie, Z., Ma, L., & Wu, Q. J. (2020). CASNet: A cross-attention siamese network for video salient object detection. IEEE transactions on neural networks and learning systems, 32(6), 2676-2690.

[36] Zhang, Q., Wang, S., Wang, X., Sun, Z., Kwong, S., & Jiang, J. (2020). A multi-task collaborative network for light field salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 31(5), 1849-1861.

[37] Fan, D. P., Wang, W., Cheng, M. M., & Shen, J. (2019). Shifting more attention to video salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8554-8564).

[38] Bi, H., Lu, D., Li, N., Yang, L., & Guan, H. (2019, September). Multi-level model for video saliency detection. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 4654-4658). IEEE.

[39] Liu, Y., Shen, C., Yu, C., & Wang, J. (2020, August). Efficient semantic video segmentation with per-frame inference. In European Conference on Computer Vision (pp. 352-368). Springer, Cham.

[40] Wang, W., Zhou, T., Porikli, F., Crandall, D., & Van Gool, L. (2021). A survey on deep learning technique for video segmentation. arXiv preprint arXiv:2107.01153.

[41] Ma, D., Zhang, F., & Bull, D. (2021). BVI-DVC: a training database for deep video compression. IEEE Transactions on Multimedia.

[42] Bross, B., Wang, Y. K., Ye, Y., Liu, S., Chen, J., Sullivan, G. J., & Ohm, J. R. (2021). Overview of the versatile video coding (VVC) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology, 31(10), 3736-3764.]

[43] Bin, Y., Shang, X., Peng, B., Ding, Y., & Chua, T. S. (2021, October). Multi-Perspective Video Captioning. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 5110-5118).

[44] Chen, J., Li, S. E., & Tomizuka, M. (2021). Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. IEEE Transactions on Intelligent Transportation Systems.

[45] Ding, Y., Ma, Z., Wen, S., Xie, J., Chang, D., Si, Z.,& Ling, H. (2021). AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Transactions on Image Processing, 30, 2826-2836.

[46] Liu, N., Zhang, N., Wan, K., Shao, L., & Han, J. (2021). Visual saliency transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4722-4732).

[47] Yang, Z., Soltanian-Zadeh, S., & Farsiu, S. (2022). BiconNet: an edge-preserved connectivity-based approach for salient object detection. Pattern recognition, 121, 108231.

[48] Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020). Multi-scale interactive network for salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9413-9422).

360-GAN: Cycle-Consistent GAN for Extrapolating 360-Degree Field-of-View

Jit Chatterjee
Department of Electrical Engineering (ESAT)
e-Media Research Lab
KU Leuven
Leuven, Belgium
jit.chatterjee@kuleuven.be

Maria Torres Vega
Department of Electrical Engineering (ESAT)
e-Media Research Lab
KU Leuven
Leuven, Belgium
maria.torresvega@kuleuven.be

***Abstract*—360-degree images, also known as panoramic, have become increasingly popular in the field of Extended Reality (XR). They offer an immersive experience to users, allowing them to explore images in a more engaging and dynamic manner. However, the visual quality associated with 360 images in XR can vary greatly depending on factors, such as image resolution with crisp details and vibrant colors. Thus, complex camera systems are required to shoot 360-degree environments. Generative adversarial networks (GANs), which have already been successfully applied to out-painting tasks and for the generation of masked regions in images, have the potential to solve the need for complex infrastructures. As such, from only a small RGB crop, the full environment could be generated. However, traditional GANs can fail to blend the input crop with the generated extrapolated region by introducing sharp vertical edges that disrupt the overall visual coherence. Another challenge in generating 360-degree images is the representation and handling of the spherical geometry of the panorama. In this work, we present 360-GAN, a cycle-consistent GAN model to generate 360-degree omnidirectional images from small RGB crops. Moreover, to maintain the spherical consistency of the generated 360 panoramic images, our method uses Structural Similarity Index (SSIM) as an added loss function. We evaluate our approach through quantitative measurements, benchmarking them against other state-of-the-art approaches. Our method generates realistic results maintaining the spherical consistency of the omnidirectional images with a Fréchet Inception Distance (FID) of 46.59, nearly 6 points better than the most current state-of-the-art methods.**

***Index Terms*—Deep Generative Networks, Cycle-GAN, 360-Degree image, SSIM.**

I. Introduction

Omnidirectional 360 images, also known as spherical or panoramic images, capture a complete 360-degree Field-of-View (FoV). One of the key benefits of 360-degree images is their ability to provide a sense of presence and immersion, allowing viewers to feel like they are physically present in the captured environment. Users can navigate through the image by panning, tilting, and zooming to explore the entire scene from different perspectives. This makes them ideal for creating virtual tours, interactive experiences, and virtual reality (VR) applications. Creating realistic panoramic images, however, typically requires specialized cameras, multiple image stitching [1] [2], and post-processing techniques.

As humans, our perception of the visual world is not solely determined by the specific FoV of our eyes. When we look at a scene, our brain combines the information from our eyes with prior knowledge and experience to construct a more comprehensive understanding of the environment. Our visual system is constantly making predictions and filling in missing information based on contextual cues, past experiences, and learned expectations. This process is known as perceptual completion or ”filling in”. For example, a person standing in a room with eyes focused on a particular object, can still perceive the surrounding environment and have a sense of what is outside our direct FoV. This is because our brain is using visual cues and context to extrapolate and imagine the rest of the scene. It takes into account factors like object continuity, perspective, and spatial layout to generate a coherent representation of the environment. As a regular camera lens has a FoV of 72 degrees, the question is if we, as humans, can somehow imagine the remaining part of the scene by predictions. Will it be possible for an intelligent system to generate a 360-degree image from a small RGB crop? While generating a complete 360-degree image from a small RGB crop is challenging, with the recent immense progress of computer vision and Artificial Intelligence (AI), mimicking human imaginary predictions becomes more of a reality. Image out-painting, or extrapolating the content outside the regular FoV of an image, allows a small crop RGB image (i.e. the FoV), to generate the full 360-degree view. In this paper, we have introduced a method based on Deep Generative Networks for 360-degree FoV extrapolation.

For several decades, the generation of photorealistic images using traditional image processing techniques [3] [4] has been a complex and time-consuming endeavor, often requiring painstaking manual adjustments and intricate algorithms to mimic the intricacies of real-world scenes. Recently, Deep Generative Networks [5] [6] has evolved with promising results but poses challenges when it comes to extrapolating the FoV for 360-degree images as it is designed to operate on planar grids. Maintaining spatial continuity across the generated images is vital for a realistic experience. Since neighboring regions in a 360-degree image are spatially connected, any inconsistencies or artifacts between these regions can disrupt

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

the immersive effect.

The purpose of this paper is to provide a model tackling both the blurriness and the spherical inconsistency. Herein, we present 360-GAN, a cycle-GAN [7]-based model, where the two GANs tend to perform domain adaptation. As described in Figure 1, the first GAN generates 360-degree images from small RGB crops and the second GAN model generates small RGB crops from 360-degree images. Both of the models have individual adversarial losses with cycle consistency loss. This procedure aims to reduce the blurriness of the output 360-degree image. Moreover, to maintain spherical consistency, SSIM is used as a new loss function. We compare our method with state-of-the-art algorithms, where 360-GAN outperforms in all cases both quantitatively and qualitatively, by generating 360-degree images.

In the following sections, we have gone through the different state-of-the-art methods for 360-degree FoV extrapolation, their limitations, and how the research has progressed over the past years. In the methodology section, we have discussed cycle-GAN [7] for panoramic images and how it's modified into 360-GAN with the added SSIM loss to maintain the spherical consistency of the generated 360-degree images. In the final section, we benchmarked our 360-GAN method with state-of-the-art methods using objective metrics. We show that our 360-GAN method can extrapolate the FoV for 360-degree image completion better than other methods.

II. Related Works

In this section, we have discussed various state-of-the-art 360-degree image synthesis methods and how different algorithms evolved over the past years. It contains the advancement of different methods starting from the planar image synthesis to in-painting algorithms and finally out-painting methods for 360-degree image generation.

A. Planar image synthesis

Earlier, different texture synthesis methods [3] [4] were used by extending the FoV of the image with the specific textures. Texture synthesis involves generating new textures that are visually similar to a given input texture. It can be used to create larger textures from a smaller sample or to generate entirely new textures based on a given style or set of exemplar textures. Texture synthesis techniques often utilize statistical models, such as Markov Random Fields (MRF) [8], to capture the characteristics of the input texture and generate coherent and visually plausible results. Recently, Generative Adversarial Networks (GANs) showed promising results [5] [9] [10] [11] [6] and have emerged as a method of choice. However, these methods consider the image scene to be planar, which is not the case for a realistic scene. We consider a scene to be spherical which means the edges of the planar representation will be merging perfectly when represented as a 360-degree view.

B. In-painting

In-painting techniques aim to fill in missing or damaged parts of an image. These methods use information from the surrounding areas to estimate the content that should be present in the missing regions. Diffusion-based in-painting algorithms [12] use partial differential equations to propagate information from the surrounding pixels into the missing region. These methods iteratively estimate the missing values based on the values of neighboring pixels, gradually filling in the in-painted area. GAN-based methods like CoModGAN [13] have been proposed for in-painting, which involves training a generator network to fill in the missing pixels inside an image, while the discriminator network evaluates the realism of the generated images and provides feedback to the generator to improve its output. For instance, CoModGAN [13] generated impressive results for image in-painting as the missing regions are constrained within the boundaries of the input image, providing a clear target for the generator to fill in.

C. FoV Extrapolation

Out-painting, which involves generating new content beyond the boundaries of an input image, is more challenging as it aims at filling in missing regions that do not exist in the original image, requiring the algorithm to create plausible and visually coherent content that seamlessly extends the scene beyond its initial confines. Several GAN methods have been proposed based on Pix2pix [14], which requires cropped images and original images paired for training. A DCGAN-based method [15] generates FoV extrapolation with hazy results. Im2Pano3D [16] predicts a comprehensive 360-degree segmentation map from a regular image, providing valuable clues about the surrounding content captured by the camera. Li *et al.* [17] utilizes a VAE-GAN (Variational Autoencoder GAN) structure that generates edges and edge transformation for FoV extrapolation. Akimoto *et al.* [18] proposed a method that is based on a two-stage conditional GAN to generate 360-degree panoramic images. Recently, Akimoto *et al.* [19] have used a transformer-based architecture to predict 360-degree FoV extrapolation for generating 3DCG backgrounds. Based on the CoModGAN [13] architecture, ImmerseGAN [20] generates plausible results for image out-painting as the generator needs to generate new content that is visually coherent and semantically meaningful while maintaining consistency with the existing content in the input image. However, the generator has little to no contextual information beyond the input image boundaries to guide the generation process. This lack of context makes it difficult to generate realistic and visually coherent content that extends beyond the input image boundaries. Moreover, low-resolution images [14] with pixelation or blurriness can reduce the visual clarity that can negatively impact the sense of presence and engagement.

III. Method

This section describes our methodology on how we have modified the cycle-GAN model [7] to generate realistic, spherically consistent 360-degree images. We have also added a new loss function based on SSIM.

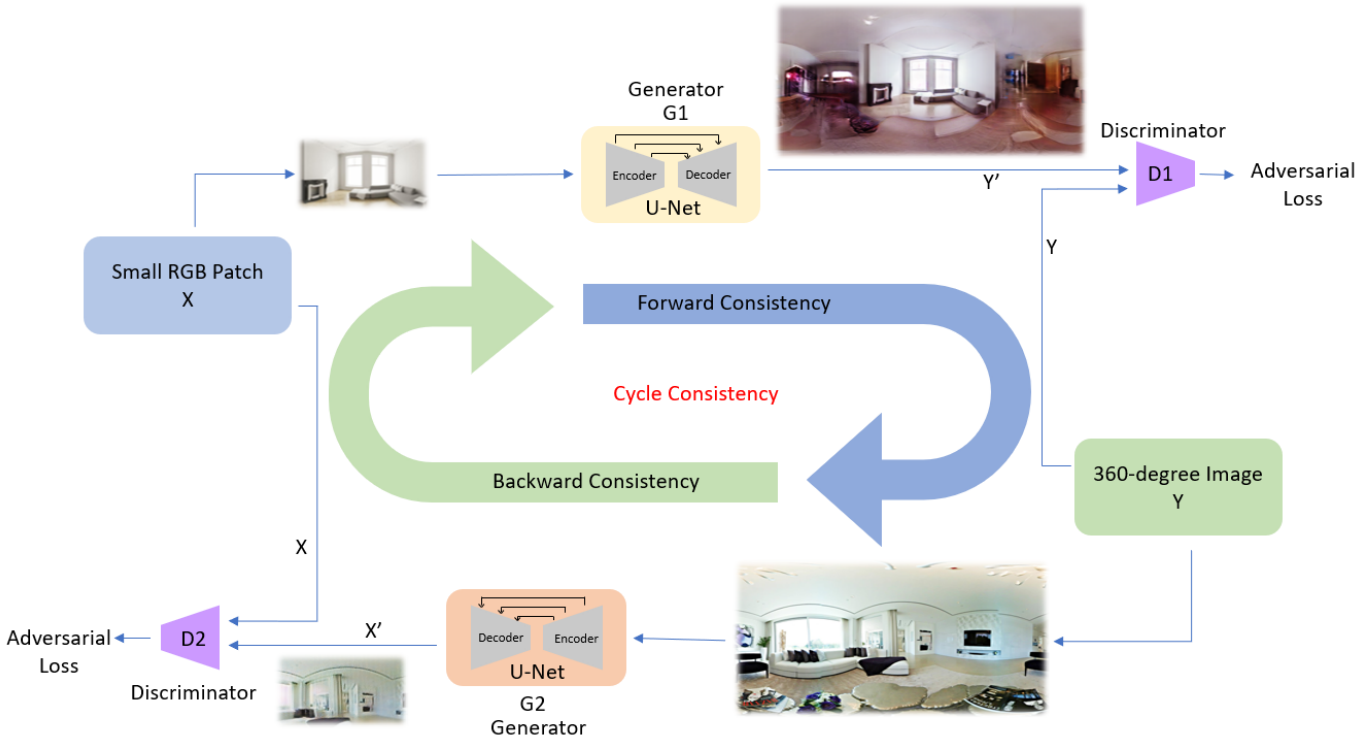


Fig. 1. Our presented method to extrapolate 360-degree Field-of-View using 360-GAN based on cycle-GAN architecture.

A. *Cycle-GAN for panoramic images*

Our work builds on the cycle-GAN [7] architecture. It typically consists of two generator networks (G_1, G_2), and two discriminator networks (D_1, D_2), one for each domain. It includes two mapping functions $G_1 : X \rightarrow Y$ and $G_2 : Y \rightarrow X$. The generator networks learn to generate fake images, while the discriminator networks learn to distinguish between fake and real images. The generator and discriminator networks are trained in an adversarial manner, where the generator tries to generate realistic images to fool the discriminator, while the discriminator tries to correctly classify fake and real images. So the generator and the discriminator are trained iteratively in a two-player minimax game setup. The adversarial loss $\mathcal{L}(G_1, D_1)$ is defined as:

$$\begin{aligned}\mathcal{L}(G_1, D_1) = \min_{\Phi_G} \max_{\Phi_D} \{&\mathbb{E}_y [\log D_1(y)] \\ &+\mathbb{E}_x [\log (1 - D_1 (G_1(x)))]\}\end{aligned} \tag{1}$$

where, Φ_G and Φ_D are the respective parameters of G_1 and D_1, and $x \in X$ and $y \in Y$ shows the unpaired training data in both domains. A similar adversarial loss is defined for the reverse mapping $\mathcal{L}(G_2, D_2)$.

The training data in Cycle-GAN is unpaired where X represents the small RGB crops and Y represents the 360-degree images. The key idea behind Cycle-GAN is the use of cycle consistency, which is achieved by training the generator networks to not only generate images from X to Y but also to be able to reverse the translation and reconstruct the original image. This is done by introducing cycle consistency loss, which penalizes the difference between the original image and the image reconstructed after going through both generator networks in a cycle. Thus the training process of Cycle-GAN involves a cycle-consistency loss term in addition to the standard adversarial loss. The cycle consistency loss encourages the generators to produce images that are consistent when translated back and forth between the two domains X and Y, ensuring that the generated images are plausible and maintain the original content.

$$\begin{aligned}\mathcal{L}(G_1, G_2, D_1, D_2) = \mathcal{L}(G_1, D_1) \\ + \mathcal{L}(G_2, D_2) + \gamma \mathcal{L}_{cycle}(G_1, G_2)\end{aligned} \tag{2}$$

where,

$$\begin{aligned}\mathcal{L}_{cycle}(G_1, G_2) = \|G_2(G_1(x)) - x\|_1 \\ + \|G_1(G_2(y)) - y\|_1\end{aligned} \tag{3}$$

is the cycle-consistency loss and γ is the cycle-loss parameter.

B. *360-GAN*

Figure 1 presents our approach. We introduce an end-to-end trainable pipeline, meticulously designed to cater to the task of a 360-degree FoV extrapolation that generates top-notch panoramas from a single RGB crop image with limited FoV. Based on the cycle-GAN architecture, we have built our 360-GAN with the addition of SSIM loss. It consists of two GANs: one to learn the features from small RGB crops to 360-degree images (forward consistency), and the other to learn the features from 360-degree images to small RGB crops (backward

Fig. 2. Qualitative Field-of-View extrapolation results.

consistency) till the point they reach cycle consistency. The SSIM [21] is a widely used image quality assessment metric that measures the structural similarity between two images. Mathematically, the SSIM loss can be defined as:

$$\mathcal{L}_{SSIM} = 1 - SSIM(I_1, I_2) \tag{4}$$

where SSIM is the Structural Similarity Index between the reference image (I_1) and the distorted image (I_2). To maintain the spherical consistency of the generated 360-degree images, we have added the SSIM loss ($L_{SSIM_{G_1}}$) between 10 edge pixels of the left (I_{left}) and right (I_{right}) sides with the whole image height of the generated panoramic image (Generator G_1). This added loss ensures the blending of the right and left edge pixels of the generated 360-degree image, hence, removing discontinuities. It encourages the generator model (G_1) to produce 360-degree images with no discontinuities at the edges.

The total loss is then calculated as:

$$\begin{aligned}\mathcal{L}_{\mathcal{TOTAL}} = \mathcal{L}\left(G_1, D_1\right) + \mathcal{L}\left(G_2, D_2\right) \\ + \gamma \mathcal{L}_{cycle}\left(G_1, G_2\right) + \lambda \mathcal{L}_{SSIM_{G_1}}\end{aligned} \tag{5}$$

where λ is the SSIM Loss parameter.

As shown in Figure 1, the Generator networks (G_1, G_2) are based on the U-Net [22] architecture. Anisotropic upsampling artifacts can occur when there is a significant difference in the scale or resolution between different dimensions of the image. To mitigate anisotropic upsampling artifacts while mapping the output to the equirectangular representation, based on [20], we made adjustments to the U-Net architecture to ensure a 2:1 aspect ratio. The skip connections, which help in preserving fine-grained details during the upsampling process, are designed to maintain the desired aspect ratio. This modification helps maintain consistent proportions and prevents distortions that may occur during upsampling, resulting in improved visual fidelity in the final equirectangular representation. The Discriminator networks (D_1, D_2) are based on the PatchGAN architecture [23] which outputs a grid of scalar values, where each scalar value corresponds to a small patch of the input image. The discriminator's job is to distinguish between the real output image and the generated output image.

IV. Experiments

This section provides a description of the dataset used for the evaluation of 360-GAN and an analysis of the results.

A. Datasets and Training

Our 360-GAN model was trained on a dataset composed of the 360-Indoor [24] and Matterport3D [25] datasets. While the 360-Indoor dataset consists of a total of 3,335 panoramic RGB images, the Matterport3D dataset consists of 10,800 indoor panoramic images. The total dataset is split into 80% train, 10% validation, and 10% test subsets. While training, random crops are computed with FoVs between 70°and 80°to ensure as diverse a set as possible.

To enhance the diversity of our dataset, we implemented data augmentation techniques including random scaling and translations, which allows for variations of up to 15% in size compared to the original image. Additionally, we applied random adjustments to the exposure and saturation of the

image, with a maximum factor of 1.2. These techniques introduce variability and augment the training data, enhancing the model's ability to generalize and learn robust features from different image variations. We trained our 360-GAN model for 200 epochs with a decaying learning rate starting with 0.0002.

B. Results

To provide realistic benchmarks, we selected the state-of-the-art methods pix2pixHD [9] and ImmerseGAN [20]. Therefore, we trained both pix2pixHD [9] and ImmerseGAN (unguided) [20] on our dataset and tested them under the same conditions as for 360-GAN. The FoV of the RGB crops ranges between 70°and 80°which is also the FoV of a normal camera lens. Qualitative results are presented in Figure 2. They show that pix2pixHD [9] generates bad quality results as the GAN model fails into mode collapse. ImmerseGAN [20], which is based on CoModGAN [13], generates results with discontinuities. Our method, 360-GAN generates realistic omnidirectional images with plausible environments maintaining the spherical consistency. As shown in Figure 2, the generated 360-degree images using our method 360-GAN ensures color accuracy, sharpness, texture, and overall visual appearance.

Moreover, to include a more quantitative analysis, we calculated the Fréchet Inception Distance (FID) [26] score for each method as shown in Table I. We have considered FID as the objective metric over peak signal-to-noise ratio (PSNR). PSNR is primarily designed to measure the pixel-level similarity between two images, emphasizing the mean squared error (MSE) between them. However, perceptual quality is not solely determined by pixel-level differences. Human perception of image quality takes into account factors like color distribution, texture, and overall visual appearance, which are not adequately captured by PSNR. FID, on the other hand, is based on feature representations extracted by a pre-trained deep neural network, which better aligns with human perception. 360-degree images are high-dimensional data due to their large spatial extent. PSNR, being a pixel-wise metric, treats all pixels equally and does not consider the spatial arrangement or structural information in the image. FID, utilizes features extracted from a deep network, and captures higher-level semantics and spatial relationships, making it more suitable for evaluating the quality of complex 360-degree images.

TABLE I
QUANTITATIVE ANALYSIS ON THE TEST SET.

Method	FID
pix2pixHD [9]	143.27
ImmerseGAN [20]	52.93
360-GAN (ours)	**46.59**

FID is a measure of the similarity between the distribution of real images and the distribution of generated images, where lower values indicate better similarity. As mentioned in [18], the PSNR does not indicate the performance of different deep generative models as our main goal is completion and not the restoration of the original images. The FID score is computed by first passing a set of real images and a set of generated images through a pre-trained Inception-v3 neural network [27] to obtain feature representations. Then, the mean and covariance of these feature representations are calculated for both sets of images, and the distance between these statistics is measured using the Fréchet distance. The FID score of pix2pixHD [9] is the worst among the three methods tested. We assume that the FID score of the state-of-the-art method ImmerseGAN [20] is not that good compared to the original paper as we have trained on our dataset till the point it starts over-fitting. Based on the quantitative analysis, our method 360-GAN outperforms the state-of-the-art method ImmerseGAN [20].

V. CONCLUSION

The contributions of this paper can be outlined as follows. To begin with, we introduce a novel method 360-GAN based on the cycle-GAN architecture to extend the FoV of a camera to a complete 360-degree panorama with SSIM as an added loss function. This method effectively controls the appearance of the extrapolated content, resulting in spherical consistent omnidirectional images that surpass the current state-of-the-art in both visual quality and the standard FID metric. We are confident that the results of our method 360-GAN can be improved if trained on a much larger dataset as the cycle-GAN model can learn more domain knowledge. The third row of Figure 2 shows that all of the tested methods produced a discontinuity for the overhead lights. We will investigate this further by using a decaying SSIM loss function to remove the discontinuities in the generated images. Moreover, our method can be used to generate realistic 360-VR scenarios which enhance the quality and user experience of omnidirectional images, making them more immersive, visually appealing, and interactive. In the future, we would like to perform some user-based studies by conducting surveys where participants are asked to rate the visual quality and realism of the generated 360-degree images.

REFERENCES

[1] A. Zisserman, "Multiple view geometry in computer vision," *Künstliche Intell.*, vol. 15, p. 41, 2001.

[2] Y.-S. Chen and Y.-Y. Chuang, "Natural image stitching with the global similarity prior," in *European Conference on Computer Vision*, 2016.

[3] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1033–1038 vol.2, 1999.

[4] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.

[5] C. H. Lin, H.-Y. Lee, Y.-C. Cheng, S. Tulyakov, and M.-H. Yang, "Infinitygan: Towards infinite-pixel image synthesis," in *International Conference on Learning Representations*, 2021.

[6] Y.-C. Cheng, C. H. Lin, H.-Y. Lee, J. Ren, S. Tulyakov, and M.-H. Yang, "Inout: Diverse image outpainting via gan inversion," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 421–11 430, 2021.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[8] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2479–2486, 2016.

[9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2017.

[10] G. Somanath and D. Kurz, "Hdr environment map estimation for real-time augmented reality," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 293–11 301, 2020.

[11] A. Nair, J. Deshmukh, A. Sonare, T. Mishra, and R. Joseph, "Image outpainting using wasserstein generative adversarial network with gradient penalty," *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1248–1255, 2022.

[12] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 451–11 461, 2022.

[13] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I.-C. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," *ArXiv*, vol. abs/2103.10428, 2021.

[14] N. Kimura and J. Rekimoto, "Extvision: Augmentation of visual experiences with generation of context images for a peripheral vision using deep neural network," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[15] M. Sabini and G. Rusak, "Painting outside the box: Image outpainting with gans," *ArXiv*, vol. abs/1808.08483, 2018.

[16] S. Song, A. Zeng, A. X. Chang, M. Savva, S. Savarese, and T. A. Funkhouser, "Im2pano3d: Extrapolating 360° structure and semantics beyond the field of view," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3847–3856, 2017.

[17] X. Li, H. Zhang, L. Feng, J. Hu, R. Zhang, and Q. Qiao, "Edge-aware image outpainting with attentional generative adversarial networks," *IET Image Process.*, vol. 16, pp. 1807–1821, 2022.

[18] N. Akimoto, S. Kasai, M. Hayashi, and Y. Aoki, "360-degree image completion by two-stage conditional gans," *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4704–4708, 2019.

[19] N. Akimoto, Y. Matsuo, and Y. Aoki, "Diverse plausible 360-degree image outpainting for efficient 3dcg background creation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 431–11 440, 2022.

[20] M. R. K. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, S. Khodadadeh, and J.-F. Lalonde, "Guided co-modulated gan for 360° field of view extrapolation," *2022 International Conference on 3D Vision (3DV)*, pp. 475–485, 2022.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, 2004.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2016.

[24] S.-H. Chou, C. Sun, W.-Y. Chang, W. T. Hsu, M. Sun, and J. Fu, "360-indoor: Towards learning real-world objects in 360° indoor equirectangular images," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 834–842, 2019.

[25] A. X. Chang, A. Dai, T. A. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, 2017.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, 2017.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2015.

All Predictions Matter: an Online Video Prediction Approach

Melan Vijayaratnam*, Marco Cagnazzo*†, Giuseppe Valenzise‡ and Enzo Tartaglione*

*LTCI, Télécom Paris, Institut Polytechnique de Paris, France
†University of Padua, Department of Information Engineering, Italy
‡Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France

***Abstract*—To effectively manage and utilize the massive amount of visual data generated by the surging number of videos, decision-making systems must predict and reason about future outcomes. This paper proposes a novel online approach for video prediction that enables continual learning in the presence of new data, as periodic training of neural networks may not be practical. We utilize all predictions, including intermediate computations obtained during the inference process, to improve the performance of video prediction. To achieve this, we incorporate a weighting scheme in the loss that accounts for all the predictions during the learning process. Additionally, we leverage semantic segmentation to assess the performance of extrapolated frames by focusing on the position of the objects in the scene. Our approach stands out from state-of-the-art methods as it uses intermediate predictions, which are available due to the iterative nature of forecasting future frames. Our method improves the offline counterpart for the same network by 1.45 dB for predicting five steps in the future.**

***Index Terms*—Extrapolation, video prediction, online learning, metric, segmentation**

I. Introduction

The human capacity to forecast future events and adapt present behavior accordingly is a well-established phenomenon in cognitive psychology and behavioral sciences [1]. As such, expecting the same for systems is key for understanding about the world that surrounds us. The applications of video prediction range from assisting in medical diagnosis [2], for autonomous driving to help the car to anticipate and react to potential hazards on the road [3] to low-latency video transmission [4]. Being a self-supervised task, the understanding only comes from the data itself, preventing the need for data labeling efforts.

Online deep learning methods have been presented as a way to scale with the stream of data [5]. It has been studied more specifically in various fields of computed vision, from classification [6] to semantic segmentation [7]. In the presented work of Zhang *et al.* [8], the authors apply online learning to video depth estimation that would normally require labeled data for the network to be updated but they devise a technique to be able to do so in a self-supervised way. Video prediction networks also benefit from the online learning paradigm, which encourage to present a novel methodology that can apply to all such networks.

To enhance the accuracy of video prediction for distant future sequences, we utilize intermediate frames in the prediction process. These intermediate frames are saved and combined with new ground truth images as they are received to update the model based on a weighting of all predictions in the loss computation. Overall, our method improves the performance of video prediction, particularly for longer temporal horizons, resulting in more accurate predictions of future frames in a video sequence.

Furthermore, we present a method for evaluating video prediction algorithms at the object level. We accomplish this by borrowing from the field of semantic segmentation and creating a pseudo label segmented image from the ground truth, which we then compare to extrapolated frames. As a result, we focus on the objects in the scenes and their locations rather than the entire scene.

II. Related Work

In the typical setup of evaluation of deep learning architectures, the model weights are typically learned on the training set, the hyperparameters are fine-tuned on the validation set, and the pre-trained weights are used on the test set. This is the batch-learning strategy, where the system learns the model only once. Before being deployed, the model is pre-trained offline, and afterward, it is frozen. Online learning techniques differ in that they continuously update and improve a model's performance as new data becomes available. The model is trained on a stream of data, with each new observation providing an opportunity for the model to learn and adapt in real-time. Interest in online learning has emerged for classification tasks [9], formally introduced in [10]. Existing video extrapolation methods [11] only considered the batch learning paradigm. Our target use case, on the other hand, has a critical distinction that allows us to progress toward a more effective framework. More precisely, for any image predicted by the extrapolator, its ground truth (the actual image) will arrive and allow a refinement of the neural network. This idea naturally leads us toward the on-line learning paradigm. The approach we propose in this paper is based on online learning [12] and allows the system to learn the model on the fly which means keeping learning even after being deployed as new data arrives.

Since video prediction is a self-supervised task [13], there is no need for human annotation as the information is already

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

present in the data. Zhang *et al.* [8] apply online adaptation to to consider the task of depth estimation as a self-supervised task in a self-supervised manner not to require depth data explicitly and adapt to evolving data streams. Later, the concept was developed for online monocular depth estimation [14]. Online learning has been shown to improve streaming policies [15]. Our work is connected to these studies as they employ video depth estimation in an online environment, similar to our objective of developing video extrapolation networks that function with online streams of video sequences.

III. Online video prediction scheme

In certain applications, such as compensating for latency through extrapolation [4], it is essential to have the ability to make predictions at a specific horizon in the future. The horizon h is defined as the number of frames we want to predict in the future. As it is well known in the literature, the larger h, the more difficult it is to get a reliable prediction. To address the decrease in prediction accuracy when dealing with large values of h, a frequently employed approach involves the iterative application of the prediction network. This entails making predictions for future frames within a shorter time horizon, and subsequently using these predictions as input to the prediction network to extrapolate frames farther away in time [11]. This iteration process can be exploited in online learning by defining a loss function that employs a weighted mean of the errors of each intermediate prediction. In an online setting, it means that as soon as a new frame from the sequence arrives, multiple forward passes coming from all approximations of the new images will occur.

Figure 1 presents the proposed scheme for online video prediction. To predict the sequence stream ahead of h frames, we start from the pre-trained weights resulting from the training process. By storing past predictions of $\hat{I}_n^h$, i.e., the predicted frames of the ground truth frame I at time step n using horizon h, we use them later when the ground truth arrives to update the prediction network. The input frames from the video prediction network, namely the context frames, can be either true (available) frames, or predicted frames from the iterative process. At each time step, the video extrapolation network $\mathcal{F}$, which would be fixed in an offline learning scenario, is updated. We denote as $\mathcal{F}_n$ the updated model at time n. By following the depicted process, the extrapolated frames for I_n can be obtained as follows (assuming as an example that $\mathcal{F}$ takes 2 context frames as input):

$$\hat{I}_n^1 = \mathcal{F}_{n-1}(I_{n-1}, I_{n-2}) \tag{1}$$

$$\hat{I}_n^2 = \mathcal{F}_{n-2}\left(\mathcal{F}_{n-2}(I_{n-2}, I_{n-3}), I_{n-2}\right) \tag{2}$$

$$\hat{I}_n^3 = \mathcal{F}_{n-3}\left(\mathcal{F}_{n-3}\left(\mathcal{F}_{n-3}(I_{n-3}, I_{n-4}), I_{n-3}\right), \hat{I}_{n-2}^1\right) \tag{3}$$

More in general, when we recursively re-circulate the last predicted output back as input h times in order to predict h steps in the future, frame n is predicted h times: at time $n-1$, $n-2, \ldots, n-h$. We can define a new overall loss $\mathcal{L}^*$ that takes advantage on one hand of all these intermediate predictions, and on the other of the availability ground truth frames:

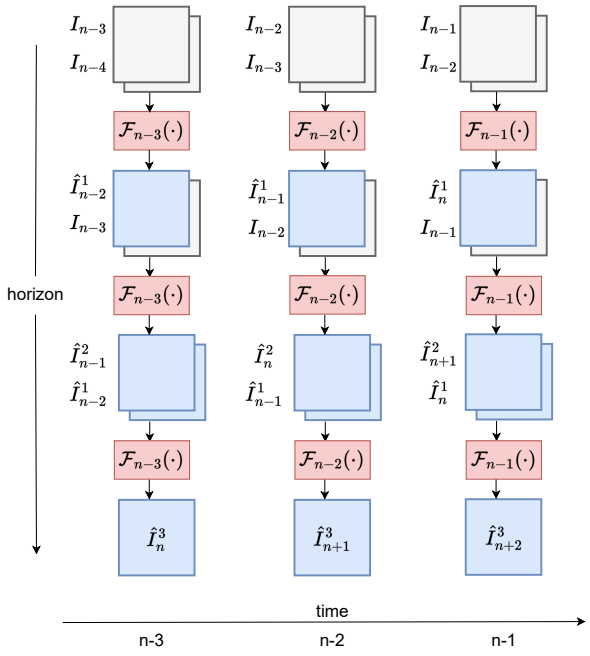


Fig. 1: Prediction of 3 steps in the future. The vertical axis represents how far we want to predict in the future and the horizontal one represents the stream of data arriving. Ground truth frames are depicted in gray and predicted frames in blue. The ground truth frames allow to get the predicted sequence 3 steps in the future $\hat{I}_n^3$, $\hat{I}_{n+1}^3$, $\hat{I}_{n+2}^3$. All the intermediate computed frames will be used to update the network as ground truth arrives.

$$\mathcal{L}^* = \sum_{i=1}^{h} \lambda_i \, \mathcal{L}\left(\hat{I}_n^i; I_n\right), \tag{4}$$

where λ refers to the weight assigned to each of the different predictions. The loss $\mathcal{L}^*$ is a weighted sum of all the per-frame losses $\mathcal{L}\left(\hat{I}_n^i; I_n\right)$ over the horizon h where $\mathcal{L}\left(\hat{I}_n^i; I_n\right)$ is often a mean squared error, but other relevant loss metrics can be used. At the arrival of new ground truth frame, the network will update itself with the loss with the formulation in Equation 4.

IV. Semantic segmentation based metric

PSNR has been criticized for not being a good objective fidelity metric [16]. Regardless, it is still widely popular and used to compare different frames from videos. It relies on every pixel of the reference frame and compares it to a target frame. Using methods from the semantic segmentation field [17], we may further verify the accuracy of the pixels at the object level in the scene. Semantic segmentation involves assigning per-pixel predictions of object categories to an image, providing a comprehensive description of the scene that includes information about the object category, location,

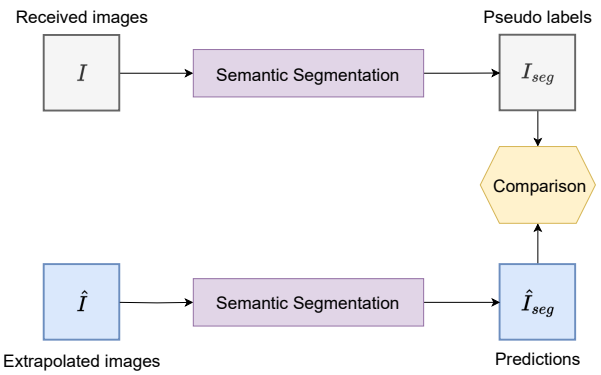


Fig. 2: Semantic segmentation based metric for video prediction.

and shape. By applying semantic segmentation to the images in question, we can observe the positions of the objects and confirm the observations made earlier. We elaborate on the method to evaluate the extrapolation methods using semantic segmentation. We demonstrate a method to evaluate the extrapolation methods using a semantic segmentation on Figure 2 scheme. The received frames I are fed to a semantic segmentation network to generate pseudo labels (since the ground truth is not provided) for segmentation $\tilde{I}_{seg}$, and compared to the segmentation maps $\hat{I}_{seg}$ computed on the extrapolated images $\hat{I}$. For this evaluation, we choose DeepLabv3+ [18] pre-trained on Cityscapes, having Resnet-101 as backbone. The adopted evaluation metric is the Intersection-Over-Union, denoted as IoU: this is a method to quantify the overlap between the target segmentation mask and our prediction segmentation output over the union of both quantities.

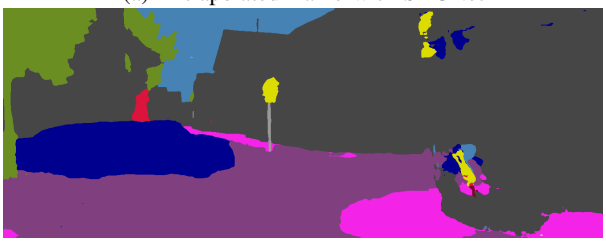
V. Experiments and Discussion

In the following experiments, we analyze the effect of the proposed online video prediction technique and evaluate them using common metrics and the segmentation-based metric introduced in this work.

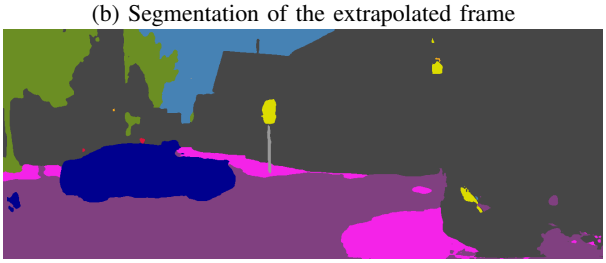
A. Datasets

We train the learning-based extrapolation methods, MCNet [19] and SDCNet [20] on the Caltech Pedestrian dataset [21], collected from a vehicle driving through regular traffic in an urban environment. The dataset consists of around 10 hours of dashcam footage with 65 different video sequences captured at 30 fps. We additionally use sequences from the Kitti [22] and DriveSeg [23] manual scene for evaluation purposes which are both datasets taken the same way as Caltech pedestrian from a moving vehicle. We use the sequence #14 from Kitti consisting of 320 frames and the first 500 frames of DriveSeg. Regarding the optical flow-based method FlowNet2 [24], we only make use of the pre-trained weights on MPI-Sintel which is an optical flow data set derived from the film Sintel [25].

(a) Extrapolated frame with SDCNet



(b) Segmentation of the extrapolated frame



(c) Segmentation of the true image

Fig. 3: Segmentation outputs for predicting one step in the future. Image taken from the Kitti dataset.

B. Choice of video prediction networks

As discussed in [26], video prediction methods can be motion-based, pixel-based, or fusion-based. Motion-based methods focus on the motion in the image which could be done with the optical flow information. Pixel-based methods generate the entirety of the pixels from scratch and finally, fusion-based methods combine both motion and pixel-based methods. We choose a technique from each class, starting with FlowNet2 [24] for predicting optical flow. Combined with a warping that moves the pixels according to the optical flow, an estimate of the next image can be obtained. MCNet [19] uses long short-term memory modules from image differences to generate a new frame. SDCNet [20] uses both optical flows and convolutional kernels from the pixels to generate the extrapolated frame. We perform offline experiments that correspond to having the weights of the neural network being frozen at validation as well as online experiments on SDCNet, with weights learning during validation.

We also include a simple frame-copy extrapolation, dubbed CopyLast. This method just copies the last available frame. Although it is not a real extrapolation method, it is often used as a reference. In particular, for understanding the visual quality of the prediction: if the predicted image is not better than CopyLast, it means that we are introducing large artifacts.

Approach	PSNR ↑			SSIM ↑			VMAF ↑		
	h=1	h=3	h=5	h=1	h=3	h=5	h=1	h=3	h=5
CopyLast	21.25	18.87	17.96	0.50	0.42	0.40	16.12	9.33	8.05
MCNet	23.19	20.66	19.36	0.60	0.52	0.49	19.84	8.91	6.47
FlowNet2 + warp	24.92	21.44	20.03	0.73	0.53	0.48	32.55	10.89	7.04
SDCNet offline	25.38	23.18	22.06	0.76	0.68	0.65	39.59	24.51	18.37
SDCNet online (ours)	**26.53**	**24.07**	**22.73**	**0.83**	**0.75**	**0.71**	**51.27**	**32.86**	**24.55**

(a) Quantitative results on Kitti scene 014

Approach	PSNR ↑			SSIM ↑			VMAF ↑		
	h=1	h=3	h=5	h=1	h=3	h=5	h=1	h=3	h=5
CopyLast	27.65	23.64	22.21	0.72	0.54	0.45	47.34	29.22	22.49
MCNet	28.84	25.20	22.68	0.89	0.74	0.61	61.05	40.78	27.70
FlowNet2 + warp	31.82	27.00	24.72	0.92	0.79	0.65	71.77	42.26	26.86
SDCNet offline	34.23	29.93	28.21	0.95	0.88	0.83	80.44	56.91	45.23
SDCNet online (ours)	**35.89**	**31.71**	**29.66**	**0.98**	**0.93**	**0.89**	**87.58**	**69.48**	**57.64**

(b) Quantitative results on DriveSeg

TABLE I: Comparison of the proposed online method with other extrapolation methods

C. Experimental results

In Table I we observe the PSNR in the YCbCr color space [27], SSIM [28], and VMAF [29], as they are widely used objective metrics. The reference extrapolated video is compared to the original input sequences. CopyLast serves as a simple baseline that uses the last available frame and corresponds to not anticipating the future while FlowNet2 combined with a warping allows predicting the future frames. For every extrapolation horizon, the weights are reinitialized from the pre-trained weights. The weights assigned to the λ are chosen so that $\lambda_i = 1 \, \forall i$, signifying that each of the parts of the sum given in the equation 4 has equal importance. The online proposed method applied to SDCNet outperforms the same network in offline mode by 0.89 dB in Kitti and 1.78 dB in DriveSeg at horizon $h = 3$, meaning predicting three steps in the future, which results in a latency compensation of 100 ms.

D. Ablation study

We perform multiple experiments to validate our proposed online approach for video prediction. To do so, we compare our proposed method, which we call "Uniform" due to the equal importance to every predictions. "First only" corresponds to considering the first prediction only and "Last only" only the last prediction. Table II shows that our approach outperforms the competing approaches, and proves the proposed approach of considering every prediction is beneficial to the network. At $h = 1$, the methods behave the same due to having a single weighting term, therefore we do not report these results as these can be found in Table I.

E. Discussion about segmentation

In Table III, we report the intersection over union (IoU) of the class "car", which is predominant in the chosen sequences. In the Kitti scene, the IoU seems to follow the same trend as the PSNR and demonstrates that the online adaptation brings an increase in performance. Concerning DriveSeg, the IoU from both methods are very close, which contradicts the PSNR results of the online outperforming CopyLast. Upon further examination, it was discovered that in the Kitti dataset, the moving cars are spaced further apart from each other compared to the DriveSeg dataset where the cars are closely grouped together. The image in Figure 3 displays an issue caused by extrapolation at the back of the car, resulting in the segmentation network incorrectly categorizing this artifact as a car.

Weighting λ_i	PSNR ↑			
	h=2	h=3	h=4	h=5
First only	24.95	23.99	23.27	22.66
Last Only	24.91	23.89	23.09	22.46
Uniform	**25.05**	**24.07**	**23.37**	**22.73**

(a) Kitti scene

Weighting λ_i	PSNR ↑			
	h=2	h=3	h=4	h=5
First only	33.24	31.59	30.46	29.57
Last Only	33.20	31.38	30.12	29.21
Uniform	**33.32**	**31.71**	**30.58**	**29.66**

(b) DriveSeg

TABLE II: Ablation study on the weighting in the online scheme

VI. Conclusion

This paper introduces an online learning algorithm for video prediction. We exploit every prediction to improve the video extrapolation network and not just the resulting frames of the desired horizon. This comes at the price of additional complexity by making use of intermediate and unused pre-

Approach	IoU car ↑ h=1	h=2	h=3	h=4	h=5
CopyLast	0.50	0.29	0.19	0.16	0.17
MCNet	0.38	0.20	0.14	0.09	0.18
FlowNet2 + warp	0.70	0.53	0.40	0.30	0.22
SDCNet offline	0.69	0.56	0.45	0.34	0.23
Ours	**0.72**	**0.58**	**0.55**	**0.48**	**0.29**

(a) IoU for Kitti scene 14

Approach	IoU car ↑ h=1	h=2	h=3	h=4	h=5
CopyLast	0.87	0.83	0.80	0.78	0.75
MCNet	0.80	0.72	0.67	0.64	0.58
FlowNet2 + warp	0.86	0.83	0.79	0.76	0.73
SDCNet offline	0.86	0.80	0.73	0.73	0.69
Ours	**0.88**	**0.84**	**0.81**	**0.78**	**0.76**

(b) IoU for DriveSeg

TABLE III: Intersection over Union comparison between CopyLast and extrapolation methods over Kitti and DriveSeg for the car class.

dicted frames but with an increase in quality as demonstrated by the experiments. The segmentation-oriented quality metric focusing on the object rather than every pixel also seems promising and may stimulate further work towards enforcing shape consistency of objects in difficult environments.

VII. Acknowledgments

This work was funded by the ANR AAPG2020 national fund (ANR-20-CE25-0014).

References

[1] T. Suddendorf and J. Redshaw, "Anticipation of future events," *Encyclopedia of animal cognition and behavior*, pp. 1–9, 2017.

[2] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, et al., "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[3] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2020.

[4] M. Vijayaratnam, M. Cagnazzo, G. Valenzise, A. Trioux, and M. Kieffer, "Towards zero-latency video transmission through frame extrapolation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2122–2126.

[5] D. Sahoo, Q. Pham, J. Lu, and S. C. Hoi, "Online deep learning: Learning deep neural networks on the fly," *arXiv preprint arXiv:1711.03705*, 2017.

[6] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.

[7] R. Volpi, P. De Jorge, D. Larlus, and G. Csurka, "On the road to online adaptation for semantic image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19184–19195.

[8] Z. Zhang, S. Lathuiliere, A. Pilzer, N. Sebe, E. Ricci, and J. Yang, "Online adaptation through meta-learning for stereo depth estimation," *arXiv preprint arXiv:1904.08462*, 2019.

[9] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[10] S. Shalev-Shwartz et al., "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[11] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros, "A review on deep learning techniques for video prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2806–2826, 2020.

[12] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," *Neurocomputing*, vol. 459, pp. 249–289, 2021.

[13] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1476–1485.

[14] Z. Zhang, S. Lathuiliere, E. Ricci, N. Sebe, Y. Yan, and J. Yang, "Online depth learning against forgetting in monocular videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4494–4503.

[15] T. Karagkioules, G. S. Paschos, N. Liakopoulos, A. Fiandrotti, D. Tsilimantos, and M. Cagnazzo, "Online learning for adaptive video streaming in mobile networks," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1, pp. 1–22, 2022.

[16] K. Navas, D. K. Gayathri, M. Athulya, and A. Vasudev, "Mwpsnr: A new image fidelity metric," in *2011 IEEE Recent Advances in Intelligent Computational Systems*. IEEE, 2011, pp. 627–632.

[17] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[19] R. Villegas, J. Yang, S. Hong, et al., "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[20] F. A. Reda, G. Liu, K. J. Shih, R. Kirby, J. Barker, D. Tarjan, A. Tao, and B. Catanzaro, "Sdc-net: Video prediction using spatially-displaced convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 718–733.

[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 304–311.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[23] L. Ding, J. Terwilliger, R. Sherony, et al., "MIT driveseg (manual) dataset for dynamic driving scene segmentation," Tech. Rep., Technical report, Massachusetts Institute of Technology, 2020.

[24] E. Ilg, N. Mayer, T. Saikia, et al., "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2462–2470.

[25] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 611–625.

[26] H. Gao, H. Xu, Q.-Z. Cai, R. Wang, F. Yu, and T. Darrell, "Disentangling propagation and generation for video prediction," in *IEEE International Conf. on Computer Vision (ICCV)*, 2019, pp. 9006–9015.

[27] G. Sullivan and K. Minoo, "Objective quality metric and alternative methods for measuring coding efficiency," in *document JCTVC-H0012, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), 8th Meeting: San Jose, CA, USA*, 2012, pp. 1–10.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[29] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.

Evaluating the Vulnerability of Deep Learning-based Image Quality Assessment Methods to Adversarial Attacks

Hanene F. Z. Brachemi Meftah[1], Sid Ahmed Fezza[2], Wassim Hamidouche[1,3], Olivier Déforges[1]
[1]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
[2]National Higher School of Telecommunications and ICT, Oran, Algeria
[3]Technology Innovation Institute P.O.Box: 9639, Masdar City Abu Dhabi, UAE

***Abstract*—Recent studies have discovered that Deep Learning (DL) models are vulnerable to adversarial attacks in image classification tasks. While most studies have focused on DL models for image classification, only a few works have addressed this issue in the context of Image Quality Assessment (IQA). This paper investigates the robustness of different Convolutional Neural Network (CNN) models against adversarial attacks when used for an IQA task. We propose an adaptation of state-of-the-art image classification attacks in both targeted and untargeted modes for an IQA regression task. We also analyze the correlation between the perturbation's visibility and the attack's success. Our experimental results show that DL-based IQA methods are vulnerable to such attacks, with a significant decrease in correlation scores. Consequently, the development of countermeasures against such attacks is essential for improving the reliability and accuracy of DL-based IQA models. To support the principle of reproducible research and fair comparison, we make the codes publicly available on https://github.com/hbrachemi/IQA_AttacksSurvey.**

***Index Terms*—Blind image quality assessment, Adversarial attacks, Robustness, Deep learning, Convolutional neural networks.**

I. Introduction

The impressive development of Deep Learning (DL) and its deployment in different fields introduced major progress in the automation process of many human-related tasks. For example, it has gained significant popularity among the Image Quality Assessment (IQA) community and has become the standard used approach.

On the other hand, Szegedy *et al.* [1] were the first to reveal the vulnerability of DL to adversarial attacks in the context of image classification. They showed that adding small yet carefully crafted perturbations to the input image can lead to its misclassification. This gave rise to serious security vulnerabilities that could be exploited for malicious purposes. Being no exception, IQA models can also fall victim to these adversarial attacks. The operational spectrum of attacks on IQA metrics ranges from inconvenience to end users to life-threatening critical risks. For instance, in content-sharing applications such as social media, tricking a metric into predicting high-quality scores for low-quality visual content can negatively impact the Quality of Experience (QoE) of end users. Conversely, predicting poor quality scores for good-quality images can trigger enhancement mechanisms and increase both energy consumption and latency, leading to reduced engagement, especially on streaming platforms. A biased quality assessment of camera feeds subsequently used by other applications could also significantly affect the reliability of the entire pipeline. The most striking example is video surveillance, where a poor quality prediction could result in storing unclear footage. Finally, an IQA metric that fails to detect medical image artifacts can lead to misdiagnosis and compromise the patient's safety and well-being.

There are different types of attacks, depending on the intentions and motivations of the adversary. They can thus be divided into different categories according to [2]: 1) their nature (poisoning and evasion attacks), and 2) the attacker's objective (targeted and untargeted attacks). Poisoning attacks involve poisoning or altering the data used during the training of the model. In contrast, evasion attacks aim to perturb the used samples during inference, making them mispredicted with high confidence. On the other hand, the main objective of a targeted attack is to produce a specific behavior, while an untargeted attack aims to decrease the performance and accuracy of the model.

In the context of IQA, we find it more interesting to investigate the robustness of current state-of-the-art solutions against evasion attacks. In a first-case scenario, the adversary can carry out a targeted attack by tricking the metric into predicting a specific score that does not reflect the true quality. The other scenario involves launching an untargeted attack to reduce the metric's performance. In other words, the adversary's objective is to make the IQA metric predict a good quality score for poor quality images and vice versa.

Although a few studies [3]–[5] tackled adversarial attacks for IQA models, these works assume the presence of a third party that verifies the integrity of images before their assessment by the IQA metric. Specifically, a human third party has to check whether the images have been changed or not. This assumption is both time-consuming and impractical. It is also quite intuitive that, in this particular context, the visibility of the adversarial perturbation is directly linked to the performance of the attack. Moreover, existing works focus more on launching the attack in a targeted fashion [3], [4], while less attention has been given to untargeted attacks [5].

This work is fully funded by both Région Bretagne (Brittany region), France, and Direction Générale de l'Armement (DGA)

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

In this paper, we explore the robustness of various Convolutional Neural Network (CNN) backbones against adversarial attacks when used in an IQA regression task. Our primary objective is to adapt existing state-of-the-art image classification attacks to IQA models in both targeted and untargeted modes. Then, we aim to examine any possible correlation between the perturbation visibility and the attack effectiveness.

The rest of this paper is organized as follows. Section II provides an overview of Blind Image Quality Assessment (BIQA) metrics, adversarial attacks, and adversarial attacks for BIQA metrics. Next, Section III formulates the problem and describes the proposed framework. Our experiments are detailed and analyzed in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Blind Image Quality Assessment

IQA has been an active research field, and over the years, a multitude of IQA metrics have been proposed. They range from distortion-specific and Natural Scene Statistics (NSS)-based metrics to Machine Learning (ML)-based metrics. However, with the introduction of CNN-based models, the performance gap has considerably widened. Many widely used CNN architectures have been initially proposed for the image classification task. However, a desirable property of deep models enables their fine-tuning on similar tasks, such as IQA [6]. Some of the first works that introduced CNNs to the IQA are [7], [8], and since then, many works have been inspired by this concept [9], [10]. Consequently, recent successful models rely on transfer learning and yield outstanding performance on publicly available IQA datasets [10]–[12]. These solutions mainly differ in the choice of the CNN backbone and the aggregation of their output features.

B. Adversarial Attacks

Ever since [1] discovered the possibility of fooling a deep classification model by carefully crafting an Adversarial Example (AE), the attention of many researchers has shifted toward the security aspect of deep neural networks. In [13], the authors demonstrated how easy it is to manipulate the model to misclassify an input image with high confidence. This work has been followed by a wave of studies, where researchers have demonstrated that the attack can be transferred between two DL architectures [14], and led to a rapid emergence of various competitive adversarial attacks [15]–[17].The adversarial attack is often formulated as an optimization problem. The objective is to minimize or maximize a loss function, i.e., the distance between the predicted and the ground truth or the targeted score, with certain constraints on the crafted AEs, depending on the context. Several popular methods used for image classification include the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [1], Fast Gradient Method (FGM) [15], Carlini and Wagner (C&W) [16], Zeroth Order Optimization (ZOO) [17], and Spatial Transformation attack [18].

C. Adversarial Attacks on BIQA Models

A few works have investigated the impact of these vulnerabilities in the context of IQA [3]–[5]. For instance, in [3], the authors studied a targeted attack to maximize the predicted image quality score and its transferability inter-model. Their main objective was to decrease the QoE by deceiving the quality assessment mechanism into predicting high-quality scores for low-quality images using imperceptible adversarial perturbations. They relied on a gradient-based attack and then used a spatial activity map to reduce its visibility. The reported results showed that increasing the predicted quality score of a given image is possible without improving its actual quality. In the same way, Shumitskaya *et al.* [4] proposed an Universal Perturbation Attack (UAP) to generate a general perturbation for a given BIQA model. They considered the UAP as a non-frozen variable updated during training. Their study demonstrated the existence of a specific gradient of the loss function with respect to the training set images for a given model whose direction leads to an augmentation of the predicted quality score. Then, they proposed to use a Contrast Sensitivity Function (CSF) as a weighting map to reduce the visibility of the crafted perturbation. Zhan *et al.* [5] reformulated the problem in a Lagrangian fashion by swapping the constraint and the objective. Their primary purpose was to maximize the gap between the predicted quality scores on the clean and perturbed images under a constraint of non-visibility of the perturbations. Their main contribution consists in estimating the distance between the clean image and the AE using a Full Reference (FR) metric to mimic the Just Noticeable Difference (JND) aspect of the Human Visual System (HVS). These works revealed the vulnerability of different BIQA metrics, including a few CNN-based architectures. However, no further investigations regarding the robustness of different CNN-based architectures have been conducted. Moreover, most proposed studies focus on achieving non-perceptible perturbations for targeted attacks. To the best of our knowledge, there has been little to no investigation into untargeted attacks on CNN-based IQA models, highlighting the need for further research in this direction.

III. Proposed Framework

In this section, we present a framework that adapts efficient and widely-used adversarial attacks in the field of image classification to the context of the IQA task in both targeted and untargeted scenarios. Then, we evaluate how well the IQA model can withstand these attacks and maintain its ability to predict image quality scores accurately. We intentionally do not set any constraints on the perceptibility of the added perturbation during the generation of the AE. Relaxing the perceptibility constraint allows us to investigate the potential correlation between the success score of the adversarial attack and its perceptibility in the IQA context.

The objective of an adversarial attack is to craft an AE x_{adv} that satisfies $M_w(x_{adv}) = y_{adv}$, where M_w is the IQA model parameterized by the vector of weights w, and y_{adv} is a score that differs from both the ground truth score y_{mos} and the

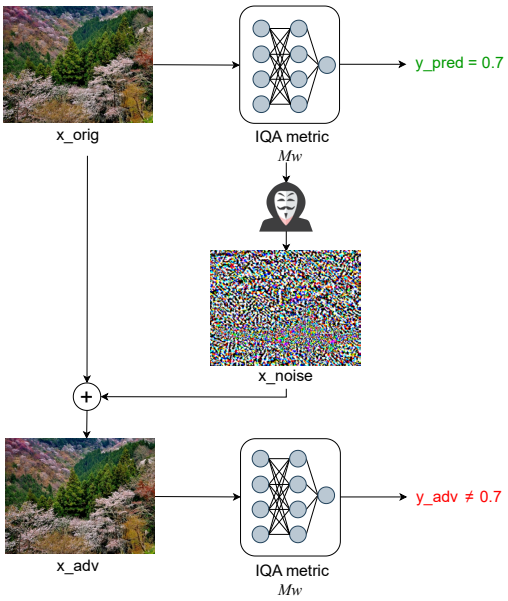


Fig. 1: Overview of the proposed framework.

predicted score on the clean image y_{pred}. The objective of the attack may vary depending on the context and the attacker's goals (availability of ground truth score, targeted or untargeted, etc.).

Our proposed framework is illustrated in Fig. 1, where the adversary uses the model's loss gradients related to the image to generate the AE. Moreover, we consider in this work both attack scenarios: targeted and untargeted.

In the targeted scenario, the adversary's objective is to deviate the model by predicting a high-quality score regardless of the actual image quality. To reach this goal, we generate the AE by adapting the optimization procedure used in previous works [3], [4] into a non-constrained problem. Specifically, we relax the optimization procedure to the following:

$$x_{adv} = \arg\min_{x} \left[y_{target} - M_w(x)\right]^2 \;\; s.t\; x_{adv} \in \mathbb{I}, \quad (1)$$

where $\mathbb{I}$ denotes the image space and y_{target} denotes the maximum achievable quality score by the model. In other words, we are looking for a solution x_{adv} that minimizes the distance between the predicted score $M_w(x_{adv})$ and y_{target}. This particular scenario where y_{target} is set as the maximum achievable quality score does not generalize to the targeted attack's context. The same procedure can be followed to achieve any different target score y_{target}.

In the untargeted scenario, on the other hand, the adversary aims to reduce the accuracy of the IQA model without necessarily deviating it to predict a specific score. For example, it can be achieved by deceiving the model to predict good-quality scores for poor-quality images and inversely. In our attempt to implement this attack, we shift the problem into the following optimization problem:

$$x_{adv} = \arg\max_{x} \left[M_w(x_{orig}) - M_w(x)\right]^2 \;\; s.t\; x_{adv} \in \mathbb{I}, \quad (2)$$

where x_{orig} refers to the original, clean image (before attacking it). This formulation presents a problem during the generation of the AE: a gradient-based optimization approach relies on the gradient of the loss function w.r.t the input image to generate the AE. In an image classification context, the loss is computed by performing the difference between the one hot encoded vector corresponding to the predicted class of x_{orig} and the actual probability vector predicted by the model. Applying it directly to a regression problem will lead to a zero loss function thus canceling out the gradient and the attack. One possible solution could be to consider the Mean Objective Score (MOS) as an equivalent to the label vector. However, this alternative supposes the availability of the MOS during the attack, which is hardly verified. To address this issue, we suppose that the adversary can estimate the MOS by using prior performance information, such as the Root Mean Squared Error (RMSE) of the target model, which is readily available. We give the formula of the RMSE by:

$$RMSE = \sqrt{\mathbb{E}[(mos^{test} - y_{pred}^{test})^2]} \quad (3)$$

where $\mathbb{E}$ denotes the mathematical expectation, mos^{test} refers to the ground truth scores and y_{pred}^{test} are the predicted scores on the test set samples. On the other hand, the standard deviation σ_S formula of a given set of points S having an average value of μ_S, is computed as follows:

$$\sigma_S = \sqrt{\mathbb{E}[(S - \mu_S)^2]} \quad (4)$$

If the set of points S follows a normal distribution $\mathcal{N}(\mu_S, \sigma_S^2)$, then any sample $s \in S$ falls within the range $[\mu_S - 3 \times \sigma_S\,,\, \mu_S + 3 \times \sigma_S]$ and more specifically within $[\mu_S - \sigma_S\,,\, \mu_S + \sigma_S]$ with a very high confidence. Similarly, given y_{pred}, the predicted score of a given sample, we assume that the true value of mos falls within $[y_{pred}(i) - 3 \times RMSE\,,\, y_{pred}(i) + 3 \times RMSE]$ and try to estimate a certain $\hat{mos}$ that is close to mos. Thus, Eq.(2) becomes:

$$x_{adv} = \arg\max_{x} \left[\hat{mos}(x) - M_w(x)\right]^2 \;\; s.t\; x_{adv} \in \mathbb{I} \quad (5)$$

Eq.(5) is derived from Eq.(2) because the attacker cannot access the MOS. Our intuition is based on the similarity between the RMSE and the standard deviation formula as explained earlier. In order to achieve this, we draw 10 samples $\hat{mos}_j \sim \mathcal{N}(y_{pred}, RMSE^2)$, then compute their average to obtain the final $\hat{mos}$.

IV. Experimental Results

A. Experimental Setup

We first selected the most commonly used CNN architectures for IQA, namely ResNet50 [19], VGG16 [20], and InceptionV3 [21]. We assume, however, that similar conclusions can be drawn about other CNNs due to the transferability of attacks across CNN models [14]. Then, we used transfer learning to adapt these pre-trained models to the BIQA task. Transfer learning was achieved by freezing the weights of the CNN backbone as a feature extractor and retraining the Fully Connected (FC) layers on the new IQA task. The training

TABLE I: Initial performance of the CNN backbones on the TID2013 and Koniq-10k datasets.

Model	Dataset	SRCC	PLCC	KRCC	RMSE
ResNet50	Koniq-10k	0.825	0.860	0.635	0.209
VGG16		0.813	0.853	0.622	0.239
InceptionV3		0.707	0.769	0.520	0.289
ResNet50	TID2013	0.942	0.948	0.789	0.332
VGG16		0.930	0.943	0.774	0.347
InceptionV3		0.875	0.911	0.695	0.386

process has been performed using default hyperparameters, specifically, a FC of two hidden layers with a hidden unit parameter of 1024 for each, a learning rate of 10^{-3}, and a batch size $b = 16$. Next, we fine-tuned the models on two widely used IQA datasets, including the TID2013 [11] dataset and the Koniq-10k [10] dataset. The first contains synthetic distortions, while the second includes images with authentic distortions. Initial performance obtained from the three considered CNN backbones on both datasets, i.e., without any attack, are reported in Table I. It is important to note that our main objective is not to get the best possible performance from the model but rather to get an efficient model that neither over-fits nor under-fits the training data.

We considered in our experiments three widely used adversarial attacks, namely Fast Gradient Method (FGM) [15], Basic Iterative Method (BIM) [22] and Projected Gradient Descent (PGD) [23] with varying values of the ϵ parameter ($\epsilon \in \{0.001, 0.01, 0.1, 1\}$) and $n_{iter} = 10$ for iterative attacks. We believe that the choice of the attack is not very crucial, as the main objective of this work is to adapt the concept of adversarial attacks to the IQA task in both targeted and untargeted settings. We used the code provided by the CleverHans software library [24] to which we added further modifications and adjustments in order to make it work in the context of our study. We compared different attacked CNN backbones to the performance of the initial model, i.e., without attack, in terms of Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (SRCC), Kendall's Rank Correlation Coefficient (KRCC), and Root Mean Squared Error (RMSE). The objective of the attack is to deviate from the model's predictions, which can be observed by the reduction of the correlation scores and the augmentation of the RMSE coefficient. In a targeted attack scenario, the objective is to achieve a correlation score closer to zero. While a lower correlation score indicates a better performance of the attack in the untargeted scenario.

We also compared the different scenarios in terms of image similarity, between x_{orig} and x_{adv}. This gives us an idea about the degree of visibility of the perturbation. Since the objective of this study also includes an investigation of the potential correlation between the visibility of added perturbations and the success of the attack. We refer for this purpose to the FR metric Learned Perceptual Image Patch Similarity (LPIPS) [6] to compute the similarity score between x_{adv} and x_{orig}. LPIPS has been shown to be effective in capturing complex

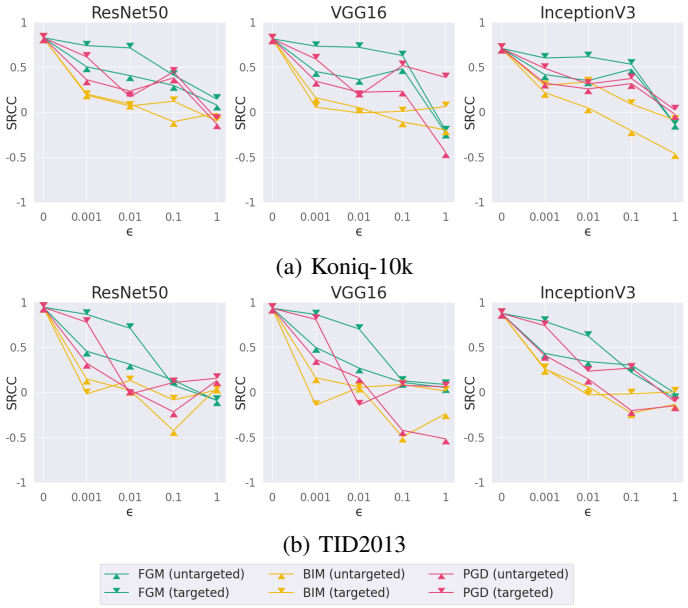


Fig. 2: SRCC on the test set with respect to the variation of ϵ and the launched attack.

and subtle differences between images, making it thus more correlated with human perceptual judgments. A higher value of LPIPS score indicates a higher visibility of the adversarial perturbation compared to the original image.

B. Results and Analysis

Tables II and III provide a summary of the reported performance of the IQA models in the targeted and untargeted scenarios, respectively. Each table represents a specific attack mode, and reports the results of the three attacks on the considered CNN architectures with variable ϵ magnitudes. In the following, we refer to the success score of an attack as the difference between the model's initial performance and its performance when subjected to the attack. To better help visualize the reported results, Fig. 2 displays the variation of the SRCC w.r.t ϵ and the launched attack. We refer to the initial performance of the IQA models, i.e., on clean images, by $\epsilon = 0$, and we provide the following equation in order to clarify the meaning of ϵ:

$$x_{adv} = x_{orig} + \epsilon \times x_{noise} \tag{6}$$

where x_{noise} refers to generated adversarial perturbation.

Tables II and III illustrate that the three tested CNN backbones show the same vulnerability to the considered attacks regardless of the architecture. This vulnerability is evident in the drastic drop in correlation scores as illustrated in Fig. 2. Furthermore, it indicates that none of the characteristics proper to any of the tested architectures represent a robustness aspect against adversarial attacks.

Tables II, III and Fig. 2 also indicate that iterative attacks, such as BIM and PGD, achieve higher success scores compared to the FGM attack and lower LPIPS. This finding is consistent with the view of the attack as an optimization

TABLE II: Performance scores obtained on the attacked CNN backbones with the variation of ϵ in a targeted mode.

Attack	Dataset	CNN	ϵ = 0.001					ϵ = 0.01					ϵ = 0.1					ϵ = 1				
			SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
FGM	Koniq-10k	ResNet50	0.737	0.799	0.549	0.199	**0.000**	0.712	0.771	0.521	0.256	0.061	0.409	0.424	0.283	0.374	0.771	0.142	0.148	0.093	0.175	**1.328**
		VGG16	0.731	0.767	0.538	0.187	**0.000**	0.718	0.730	0.523	0.269	0.061	0.627	0.644	0.445	0.321	0.739	-0.211	0.185	-0.143	0.074	**1.341**
		InceptionV3	0.600	0.686	0.427	0.247	**0.000**	0.613	0.696	0.437	0.250	0.057	0.530	0.553	0.371	**0.454**	0.709	-0.137	0.123	-0.091	0.198	**1.180**
	TID2013	ResNet50	0.863	0.878	0.677	0.814	**0.313**	0.709	0.764	0.526	1.423	0.355	**0.075**	0.248	**0.044**	0.454	0.849	-0.091	0.290	-0.057	1.455	1.256
		VGG16	0.862	0.884	0.675	0.731	**0.313**	0.701	0.776	0.515	1.448	0.351	0.124	0.231	0.075	0.634	0.829	0.084	0.250	0.051	0.480	**1.278**
		InceptionV3	0.785	0.865	0.597	0.510	0.314	0.621	0.760	0.449	0.682	0.355	0.217	0.273	0.146	0.416	0.812	**-0.070**	0.095	-0.047	0.101	**1.138**
BIM	Koniq-10k	ResNet50	**0.186**	**0.286**	**0.125**	**0.325**	0.006	**0.066**	**0.083**	**0.044**	**0.540**	0.294	**0.119**	**0.147**	**0.078**	**0.844**	1.178	-0.100	**0.100**	-0.066	**1.127**	1.332
		VGG16	**0.052**	**-0.021**	**0.035**	**0.428**	0.005	**-0.011**	**0.030**	**-0.008**	**1.055**	0.290	**0.005**	**0.022**	**0.003**	**1.152**	1.234	**0.059**	**0.076**	**0.039**	**1.033**	1.376
		InceptionV3	**0.296**	**0.447**	**0.200**	**0.391**	0.006	0.333	**0.361**	0.225	0.372	0.275	**0.086**	**0.091**	**0.057**	0.203	1.184	-0.091	0.077	-0.060	0.178	1.308
	TID2013	ResNet50	**-0.020**	**0.413**	**-0.016**	**1.360**	0.320	0.129	**0.131**	0.086	**2.993**	0.494	-0.088	0.118	-0.061	**3.603**	1.160	**0.022**	**0.134**	**0.015**	**13.866**	**1.212**
		VGG16	**-0.146**	**0.189**	**-0.094**	**1.232**	0.319	**0.051**	**0.081**	**0.033**	**3.908**	0.486	0.083	0.118	0.056	**5.444**	1.195	**0.012**	**0.060**	**0.007**	**25.786**	1.312
		InceptionV3	**0.262**	**0.489**	**0.180**	**1.182**	0.320	**-0.032**	**0.079**	**-0.020**	**2.047**	0.484	**-0.020**	**0.071**	**-0.012**	1.108	1.117	0.002	**0.041**	**0.002**	**2.385**	1.258
PGD	Koniq-10k	ResNet50	0.612	0.662	0.440	0.200	**0.000**	0.167	0.312	0.112	0.344	**0.009**	0.444	0.534	0.307	0.693	**0.365**	**-0.081**	0.116	**-0.053**	0.371	1.378
		VGG16	0.589	0.631	0.417	0.177	**0.000**	0.184	0.220	0.124	0.281	**0.010**	0.516	0.533	0.358	0.271	**0.365**	0.384	0.363	0.262	0.108	1.377
		InceptionV3	0.485	0.604	0.338	0.237	**0.000**	**0.311**	0.481	**0.211**	**0.377**	**0.010**	0.371	0.423	0.253	0.464	**0.366**	0.022	0.034	0.015	0.487	1.378
	TID2013	ResNet50	0.776	0.791	0.577	1.037	**0.313**	**-0.023**	0.180	**-0.017**	1.297	**0.325**	0.106	**0.164**	0.072	1.021	**0.542**	0.153	0.222	0.102	0.466	1.283
		VGG16	0.807	0.824	0.608	0.924	**0.313**	-0.145	0.212	-0.089	1.245	**0.325**	**0.074**	**0.096**	**0.053**	0.777	**0.542**	0.058	0.123	0.039	0.571	1.283
		InceptionV3	0.738	0.806	0.549	0.546	**0.313**	0.232	0.489	0.161	1.128	**0.326**	0.266	0.426	0.181	**1.330**	**0.543**	-0.108	0.150	-0.075	1.481	1.283

TABLE III: Performance scores obtained on the attacked CNN backbones with the variation of ϵ in an untargeted mode.

Attack	Dataset	CNN	ϵ = 0.001					ϵ = 0.01					ϵ = 0.1					ϵ = 1				
			SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
FGM	Koniq-10k	ResNet50	0.501	0.556	0.354	0.494	**0.000**	0.404	0.409	0.285	0.658	0.063	0.294	0.315	0.202	0.402	0.774	0.076	0.086	0.050	0.173	1.327
		VGG16	0.449	0.494	0.321	0.639	**0.000**	0.361	0.329	0.257	0.818	0.064	0.479	0.496	0.329	0.385	0.749	-0.236	0.201	-0.159	0.074	1.345
		InceptionV3	0.415	0.462	0.288	0.597	**0.000**	0.348	0.383	0.239	0.632	0.058	0.474	0.480	0.328	0.472	0.710	-0.138	0.131	-0.091	0.198	**1.180**
	TID2013	ResNet50	0.453	0.424	0.339	2.071	**0.000**	0.308	0.282	0.228	3.050	0.044	0.130	0.226	0.085	0.454	0.719	-0.096	0.292	-0.060	1.429	1.295
		VGG16	0.496	0.462	0.372	2.085	**0.000**	0.266	0.264	0.194	3.526	0.040	0.107	0.126	0.068	0.714	0.692	0.049	0.224	0.027	0.475	1.317
		InceptionV3	0.430	0.428	0.312	1.432	**0.000**	0.338	0.299	0.238	1.640	0.042	0.299	0.335	0.204	0.426	0.687	-0.012	0.043	-0.008	0.103	**1.161**
BIM	Koniq-10k	ResNet50	**0.198**	**0.142**	**0.137**	**2.031**	0.005	**0.088**	**0.063**	**0.061**	**2.617**	0.269	**-0.109**	**0.214**	**-0.082**	**1.814**	1.109	-0.015	**0.282**	-0.008	**6.751**	**1.303**
		VGG16	**0.157**	**0.080**	**0.110**	**2.722**	0.006	**0.047**	**0.070**	**0.035**	**6.042**	0.255	**-0.112**	**0.162**	**-0.078**	**15.413**	1.202	-0.201	**0.288**	-0.136	**22.329**	**1.294**
		InceptionV3	**0.215**	**0.203**	**0.145**	**1.734**	0.006	**0.044**	**-0.014**	**0.031**	**1.835**	0.273	**-0.211**	0.333	**-0.141**	0.959	1.184	**-0.463**	0.474	**-0.313**	0.374	1.307
	TID2013	ResNet50	**0.144**	**0.118**	**0.103**	**10.539**	0.005	**0.019**	**0.073**	**0.015**	**15.815**	0.198	**-0.429**	**0.553**	**-0.337**	**7.335**	1.156	**0.046**	**0.044**	**0.031**	**22.722**	**1.280**
		VGG16	**0.158**	**0.182**	**0.119**	**11.739**	0.004	**0.057**	**0.094**	**0.042**	**35.436**	0.160	**-0.499**	**0.484**	**-0.379**	**46.853**	1.109	-0.240	**0.335**	-0.164	**173.622**	**1.268**
		InceptionV3	**0.253**	**0.161**	**0.178**	**5.235**	0.004	**0.057**	0.119	**0.041**	**6.507**	0.203	**-0.233**	0.501	**-0.159**	1.703	1.168	-0.138	**0.534**	-0.093	**1.409**	1.269
PGD	Koniq-10k	ResNet50	0.357	0.399	0.258	0.699	**0.000**	0.230	0.172	0.158	1.928	**0.010**	0.379	0.285	0.273	1.027	**0.368**	**-0.135**	0.414	**-0.090**	0.519	1.377
		VGG16	0.341	0.346	0.245	0.910	**0.000**	0.219	0.121	0.155	2.537	**0.010**	0.229	0.117	0.179	0.839	**0.368**	**-0.453**	0.604	**-0.292**	0.725	1.377
		InceptionV3	0.319	0.339	0.223	0.804	**0.000**	0.256	0.223	0.175	1.704	**0.010**	0.314	**0.195**	0.217	**1.130**	**0.367**	-0.026	**0.348**	-0.017	**0.721**	1.378
	TID2013	ResNet50	0.320	0.291	0.235	2.682	**0.000**	0.023	0.108	0.018	9.619	**0.009**	-0.220	0.606	-0.138	5.158	**0.279**	0.127	0.209	0.087	1.175	1.342
		VGG16	0.362	0.329	0.270	2.594	**0.000**	0.157	0.133	0.115	10.673	**0.008**	-0.426	0.587	-0.314	5.624	**0.279**	**-0.521**	0.774	**-0.338**	1.579	1.343
		InceptionV3	0.408	0.362	0.298	1.823	**0.000**	0.145	**0.059**	0.106	4.912	**0.008**	-0.206	**0.371**	-0.143	**3.116**	**0.279**	**-0.153**	0.687	**-0.113**	0.738	1.342

Fig. 3: Attack's success scores w.r.t LPIPS.

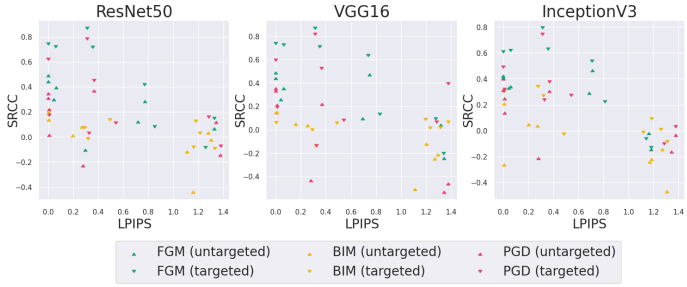


Fig. 4: SRCC scores w.r.t LPIPS.

problem, where iterative methods can achieve better results than a single-step method like the FGM attack. Moreover, the results also exhibit a trade-off between the attack success (drop in correlation scores) and perturbation visibility (LPIPS score). The best-performing attack in terms of success score (whose correlation scores are displayed in bold in Tables II and III) may not necessarily achieve the lowest LPIPS scores and vice versa. Overall, the attack's magnitude ϵ and LPIPS score are inversely proportional. Figures 3 and 4 illustrate the relationship between LPIPS score and, respectively, the attack success score and the SRCC. Surprisingly, there appears to be a low correlation between the LPIPS score and the resulting SRCC and success score. This is highlighted in Figs. 3 and 4 by scattered plotted points that do not show a clear pattern.

This correlation highlights the importance of carefully balancing the attack magnitude and visibility in order to achieve optimal attack performance. Nevertheless, while an increase in the LPIPS score may generally indicate better attack performance, it is not always the case. According to Fig. 2, an increase in ϵ can have the opposite effect, BIM's correlation score for example increases for ϵ = 1e-2 on VGG16 and Resnet50 in the TID2013 dataset when compared to $\epsilon = 1e{-}3$. This finding calls into question our initial hypothesis that attack effectiveness is highly correlated with the magnitude and thus the visibility of the attack. Furthermore, we observe that untargeted attacks have a tendency to converge faster than targeted attacks. This can be justified by the fact that targeted attacks aim to achieve a specific goal, making them more difficult to optimize.

In Fig. 5, we illustrate PGD attacked versions of a few samples with varying values of ϵ. We note that quality scores

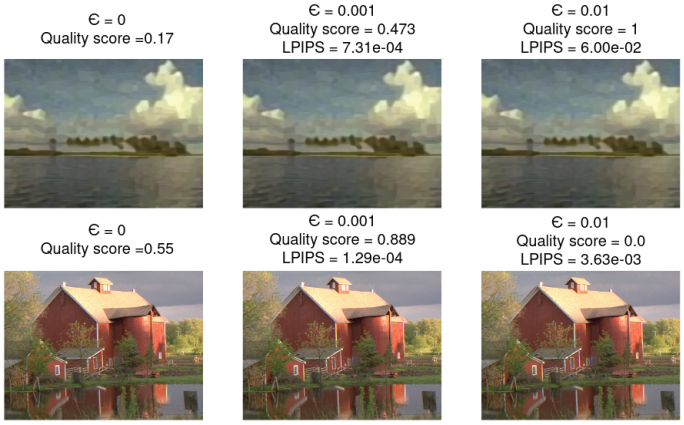


Fig. 5: Visual comparison of original images with their attacked versions under variable ϵ values.

are normalized for an easier comparison between the datasets. It is noteworthy that the attack achieves optimum results from $\epsilon = 0.01$, but perturbations start slightly appearing in low texture regions from this specific value. This suggests that ϵ values in the range of 0.01 establish a favorable balance between perturbation visibility and attack performance. Upon further visualizing the resulting samples, we also noticed that in some cases, targeted attacks fail to maximize the quality scores and instead have the opposite effect. Conversely, untargeted attacks can outperform targeted attacks on poor-quality images and achieve higher predicted quality score.

V. Conclusion

In this work, we investigated the vulnerability of deep CNN-based BIQA metrics to adversarial attacks. We summarize our key contributions in what follows: Firstly, we proposed a framework to generate AEs by adapting widely used image classification attacks to the IQA regression context. Secondly, we introduced the untargeted mode in the context of BIQA and showed that it can outperform targeted attacks in certain scenarios. Thirdly, we investigated the correlation between the attack magnitude and attack success. Then concluded that although a trade-off between the visibility of the perturbation and the effectiveness of the attack exists, higher attack magnitude does not necessarily imply higher attack success scores. Finally and most importantly, we demonstrated the widespread vulnerability of CNNs used in BIQA context, which opens new challenges to develop defense techniques and highlights the importance of ensuring the robustness of BIQA models.

Moving forward, this work also leaves room for further perspectives. For instance, the investigation of the robustness of vision transformers-based IQA metrics against adversarial attacks is much desired. Moreover, the development of defense mechanisms to secure the IQA models is a promising avenue of paramount importance, as many systems rely on their performance.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] N. Pitropakis *et al.*, "A taxonomy and survey of attacks against machine learning," *Computer Science Review*, vol. 34, p. 100199, 2019.

[3] J. Korhonen and J. You, "Adversarial attacks against blind image quality assessment models," in *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 3–11.

[4] E. Shumitskaya, A. Antsiferova, and D. Vatolin, "Universal perturbation attack on differentiable no-reference image-and video-quality metrics," *arXiv preprint arXiv:2211.00366*, 2022.

[5] W. Zhang, D. Li, X. Min, G. Zhai, G. Guo, X. Yang, and K. Ma, "Perceptual attacks of no-reference image quality models with human-in-the-loop," *arXiv preprint arXiv:2210.00933*, 2022.

[6] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[7] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2014.

[8] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3773–3777.

[9] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, pp. 355–362, 2018.

[10] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[11] N. Ponomarenko *et al.*, "Color image database tid2013: Peculiarities and preliminary results," in *IEEE European workshop on visual information processing (EUVIP)*. IEEE, 2013, pp. 106–111.

[12] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[13] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[14] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[16] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[17] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[18] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International conference on machine learning*. PMLR, 2019, pp. 1802–1811.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learningfor image recognition," *CoRR, abs/1512*, vol. 3385, p. 2, 2015.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision. corr abs/1512.00567 (2015)," 2015.

[22] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[24] N. Papernot *et al.*, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.

Blind Video Stabilization Quality Assessment based on convolutional LSTM

Mohamed Riad Yagoubi[1], Seyed Ali Amirshahi[1], Steven Le Moan[1], Azeddine Beghdadi[1,2], Erik Rodner[3]
[1] *Norwegian University of Science and Technology, Gjøvik, Norway*
[2] *Institut Galilee, University Sorbonne Paris Nord, Paris, France*
[3] *University of Applied Sciences, Berlin, Germany*
{mohamed.r.yagoubi,s.ali.amirshahi, steven.lemoan}@ntnu.no, beghdadi@sorbonne-paris-nord.fr, erik.rodner@htw-berlin.de

***Abstract*—Assessing the performance of video stabilization algorithms is still in its infancy compared with the state-of-the-art in video quality assessment, where the focus is on classic distortions such as noise, blur, lighting problems, and coding artifacts. In this study, we present a new blind video stabilization quality assessment metric based on learning sensor outputs via a novel deep neural network. We build an architecture based on a convolutional LSTM to estimate accelerations given by an inertial measurement unit sensor with no additional information than video frames. The experimental study was carried out on a new dedicated dataset, which showed the effectiveness of our metric in characterizing the level of shakiness of videos, in addition, we demonstrate the efficiency of our system to learn acceleration data provided from sensors, which opens ambitious perspectives in retrieving sensors data from any video.**

***Index Terms*—VSQA, Video Quality, Stabilization, LSTM, ConvLSTM2D, IMU, Accelerometer**

I. Introduction

A significant number of videos are produced every day as a result of the pervasive usage of mobile video recording equipment. However, such videos frequently experience video instability or shakiness. This distortion is often responsible for various artifacts and degradations, such as loss of resolution, blur, and geometric distortions, which inevitably affect video quality and cause visual fatigue and discomfort.

These spatio-temporal distortions also make high-level tasks such as object recognition and visual tracking difficult. This is mainly due to the speed observers process the visual stimuli, particularly in the presence of visual discomfort and highly variable visual content over time [1]. It should be noted that such limitations also exist in the field of computer vision, but not for the same reasons, which are purely psychovisual and physiological in the case of the brain. For example, the performance of algorithms for object detection and identification, visual tracking, segmentation, and coding is affected by video shakiness. It is therefore important to have metrics that can measure the level of instability in videos allowing us to assess the video quality and develop techniques for correcting such instabilities in real time or in off-line video analysis mode.

In recent years, numerous effective and economical techniques, known as Digital Video Stabilization (DVS), have been developed to overcome frame-to-frame shakiness [2]. However, so far, there is still a need to develop a reliable objective metric to quantify the performance of these increasing numbers of DVS approaches [3], [4]. Since this field remains far less explored, in this paper, we propose a no-reference video quality assessment metric dedicated to Video Stabilization Quality Assessment (VSQA).

VSQA is the process of evaluating the performance of DVS algorithms in terms of perceptual quality [5]–[7]. It should be noted that the evaluation of the perceptual quality of video stabilization results can be carried out using the same protocols and conventional approaches used for Image and Video Quality Assessment (IQA and VQA). Three categories can be distinguished according to the availability of the original version (stabilized video), reduced information, or the absence of stabilized video. The Full Reference (FR) VSQA consists of estimating the difference between the original video and its stabilized version. However, for the case of Reduced Reference (RR) a reduced amount of information is available in the case of the original video. The No-Reference (NR) approach focuses on estimating the level of shakiness without using any prior information about the original version of the video.

The main challenge in designing a FR VSQA is the availability of a physically stable version to be used as ground-truth data. Most NR VSQA methods for assessing the quality of visual content require a priori knowledge of the original version or are based on distortion models to measure their level of severity and their impact on the observed video signal, making them RR and not NR methods.

The aim of this study is to develop a stabilization metric on which DVS approaches could rely through their benchmark study. The main contributions of this study are:

- We develop the first blind VSQA approach based on learning IMU (Inertial Measurement Unit) sensor data via a Recurrent Neural Network (RNN). Our system can provide an estimation of the accelerometer data between video frames with no additional information. We develop an algorithm based on the sensor data estimation to quantify video shakiness and provide a score for the stabilization quality.
- A new database has been created by adding several levels of shakiness to a common database previously used to assess DVS techniques [7]. This provides the scientific community with a complete and realistic database for assessing not only VSQA metrics, but also for testing video stabilization algorithms.

- The proposed approach is compared with the most common VSQA approaches and efficient video quality metrics that consider temporal distortions.

This paper is organized as follows. Section II is dedicated to a brief overview of the VSQA methods proposed in the field. We then discuss in detail the proposed approach in Section III followed by the experimental results in Section IV. Finally, a conclusion of the work is given in Section V.

II. Related work

Although different video stabilization methods have been proposed over the years [2]–[4], the benchmark is usually done through subjective evaluations due to the lack of reliable objective evaluation metrics. One of the most important techniques was proposed by Zhang et al. [7] which is based on computing the geodesic distance between the motion paths of the stable and stabilized videos via the Riemannian metric defined on the manifold of spatial transformations. Keeping in mind that in most cases the stable video is not on hand, this FR technique is often difficult to apply in most real-life applications. A NR stabilization assessment technique has been introduced by the same group [6] using the geodesic curvature to estimate their metric. The same database from their previous study has been used for the experimentation; however, the lake of a public implementation makes it hard for DVS approaches to rely on this metric for a benchmark study.

In the literature, the widely used NR metrics in VSQA are the inter-frame PSNR-based/SSIM-based metrics. Despite its simplicity, the Interframe Transformation Fidelity (ITF) [8] seems to be the most common method in this field of work. ITF is based on the video inter-frame calculation expressed by the average inter-frame PSNR on the whole video. Interframe Similarity Index (ISI) is another widely used score in DVS benchmark studies. Similar to ITF, ISI is an interframe metric that, instead of PSNR, is based on the well-known Structural Similarity Index (SSIM) [5]. By definition, ISI simply measures the average of the inter-frame SSIM through video frames.

Finally, we should point out that spatio-temporal video quality metrics such as VSFA [9] or VIIDEO [10] could be used for assessing instability (shakiness) in videos. However, since these metrics are initially designed to deal with several types of distortions, their performance against video instability are very limited.

III. Proposed approach

Video shakiness is mostly due to random micro-motions of the camera during the recording process. This phenomenon is linked to various parameters, including the trepidation of the hand or environmental aspects such as recording on a moving vehicle. Shakiness is one of the rare distortions that uniformly affects all pixels of a frame [2]. In this study, we develop the first NR VSQA approach based on learning IMU data via an LSTM network. The estimated accelerometer data from the IMU measurements which is fed into the network enable the camera shake levels to be quantified objectively and reliably through the learning process. A perceptual quality score, which is to be analyzed in relation to the subjective appreciation of the observer, can thus be calculated.

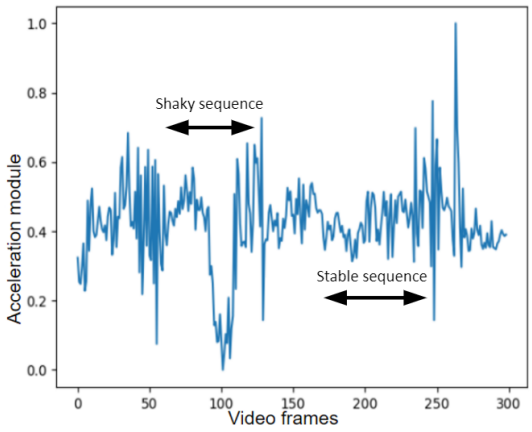


Fig. 1: Accelerometer data of a video with both shaky and stable sequences.

The idea behind relying on acceleration is based on the nature of the shakiness distortion. Indeed, video shaking is generally the result of random oscillations related to unintentional and uncontrollable movements of various magnitudes and orientations. In certain specific situations, they can occur according to a pseudo-periodic process; this is, for example, the case when shooting from a car moving over a more or less flat terrain. One of the major difficulties is then finding a reliable solution that allows one to differentiate between these unintentional camera movements and those of the various moving objects in the scene. The distribution of some motion features, such as velocity and acceleration, of points of interest could help in designing a robust camera instability estimation. Using IMU acceleration measurements is a plausible solution to estimate the level of shakiness in the video. Figure 1 illustrates the acceleration profile of a 10-second video sequence.

A. IMU sensor Accelerometer data for shaking estimation

Lets denote $F = \left[f_{t_0}^0, f_{t_1}^1, \dots f_{t_i}^i \dots\right]$ vector containing all frames of the video, while $f_{t_i}^i$ is the i^{th} frame of size $M \times N$ taken at the timestamp t_i. Furthermore, we denote Acc_x^i, Acc_y^i, and Acc_z^i corresponding to the Cartesian acceleration vector of the frame i from the precedent frame $i-1$.

Since the main goal is to take advantage of the acceleration data to detect vibrations (shakiness) in the video frames, we propose to model vibrations of the camera by a simple acceleration vector module. Despite the fact that space information of the acceleration are lost after such simplification, the benefit consists of reducing the output complexity on one-dimensional data instead of three. This drastically improves the performance of our learning process. Eq. (1) provides information on how to estimate the acceleration module a_i through $\|.\|_2$.

$$a_i = \| \left[Acc_x^i, Acc_y^i, Acc_z^i\right] \|_2 \tag{1}$$

Now, let $A = \{a_{\tau_0}, a_{\tau_1}, ..., a_{\tau_j}...\}$ where a_{τ_j} corresponds to the acceleration given by the accelerometer sensor exactly at the timestamp τ_j. In real-world applications, sensor timestamps τ and frame timestamps t are very rarely synchronized. For that, we propose the following equalization process expressed as:

$$p = \arg\min_j |t_i - \tau_j| \tag{2}$$

$$\hat{a}_{t_i} = a_{\tau_p} \quad . \tag{3}$$

While $\hat{a}_{t_i}$ is the estimation of the acceleration sensor at the exact timestamp t_i. We then denote $\hat{A} = [\hat{a}_{t_0}, \hat{a}_{t_1}, ...\hat{a}_{t_i} \ldots]$ the estimated acceleration vector synchronized with the frame vector timestamps.

The aim now is to design a system that can estimate the acceleration feature from frames changing velocity. However, due to the non-linearity of the data and the complexity of retrieving accelerations between frames, tackling such a problem via a deep learning network so far seems to be the most appropriate option. The idea is to inject the frame sequence vector F as input and fix the acceleration vector $\hat{A}$ in the output to develop a system that can estimate the acceleration without the information given by the IMU. Since the network should have the ability to estimate velocity changes between frames, RNN would be the ideal candidate due to its ability to establish connections between nodes to create a cycle, allowing output from some nodes to affect subsequent input to the same nodes. This special feature allows RNN to exhibit a temporal dynamic behavior. For this purpose, in this study, we select the Conv2DLSTM Recurrent network.

B. Convolutional Long Short Term Memory Network (ConvLSTM2D) - data ingestion

Inspired by the success of CNN and RNN, Conv2D-LSTM network was proposed in [11], which mainly consists of 2-Dimensional Convolutional Neural Networks (Conv2D) instead of 1D input fed into LSTM network. Convolutional layers take advantage of grasping the feature spatially, and an LSTM system is, so far, one of the best at analyzing complex time sequences.

The convolutional layers are able to learn the relevant features from an image at different levels. In addition, LSTM is able to bridge long time lags between inputs over arbitrary time intervals. In our case, the velocity change detected by the sensor is computed between two consecutive timestamps. Therefore, the time interval is obviously fixed at two, which means that we inject and learn two consecutive frames ($f^i_{t_i}$ and $f^{i+1}_{t_{i+1}}$) synchronized with the acceleration output $\hat{a}_{t_{i+1}}$ of the $i+1^th$ frame. For that, the output acceleration vector $\hat{A}$ is sampled by two, then, to simplify our annotations, we propose to multiply the indexes of F by two instead of using the dividing indexes in $\hat{A}$. Our system is then summarized in Eq. 4. Figure 2 shows the architecture of the proposed ConvLSTM2D. We note that ft, Ot, it and C_t are well-known LSTM parameters, while their respective expressions could be found in [11].

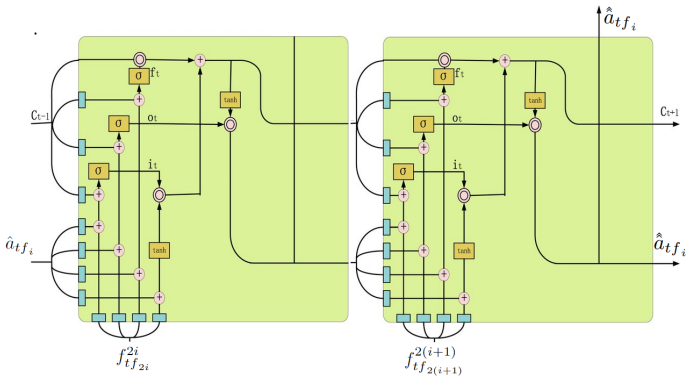


Fig. 2: Proposed architecture for ConvLSTM2D.

$$\begin{aligned} \text{input:} F &= \left[f^0_{t_0}, f^1_{t_1}, ...f^{2i}_{t_{2i}}, f^{2i+1}_{t_{2i+1}}....\right] \\ \text{output:} \hat{A} &= [\hat{a}_{t_0}, \hat{a}_{t_1}, ...\hat{a}_{t_i}....] \\ \text{predicted output:} \hat{\hat{A}} &= [\hat{a}_{t_0}, \hat{\hat{a}}_{t_1}, ...\hat{\hat{a}}_{t_i}....] \\ card(F) &= 2 \cdot card(\hat{A}) \end{aligned} \tag{4}$$

C. Proposed architecture

We propose to construct a architecture around three ConvLSTM2D layers. We inject our input into the first ConvLSTM2D layer with 10 filters, then a batch normalization is performed to provide resistance to the vanishing gradient during training, which positively impacts the convergence [12]. A max-pooling layer is performed thereafter, despite the main reason of the pooling which consists in reducing the dimensionality,it is well known that max-pooling layers often improve generalization and provide resistance to micro-distortions [13]. These three layers (ConvLSTM2D, BatchNormalization and Max-pooling) are performed sequentially three times.

We then perform a 3D Convolutions layer with 10 filters in order to regroup all features extracted in one cube. This cube is after that flattened into a 1D vector, which is sent into a Dropout layer with 0.6 dropout probability. The main reason for putting a Dropout layer is its capability to drop nodes randomly, which prevents overfitting of our architecture. The activation function of the model is Sigmoid to provide smooth gradient, thereby, preventing jumps in output values which can be close from acceleration features that our network has to learn.

Finally, despite the robustness of the proposed architecture as shown in the experiment, its simplicity opens a new perspective to perform our metric in real-time and provide a real-time shakiness quality measure of the captured video. Figure 3 summarizes the architecture constructed to learn IMU data from video frames.

D. Shakiness metric

The aim of our study is to develop a metric that can be correlated with the uncomfortable shaky sequences in the video.

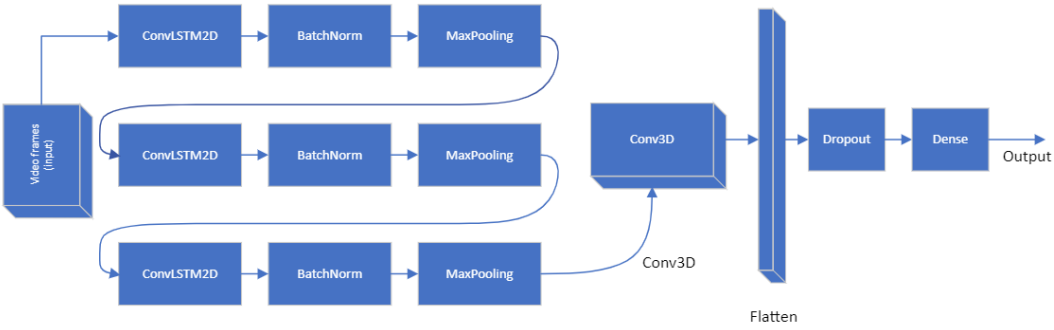


Fig. 3: Proposed IMU estimation architecture

The main challenge is to differentiate vibrations from scene changing or usual device motions. Shakiness is characterized by many motions, with random amplitudes and orientations. We propose to characterize these special types of fluctuations by the standard deviation of acceleration derivative. A pooling step seems to be unavoidable in order to express a global score for the whole or a sequence of the video. We denote the length of the pooling window by L. Our ShaKiness Metric (SKM) is expressed as follows:

$$\text{SKM} = \sqrt{\frac{\sum_0^{L-1} (\nabla(\hat{a})_{t_i} - \overline{\nabla(\hat{a})_t})^2}{L}} \tag{5}$$

While $\nabla(\hat{a})$ is the acceleration signature gradient. Furthermore, if the aim is to provide a global quality score for a video, the window L is set to the length of the whole video.

IV. Experimental Study

In the current section, we introduce our collected dataset for training and validating the proposed approach. Our experimentation starts by demonstrating the efficiency of the proposed architecture in learning accelerometer sensors data from video frames; thereafter, we compare our metric (no-reference) with, the most common shakiness metrics in another dataset rather than the training data.

All our experiments were performed on an Intel(R) Core(TM) i7-8086K CPU @ 4.00GHz 12 Cores, 32Gb of RAM, with the GPU: Nvidia GeForce RTX 2080Ti card. Our training dataset has been constructed with a SAMSUNG Galaxy Note S10 device containing IMU (Initial Measurement Units) sensors. We notice that IMU is used only for accelerometer data.

A. Dataset

1) Training dataset: Our dataset is constructed from five two-minute videos recorded at 30 frames/seconds. The collection process consists of recording videos of moving objects intermittently with and without shakiness, meanwhile, the IMU sensor is collecting the accelerometer data. After the collection process, as stated in Section III-A a time equalization step should be performed to synchronize frames and accelerometer data timestamps resulting in a total of 18000 frames (five videos $\times$ two minutes $\times$ 60 seconds $\times$ 30 frames) in our dataset. It is important to note that the collected dataset is only used for training and validating the convergence of our deep learning model. The proposed stabilization metric is compared with the state-of-the-art in a new dataset inspired by [7] and completely different from the training dataset.

2) Testing and validation datasets: The main challenge in assessing DVS approaches is the extreme difficulty of creating ground-truth data. Wang et al. [7] have introduced a dataset which consists of pairwise stable/shaky videos, out of which each shaky video is assigned with an ideally stable video that has the same content as the reference. This has been done by attaching two cameras, Dji Osmo (for stable output) and GoPro (for shaky outputs), to the same rod while collecting data simultaneously. Subsequently, a data refinement process [14] is then performed to superpose shaky/stable outputs. To encompass a variety of shaky motions and scene complexities, authors performed a collection process in nine scenarios while each of them contains five short shaky videos and their stable versions: eight types $\times$ five videos $\times$ two variants (shaky/stable): walking, climbing, running, riding, driving, crowd, near-range object and dark.

One proposes to create several nuances of shakiness by generating different scales of frame vibrations artificially. To do so, we start by selecting an empty Sub-Frame sf^i of size $(M - b) \times (N - b)$ centered inside a frame f^i. The sub-frame is then slid and moved vertically by $\delta x \in [-\frac{b}{2}, \frac{b}{2}]$ and horizontally by $\delta y \in [-\frac{b}{2}, \frac{b}{2}]$ on the border frame, per frame. Due to the uniform nature of shakiness [2], the random nature of δx and δy following a Gaussian distribution perfectly simulates shakiness within the video; and, due to the standard deviation of the Gaussian distribution, the aggressiveness of the shakiness is calibrated. The entire data generation process is expressed in algorithm 1 and figure 4 illustrates the shakiness generation process to construct our database.

B. Results

1) IMU sensor data - learning performance: We demonstrate in this section the capacity of our architecture to learn data from accelerometer sensors from video frames. Our dataset contains (18000 frames) on which the estimation of the accelerometer sensor is associated to each frame. The data

$\Sigma = [\sigma_1, \sigma_2, ..., \sigma_n]$;
for *each video in dataset* **do**
  $L \leftarrow$ **number of frames**;
  for *each* σ_i *in* Σ **do**
    $X \leftarrow$ RandomGaussianVector$(size = L, min = 0, max = b, std = \sigma_i)$;
    $Y \leftarrow$ RandomGaussianVector$(size = L, min = 0, max = b, std = \sigma_i)$;
    for *each frame* $i \leq L$ **do**
      $\delta_x = X[i]$;
      $\delta_y = Y[i]$;
      $\text{sf}^i = f^i[\delta_x : \delta_x + M, \delta_y : \delta_y + N]$;
    end
    Write the new video from $sf[...]$
  end
end

Algorithm 1: Shakiness generation algorithm

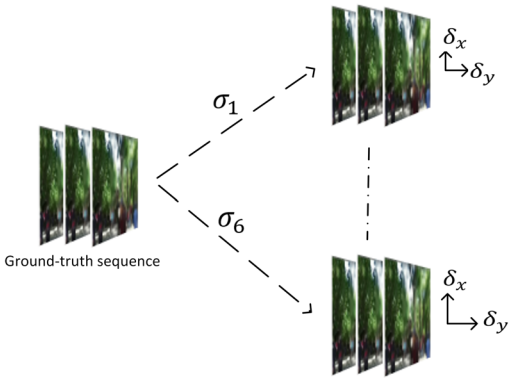


Fig. 4: Illustration of shakiness level generation.

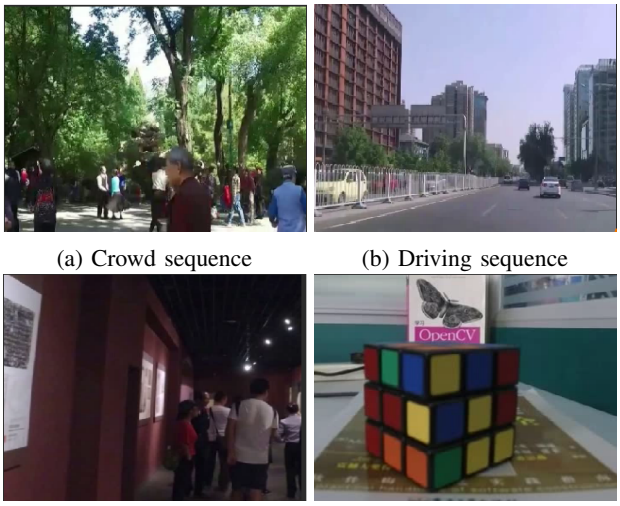
distribution has been carried out as follows: 80% of the data are dedicated to the training set, of which 10% is used for validation and 20% of the whole data are used for testing.

Table I demonstrates the performance parameter of the proposed model. In addition, it is worth to note that our goal is to estimate the graph shape of the acceleration instead of retrieving exact amplitudes. The proposed metric relies on acceleration fluctuations and does not take into account the intensity since it is considered as a relative value.

2) Shakiness metric - Experimentation: We demonstrate in this section the effectiveness of our shakiness metric against existing state-of-the-art metrics (ITF [8] (Eq. 6) and ISI [5] (Eq. 7) metrics). We add to our comparative study two robust and well-known video quality metrics (VIIDEO metric [10]

TABLE I: Proposed model - KPIs

	Training	Validation	Test
Mean absolute error (Average) $\lvert\hat{a}_i - \hat{\hat{a}}_i\rvert$	0,061	0,073	0,088
Mean absolute relative error (Average) $\frac{\lvert\hat{a}_i - \hat{\hat{a}}_i\rvert}{\hat{a}_i}$	1,1%	1,4%	1,8%



(a) Crowd sequence (b) Driving sequence

(c) Dark sequence (d) Object sequence

Fig. 5: Examples of our dataset used for evaluation

and VSFA metric [9]). VIIDEO and VSFA are well known to be designed to quantify the spatio-temporal aspect of a video which could represent shakiness distortions. Table II presents in detail the results obtained from the assessed metrics through a progressive shakiness intensity given by σ, while σ_{i+1} is more shaky than σ_i.

$$ITF = \frac{1}{N-1}\sum_{i=1}^{N-1} \text{PSNR}(f^i, f^{i+1}) \tag{6}$$

$$ISI = \frac{1}{N-1}\sum_{i=1}^{N-1} \text{SSIM}(f^i, f^{i+1}) \tag{7}$$

The results shown in Table II confirm the superiority of the proposed metric. The proposed approach outperforms all metrics in every category, except for riding and crowd videos. In these specific categories ISI and VIIDEO metrics show, respectively, a better result. This is probably due to the nature of our learning dataset, which does not include such kind of shakiness since these videos are characterized by specific motions of camera with various objects in motion. In fact, even if the proposed scheme does not have the best performance, we observe acceptable achievement from ITF and ISI which explains their wide use in DVS benchmark study. However, our study shows the limits of the VFSA and VIIDEO video quality metrics against the shakiness noise. Indeed, performance of the VFSA metric is totally uncorrelated in some categories (eg. Riding), while VIIDEO shows slightly better performance despite a very high time-consuming process. This proves once again the need to develop specific shakiness video quality measures.

In fact, time-consuming is another drawback of all these metrics expect ITF (see Table III). The proposed metric shows similar time execution performance as ITF, while performance of the proposed approach is drastically better. This characteristic allows the proposed metric to be implemented in real-

TABLE II: Ranking on progressive shakiness aggressiveness and Spearman correlation coefficients for ITF [8], ISI [5], VFSA [9] and VIIDEO [10] on each category. Spearman correlation is estimated between the shakiness magnitude expressed via σ_i and the quality metric outputs

Category	Metric	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	SpearCorr
Object	ITF [8]	2	4	5	3	6	1	0.09
	ISI [5]	1	2	4	6	5	3	0.60
	VIIDEO [10]	4	1	3	5	2	6	0.43
	VFSA [9]	2	1	3	4	5	6	0.94
	Our metric	1	2	3	4	6	5	**0.94**
Dark	ITF [8]	4	1	3	2	5	6	0.60
	ISI [5]	4	1	2	5	3	6	0.54
	VIIDEO [10]	2	1	4	5	6	3	0.78
	VFSA [9]	2	3	1	6	4	5	0.66
	Our metric	2	1	3	4	5	6	**0.94**
Crowd	ITF [8]	2	1	3	5	6	4	0.77
	ISI [5]	3	1	2	4	6	5	0.77
	VIIDEO [10]	2	1	3	4	5	6	**0.94**
	VFSA [9]	1	2	4	6	3	5	0.71
	Our metric	1	2	3	5	6	4	0.83
Running	ITF [8]	1	2	3	4	5	6	0.94
	ISI [5]	1	2	5	3	6	4	0.71
	VIIDEO [10]	1	2	3	6	5	4	0.77
	VFSA [9]	3	4	2	1	5	6	0.49
	Our metric	1	2	3	4	5	6	**1.00**
Riding	ITF [8]	2	1	3	6	5	4	0.71
	ISI [5]	2	1	3	4	6	5	**0.89**
	VIIDEO [10]	1	2	3	6	5	4	0.77
	VFSA [9]	3	6	5	4	2	1	-0.65
	Our metric	3	1	2	5	6	4	0.66
Climbing	ITF [8]	1	2	4	5	3	6	0.83
	ISI [5]	1	2	3	4	5	6	**1.00**
	VIIDEO [10]	1	2	3	6	4	5	0.83
	VFSA [9]	1	3	2	6	5	4	0.71
	Our metric	1	2	3	4	5	6	**1.00**
Driving	ITF [8]	4	2	5	1	3	6	0.26
	ISI [5]	2	1	4	3	5	6	0.88
	VIIDEO [10]	1	2	5	1	6	3	0.60
	VFSA [9]	6	3	2	4	5	1	-0.48
	Our metric	1	2	3	4	5	6	**1.00**
Walking	ITF [8]	1	2	4	5	6	3	0.66
	ISI [5]	1	2	3	5	6	4	**0.83**
	VIIDEO [10]	1	2	3	6	5	4	0.77
	VFSA [9]	2	3	5	1	4	6	0.54
	Our metric	1	2	3	5	6	4	**0.83**

TABLE III: Average time consuming performance of ITF [8], ISI [5], VFSA [9] and VIIDEO [10]

	ITF	ISI	Oracle	VFSA	Our metric
Avg time consuming(s) Resolution: 360*480 Size of the video: 16s Frame rate: 30fps	0.85 s	11.54 s	273 s	9.19 s	1.92 s

time in order to give an explicit output about quality of streamed videos, which opens up other perspectives for VSQA techniques.

V. Conclusion

In this study, we developed a blind Video Stabilization Quality Approach (VSQA) that estimates acceleration given by real sensors with no additional information more than video frames. The proposed metric shows its superiority in assessing video stability while at the same time being drastically better compared to the state-of-the-art approaches when it comes to time consumption. In addition, our experiments demonstrate the capacity of our architecture to learn data provided by the IMU sensor. This sensor emulation could open ambitious perspectives to retrieve sensors data from offline videos.

Acknowledgement

This work has received funding from the Research Council of Norway (project number 329034).

References

[1] P. R. Boyce and A. Wilkins, "Visual discomfort indoors," *Lighting Research & Technology*, vol. 50, no. 1, pp. 98–114, 2018.

[2] W. Guilluy, L. Oudre, and A. Beghdadi, "Video stabilization: Overview, challenges and perspectives," *Signal Processing: Image Communication*, vol. 90, p. 116015, 2021.

[3] M. Roberto e Souza, H. d. A. Maia, and H. Pedrini, "Survey on digital video stabilization: concepts, methods, and challenges," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–37, 2022.

[4] Y. Wang, Q. Huang, C. Jiang, J. Liu, M. Shang, and Z. Miao, "Video stabilization: A comprehensive survey," *Neurocomputing*, 2022.

[5] W. Guilluy, A. Beghdadi, and L. Oudre, "A performance evaluation framework for video stabilization methods," in *2018 7th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2018, pp. 1–6.

[6] L. Zhang, Q.-Z. Zheng, and H. Huang, "Intrinsic motion stability assessment for video stabilization," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 4, pp. 1681–1692, 2018.

[7] L. Zhang, Q.-Z. Zheng, H.-K. Liu, and H. Huang, "Full-reference stability assessment of digital video stabilization based on riemannian metric," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6051–6063, 2018.

[8] L. Marcenaro, G. Vernazza, and C. Regazzoni, "Image stabilization algorithms for video-surveillance applications," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, vol. 1, 2001, pp. 349–352 vol.1.

[9] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2351–2359. [Online]. Available: https://doi.org/10.1145/3343031.3351028

[10] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

[11] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[12] F. Schilling, "The effect of batch normalization on deep convolutional neural networks," 2016.

[13] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2011, pp. 342–347.

[14] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," *ACM transactions on graphics (TOG)*, vol. 32, no. 4, pp. 1–10, 2013.

VStab-QuAD: A New Video-Stabilization Quality Assessment Database

Borhen-eddine Dakkar[1], Amine Bourki[2], Azeddine Beghdadi[1] and Faouzi Alaya Cheikh[3]

[1]University Sorbonne Paris Nord, France
[2]VizioSense AI Lab, France
[3]Norwegian University of Science and Technology (NTNU), Norway
https://www.l2ti.univ-paris13.fr/databases/

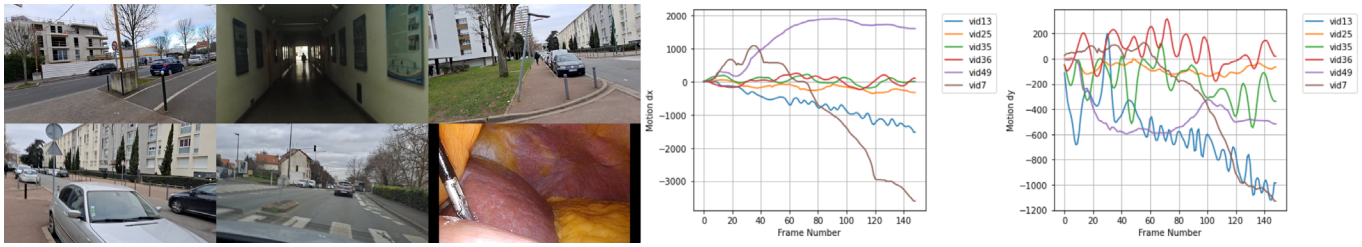


Fig. 1: Overview of the VStab-QuAD database for video stability quality assessment. It includes 320 stabilized video sequences, which stem from five different video stabilization methods that were applied to 64 unstable video sequences associated with various challenging scenarios (e.g., urban, selfies, zoom, medical imagery), image distortions, and perceptual quality scores.

***Abstract*—We introduce VStab-QuAD, a novel Video-Stabilization Quality Assessment Database, consisting of 320 stabilized video sequences obtained from 64 unstable video sequences with different contents (crowd, parallax, running, zooming, etc) using five different strong performing video-stabilization methods. The stabilized videos that are considered contain residual instability-induced image distortions and additional distortions introduced as a side effect of the digital stabilization algorithms applied. In addition, this study presents a comprehensive objective analysis of the video sequences and their validity for video-stabilization quality assessment. Subjective experiments were performed in a controlled environment using the single stimulus continuous quality evaluation (SSCQE) protocol, and their results are included in the database. Therefore, our proposed VStab-QuAD contains (i) the 64 original videos with in-capture distortions due to initial video instability, (ii) the 320 stabilized videos produced with the five different algorithms, and (iii) associated perceptual quality scores.**

This paper hence contributes VStab-QuAD as a comprehensive public benchmark database designed to facilitate the development of powerful video-stabilization methods and quality assessment metrics in real-life video acquisition scenarios.

***Index Terms*—Video Stabilization, Subjective Quality Assessment, Image Distortions, Datasets and Evaluation, Video Processing.**

I. Introduction

Thanks to the growing number of devices equipped with cameras, the production of videos has registered a notable increase [1]. In 2020, there were over three billion internet users who watched or downloaded a video content [1]. Most of the hand-held devices make it easy to capture video. The captured video with such mobile devices, are generally shaky and suffer from unwanted motions. As the human visual system is very sensitive to distortions in such visual material, the instability of a video is perceived as a disturbing visual degradation affecting the view experience.

This made video-stabilization one of the most important applications of video processing. It aims to improve the visual quality of video sequences by eliminating the unwanted motion. A video stabilization method generally consists of two major steps: motion estimation step and motion compensation step. Motion estimation allows the estimation of the global camera motion vectors. While in the motion compensation, global motion vectors are compensated for to remove the frames jitter and to produce smooth motion trajectories.
A robust video stabilization method should satisfy three major criteria regarding the perceptual quality: stability, cropping and visual distortion [2]. An unstable video means the presence of object instabilities in the whole frame. Cropping is characterised by irregular boundaries after warping frames. As the warped frames are cropped to obtain a rectangular boundary for normal video display. The visual distortion criteria could be any perceptual visual discomfort, e.g, blur, noise, objects geometry or illumination.

Most of the existing datasets are destined to video-stabilization methods development and testing [3]. To the best of our knowledge, no dataset were specifically dedicated to video-stabilization quality assessment. Our VStab-QuAD is the first one to fill that space. It aims to facilitate the quantification and benchmarking of video-stabilization algorithms. A new subjective methodology is proposed, that takes into account the complexity of the perceptual video stabilization problems.

The main contributions of this work are:

- VStab-QuAD, a novel database containing multiple challenging scenarios that can be used to assess the performance of video stabilization methods.
- A subjective video quality study using a new protocol for the evaluation of the distortions that are typically the by-product of digital video-stabilization.
- A new subjective quality score measure using a fusion scheme based on human observers' preferences. It is expressed as an explicit combination of the visual- and stability quality scores.

II. Related Work

Video Stabilization Quality Assessment (VSQA) methods can be divided into two main categories based on subjective, and objective assessment strategies.

Subjective assessment involves the investigation of the quality by asking people's judgement. A final score representing the quality is given for each sequence. The authors of [4] proposed a comparison of stabilized versions of synthetically distorted endoscopic videos and their synthetic full-reference ones. They have generated synthetically unstable videos. Then, four video-stabilization methods have been used to stabilize the generated videos. The quality were determined by comparing the videos to their respective full-reference ones. This dataset was limited only to the endoscopic Videos. In [5], a dataset of stable and shaky videos was constructed by capturing realistic stable/shaky videos. Then, digital video-stabilization algorithms have been running on shaky videos to obtain the stabilized sequences. They have used the Double Stimulus Continuous Quality Scale (DSCQS) to assess the subjective scores . A mean opinion score (MOS) is associated to each video (ranging from 1, being perfectly stable, to 5, extremely shaky). The dataset assesses only the stability and does not take into account other stabilization distortions. In [6], they propose a user study based on MOS to measure the subject preference based on the paired comparison between different methods. Observers are shown several videos and are asked to select the best one according to some predefined metrics. Zhang et al. [7] proposed a dataset composed of shaky videos and their corresponding stabilized ones by running some classic video stabilization methods. The stability is ranked by choosing the best stable video. The average rank is computed for each video to give the final order of the subjective judgement.

All the state-of-the-art proposed datasets evaluate only the stability [2], [8]–[10]. However, video-stabilization methods introduce other degradations that affect the perceptual quality. Neglecting these degradations gives a biased assessment. In this work, we intend to investigate all the video-stabilization impairments.

III. The VStab-QuAD Database

We have constructed a new Video-Stabilization Quality Assessment (VStab-QuAD) dataset, and then conducted a human subjective study. A new protocol dedicated to the video-stabilization QA has been adopted.

A. Acquisition of the Initial Unstable Videos

There are a total of 64 videos in different categories, including selfie, rotation, zooming, running, climbing, driving, endoscopic, rolling shutter, dark, and crowd with large parallax, among others. These videos are provided in raw AVI format, with frame rates ranging from 24 to 30 frames per second, and durations ranging from 5 to 10 seconds. Out of the 64 videos, 46 are novel acquisitions, and 18 were obtained from related to existing video-stabilization method papers [11]–[14].

B. Generation of Stabilized Videos

To produce stabilization versions of the initially unprocessed, real, video data, we consider 5 well established methods. The 5 softwares are (i) Adobe After Effects (Ae) warp stabilizer, (ii) Google Photos application, (iii) VirtualDub Deshaker, (iv) vReveal and (v) VideoProc Converter [15].

AE stabilizer is based on the subspace method in [4], and YouTube implements the method of L1-optimization on inter-frame transformations [2]. 320 stabilized videos are then produced by using these methods leading to a very broad range of motions trajectories and novel visual artefacts that typically stem from video stabilization (Figure 2).

C. Dataset Characteristics Analysis

It is important to have different scenarios but also visual contents rich in spatio-temporal structures at different scales of observation and under various lighting and viewing angles. The richness or diversity in videos is computed through spatial and spatio-temporal descriptors. A set of criteria and measures to quantify and analyze the richness and representativeness of image and video databases dedicated to perceptual quality assessment has been proposed in [16], [17]. In what follows we will recall the Saptio-temporal Descriptors and apply them to the analysis of the database.

D. Saptio-temporal Descriptors

The richness of a video signal in terms of visual time-varying content is measured using the Spatio-temporal information.The VQEG group formula has been used [18].

1) Spatial perceptual information: We have used the Sobel operator to enhance the edges as in [18]. We applied a Sobel filter for each video frame (F_n). The standard deviation of the magnitude of the Sobel response is then computed. The maximum value over time is used to represent the spatial information content of the video sequence.

$$\mathrm{SI} = \max_n\{\mathrm{std}[\mathrm{Sobel}(F_n)]\} \quad (1)$$

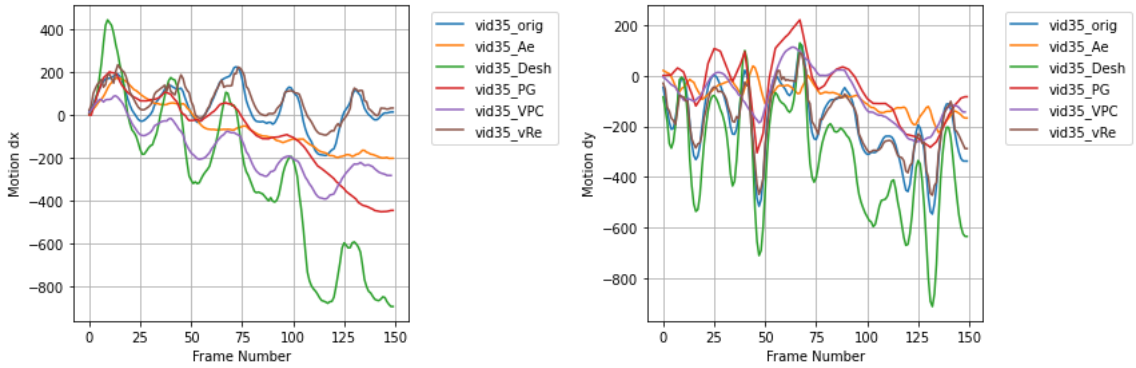


Fig. 2: An example video with its associated motion trajectory according to its horizontal (dx) and vertical (dy) components. The original trajectory is in blue, and the 5 video stabilization methods we consider, present the remaining colors.

2) Temporal perceptual information: To capture the motion difference at a location (i, j), the temporal perceptual information denoted $M_n(i, j)$ is used. It is computed as follows:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j) \tag{2}$$

where $F_n(i, j)$ is the pixel value at the (i, j) location of n^{th} frame. The TI measure is computed as the maximum over time (n) of the standard deviation over space of inter-frame difference M_n:

$$\text{TI} = \max_n\{\text{std}[M_n]\} \tag{3}$$

High motion Videos are characterized with high TI values. Figure 3 presents a scatter plot of the spatio-temporal descriptors SI and TI of the original shaky videos. It can be seen that the still scenes and those with very limited motion are found when SI is close to 0 (for samples 18 and 14), while scenes with a lot of motion are found near the upper part of the plot (for 63, 28, 32). Scenes with minimal spatial detail are located on the left side of the plot (18, 14, 19), while scenes with the most spatial detail are located on the right side of the plot (12, 32, 28).

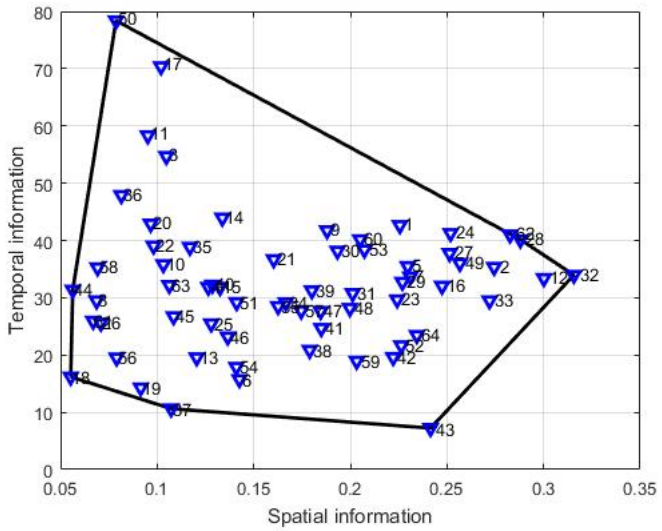


Fig. 3: Spatio-temporal diversity of the VStab-QuAD videos by means of their perceptual Spatial (SI) and Temporal (TI) criteria

E. Motion trajectory

Optical flow has been used to estimate the motion trajectory in the original videos. First, features are extracted from each frame and tracked using the Lucas-Kanade Optical Flow algorithm. The motion is then estimated using a rigid Euclidean transformation. The motion trajectory of some original videos in the x and y directions are presented in Figure 1.

IV. Subjective Study for a Joint Visual and Stabilization QA

A. Experimental Setup

The testing environment included a i7-7700 CPU @ 3.60GHz PC equipped with 1920×1080 monitor. A user interface was designed whereby the subjects could view and rate the videos. The observers begin the test by registering their information. In the next tab, the test instructions are explained. Two sessions are proposed in the following: Train session and Test session respectively. The aim of the Train session is to familiarize the observers with the test expectations. The Test session contains the real test where each subjective score is recorded. At the end, the observers are asked to give their preferences between the visual quality and the stabilization quality.

B. Experimental Protocol

Most of the existing video-stabilization metrics or datasets only consider stabilization-related evaluation criteria. However, other degradations may appear as a by-product of video stabilization. In order to obtain representative data on how humans perceive the video-stabilization distortions, we have conducted a subjective study to obtain subjective scores. For this, we have adopted a new protocol based on single stimulus continuous quality evaluation (SSCQE) to obtain the subjective

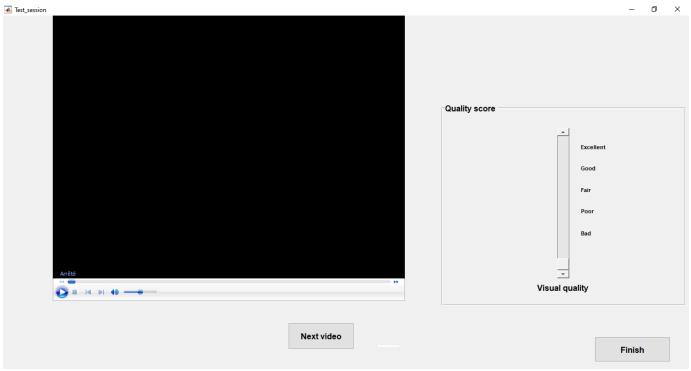


Fig. 4: GUI of our SSCQE protocol to construct our subjective quality ratings.

quality scores. It rates all of the stabilized videos. As we have reported before, evaluating a video-stabilization method regarding only its stability is not sufficient. In this study, the overall quality score is obtained by evaluating two criteria, the visual quality and the stabilization quality.

1) **Visual quality:** The removal of instabilities is often accompanied by the introduction of some side effects that affect the visual quality of the video sequences. In this part of the study, the observers are asked to evaluate the perceptual quality with regard to the following distortions:
   - Blur: it is characterized by the lack of sharpness.
   - Twisted backgrounds: this distortion mainly occurs on backgrounds. It represents a wave-like movement on the background.
   - Vibration (shudder): it appears as tremors in the sequences.
2) **Stabilization quality:** The main objective of digital video stabilization is the improvement of the video quality by removing unwanted camera motion. Two scenarios are possible after the application of a such video-stabilization method:
   i- The video remains unstable. It means that the stabilization process is inefficient with regard to the jitters presented in the video.
   ii- The moving objects present some latency. This distortion is the consequence of the suppression of the wanted motion in the video.

The experiments are performed to evaluate the visual quality and the stabilization quality. A continuous quality slider is used to derive the subjective scores. The quality slider is labeled with five adjectives as mentioned in Table I. All subjects were instructed to give an opinion quality score of the overall perceived video sequences. They were seated to perform the experiments at a fixed distance of twice the screen height. While this contrasts slightly with our overall experimental setup which mainly follows [18], we have decided to consider this typical distance relative to the considered screen resolution in order the induce more critical disparities within the gathered

TABLE I: Adjectival categorical judgement

Stabilization quality	**Visual quality**
Very annoying	Bad
Annoying	Poor
Slightly annoying	Fair
Perceptible, but not annoying	Good
Imperceptible	Excellent

ratings by allowing the observers to assess more clearly the distortions introduced by the stabilization process. This viewing distance was maintained during each test session, and throughout the considered observers. All the observers had either normal vision or corrected to normal vision.

A total of 15 subjects participated in the study, most of them are graduate students. The subjective test is started with general information about the study and instructions on how to participate to the video-stabilization task. In a short training session, 10 videos were played to familiarize subjects with the user interface. Following this, a test session was initiated. 160 stabilized videos were randomly and equally divided into 2 sessions. The subjects participated in 4 sessions, 2 for visual quality and 2 for stabilization quality. These sessions were separated by at least 24 hours. After completion of both testing sessions, subjects were asked to indicate their preference between stability and visual quality. This information will be used as a weighted trade-off parameter for the combination score.

C. Joint Visual and Stabilization Subjective Scores

The subjective score is computed as indicated in the ITU-R recommendation [19]. The ratings was converted into mean opinion scores (MOS).

$$MOS_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij} \tag{4}$$

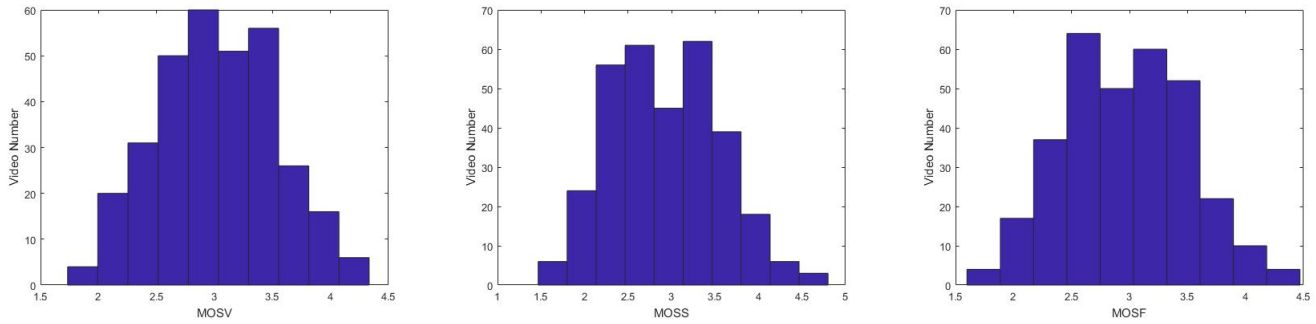
where r_{ij} represents the ratings of the j th image given by the i th subject and N is the number of subjects.
We named $MOSV$ and $MOSS$ the obtained score form the visual quality test and the obtained score from the stabilization quality test respectively.
The overall score of one video $MOSF$ is obtained using a the preference weight as follows:

$$MOSF_j = \alpha MOSV_j + (1-\alpha)MOSS_j \tag{5}$$

where α is a tuning parameter which encodes the relative preference between the perceptual visual quality and stabilization quality. It is inferred by means of a subjective survey as detailed further in the next sub-section.

D. Subjective scores

Figure 5 displays the histogram of Mean Opinion Scores (MOSs) across the entire database. The plot shows a wide range of perceptual quality scores, spanning all possible values. This indicates that the dataset contains a rich variety of MOS levels throughout.



(1) (2) (3)

Fig. 5: **Histogram of MOS scores covering the VStab-QA.** This includes MOSV for visual quality (1), MOSS for stabilization quality (2), and MOSF for the final joint MOS (3).

TABLE II: Relative preference weight between Visual and Stabilization quality, as estimated by surveying a human cohort.

Survey size	Visual quality	Stabilization quality
15	33.33%	66.66%

Table II presents the results of the subject preferences. We can notice that 66% of the subjects prefer a stable video, while 33% prefer a better visual quality. Regarding the preference results, we have fixed the preference weight to $\alpha = 0.33$. This parameter has been used to compute the overall MOS score.

V. Conclusion and Perspectives

We have introduced VStab-QuAD, a novel database dedicated to quality assessment of video stabilization methods. It is comprised of 320 stabilized video sequences from five different video stabilization methods that were applied to 64 unstable video sequences associated with various challenging scenarios, along visual- and stability focused MOSs. We proposed an experimental study using subjective scoring, as well as a novel combined score that jointly considers perceptual visual quality and video stabilization quality in a structured, unbiased way. In particular, our combined score will pave the way for the ever-growing body of literature on video stabilization in a blind, reference-free context [20]–[22] while taking into account their impact on visual quality.

Future work includes studying the correlation of objective metrics with our subjective unified MOS, which encompasses the considerations of stabilization and visual quality, to help cope with the by-product artifacts that are typically generated by VS methods.

References

[1] "statista," https://www.statista.com/statistics/1061017/digital-video-viewers-number-worldwide/, 2023.

[2] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," ACM Trans. Graph., vol. 32, no. 4, jul 2013. [Online]. Available: https://doi.org/10.1145/2461912.2461995

[3] Y. Wang, Q. Huang, C. Jiang, J. Liu, M. Shang, and Z. Miao, "Video stabilization: A comprehensive survey," Neurocomput., vol. 516, no. C, p. 205–230, dec 2022. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.10.008

[4] M. C. Offiah, N. Amin, T. Gross, N. El-Sourani, and M. Borschbach, "An approach towards a full-reference-based benchmarking for quality-optimized endoscopic video stabilization systems," in Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, ser. ICVGIP '12. New York, NY, USA: Association for Computing Machinery, 2012. [Online]. Available: https://doi.org/10.1145/2425333.2425398

[5] L. Zhang, Q.-Z. Zheng, H.-K. Liu, and H. Huang, "Full-reference stability assessment of digital video stabilization based on riemannian metric," IEEE Transactions on Image Processing, vol. 27, no. 12, pp. 6051–6063, 2018.

[6] W. Guilluy, A. Beghdadi, and L. Oudre, "A performance evaluation framework for video stabilization methods," in 2018 7th European Workshop on Visual Information Processing (EUVIP), 2018, pp. 1–6.

[7] L. Zhang, Q.-Z. Zheng, and H. Huang, "Intrinsic motion stability assessment for video stabilization," IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 4, pp. 1681–1692, 2019.

[8] A. Goldstein and R. Fattal, "Video stabilization using epipolar geometry," ACM Transactions on Graphics (TOG), vol. 31, no. 5, pp. 1–10, 2012.

[9] Y. J. Koh, C. Lee, and C.-S. Kim, "Video stabilization based on feature trajectory augmentation and selection and robust mesh grid warping," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5260–5273, 2015.

[10] M. Wang, G.-Y. Yang, J.-K. Lin, S.-H. Zhang, A. Shamir, S.-P. Lu, and S.-M. Hu, "Deep online video stabilization with multi-grid warping transformation learning," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2283–2292, 2018.

[11] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," ACM Trans. Graph., vol. 32, no. 4, jul 2013. [Online]. Available: https://doi.org/10.1145/2461912.2461995

[12] J. Yu, R. Ramamoorthi, K. Cheng, M. Sarkis, and N. Bi, "Real-time selfie video stabilization," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12 031–12 039.

[13] Z. Shi, F. Shi, W.-S. Lai, C.-K. Liang, and Y. Liang, "Deep online fused video stabilization," in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 865–873.

[14] A. Beghdadi, M. A. Qureshi, B.-E. Dakkar, H. H. Gillani, Z. A. Khan, M. Kaaniche, M. Ullah, and F. A. Cheikh, "A new video quality assessment dataset for video surveillance applications," in 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 1521–1525.

[15] "Videoproc," https://www.videoproc.com, 2023.

[16] S. Winkler, "Analysis of public image and video databases for quality assessment," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pp. 616–625, 2012.

[17] A. Beghdadi, M. A. Qureshi, B. Sdiri, M. Deriche, and F. Alaya-Cheikh, "Ceed-a database for image contrast enhancement evaluation," in Colour and Visual Computing Symposium (CVCS). IEEE, 2018, pp. 1–6.

[18] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," International telecommunication union, 1999.

[19] B. Series, "Methodology for the subjective assessment of the quality of television pictures," Recommendation ITU-R BT, vol. 500, pp. 500–13, 2012.

[20] J. Yu, R. Ramamoorthi, K. Cheng, M. Sarkis, and N. Bi, "Real-time selfie video stabilization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 036–12 044.

[21] A. Ciancio, E. A. da Silva, A. Said, R. Samadani, P. Obrador, et al., "No-reference blur assessment of digital pictures based on multifeature classifiers," IEEE Transactions on image processing, vol. 20, no. 1, pp. 64–75, 2010.

[22] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3677–3686.

[23] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in 27th ACM International Conference on Multimedia, 2019, pp. 2351–2359.

[24] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," IEEE Access, vol. 9, pp. 72 139–72 160, 2021.

[25] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "Cvd2014—a database for evaluating no-reference video quality assessment algorithms," IEEE Transactions on Image Processing, vol. 25, no. 7, pp. 3073–3086, 2016.

[26] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in 9th International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–6.

[27] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 9, pp. 2061–2077, 2017.

[28] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," IEEE Transactions on Image Processing, vol. 28, no. 2, pp. 612–627, 2018.

[29] R. Kumar, H. Sawhney, S. Samarasekera, S. Hsu, H. Tao, Y. Guo, K. Hanna, A. Pope, R. Wildes, D. Hirvonen, et al., "Aerial video surveillance and exploitation," Proceedings of the IEEE, vol. 89, no. 10, pp. 1518–1539, 2001.

[30] I. Bezzine, Z. A. Khan, A. Beghdadi, N. Al-Maadeed, M. Kaaniche, S. Al-Maadeed, A. Bouridane, and F. A. Cheikh, "Video quality assessment dataset for smart public security systems," in 23rd International Multitopic Conference (INMIC). IEEE, 2020, pp. 1–5.

[31] Z. A. Khan, A. Beghdadi, F. A. Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, Å. A. Fretland, B. Edwin, and O. J. Elle, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, vol. 11316. International Society for Optics and Photonics, 2020, p. 113160P.

[32] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of h. 264/avc video sequences transmitted over a noisy channel," in International Workshop on Quality of Multimedia Experience. IEEE, 2009, pp. 204–209.

[33] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "Ivp subjective quality video database," The Chinese University of Hong Kong, http://ivp. ee. cuhk. edu. hk/research/database/subjective, 2011.

[34] M. A. Qureshi, A. Beghdadi, and M. Deriche, "Towards the design of a consistent image contrast enhancement evaluation measure," Signal Processing: Image Communication, vol. 58, pp. 212–227, 2017.

[35] Z. Mortezaie, H. Hassanpour, and A. Beghdadi, "People re-identification under occlusion and crowded background," Multimedia Tools and Applications, pp. 1–21, 2022.

[36] P. Bouttefroy, A. Bouzerdoum, S. Phung, and A. Beghdadi, "Abnormal behavior detection using a multi-modal stochastic learning approach," in 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing. IEEE, 2008, pp. 121–126.

[37] P. L. M. Bouttefroy, A. Bouzerdoum, S. L. Phung, and A. Beghdadi, "Vehicle tracking using projective particle filter," in 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. IEEE, 2009, pp. 7–12.

[38] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in Human vision and electronic imaging VIII, vol. 5007. International Society for Optics and Photonics, 2003, pp. 87–95.

[39] H. Yu and S. Winkler, "Image complexity and spatial information," in 5th International Workshop on Quality of Multimedia Experience (QoMEX). IEEE, 2013, pp. 12–17.

[40] K. Matkovic, L. Neumann, A. Neumann, T. Psik, and W. Purgathofer, "Global contrast factor-a new approach to image contrast." Computational Aesthetics, vol. 2005, no. 159-168, p. 1, 2005.

[41] K. Garg and S. K. Nayar, "When does a camera see rain?" in 10th International Conference on Computer Vision (ICCV). IEEE, 2005, pp. 1067–1074.

[42] A. Beghdadi, M. Asim, N. Almaadeed, and M. A. Qureshi, "Towards the design of smart video-surveillance system," in NASA/ESA Conference on Adaptive Hardware and Systems (AHS). IEEE, 2018, pp. 162–167.

[43] A. Beghdadi, I. Bezzine, and M. A. Qureshi, "A perceptual quality-driven video surveillance system," in 23rd International Multitopic Conference (INMIC). IEEE, 2020, pp. 1–6.

[44] A. Beghdadi, M. A. Qureshi, S. A. Amirshahi, A. Chetouani, and M. Pedersen, "A critical analysis on perceptual contrast and its use in visual information analysis and processing," IEEE Access, vol. 8, pp. 156 929–156 953, 2020.

[45] E. Kalalembang, K. Usman, and I. P. Gunawan, "Dct-based local motion blur detection," in International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering. IEEE, 2009, pp. 1–6.

[46] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," IEEE Transactions on Image Processing, vol. 29, pp. 6054–6068, 2020.

[47] J. Choi, J. Park, and I. S. Kweon, "Self-supervised real-time video stabilization," arXiv preprint arXiv:2111.05980, 2021.

[48] M. K. Ali, S. Yu, and T. H. Kim, "Deep motion blind video stabilization," arXiv preprint arXiv:2011.09697, 2020.

[49] Z. Shi, F. Shi, W.-S. Lai, C.-K. Liang, and Y. Liang, "Deep online fused video stabilization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1250–1258.

[50] Y.-T. Chen, K.-W. Tseng, Y.-C. Lee, C.-Y. Chen, and Y.-P. Hung, "Pixstabnet: Fast multi-scale deep online video stabilization with pixel-based warping," in 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 1929–1933.

[51] Y.-L. Liu, W.-S. Lai, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Hybrid neural fusion for full-frame video stabilization," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2299–2308.

[52] M. Zhao and Q. Ling, "Pwstablenet: Learning pixel-wise warping maps for video stabilization," IEEE Transactions on Image Processing, vol. 29, pp. 3582–3595, 2020.

[53] S. Liu, P. Tan, L. Yuan, J. Sun, and B. Zeng, "Meshflow: Minimum latency online video stabilization," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. Springer, 2016, pp. 800–815.

[54] L. Zhang, Q.-Z. Zheng, H.-K. Liu, and H. Huang, "Full-reference stability assessment of digital video stabilization based on riemannian metric," IEEE Transactions on Image Processing, vol. 27, no. 12, pp. 6051–6063, 2018.

[55] Q. Rao, X. Yu, S. Navasardyan, and H. Shi, "Sim2realvs: A new benchmark for video stabilization with a strong baseline," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5406–5415.

[56] M. Sarıgül, "A survey on digital video stabilization," Multimedia Tools and Applications, pp. 1–27, 2023.

Improving Viewer Training in Visual Assessment

Mathias Wien
Lehrstuhl für Bildverarbeitung
RWTH Aachen University
Aachen, Germany
wien@lfb.rwth-aachen.de

Vittorio Baroncini
VABTech
London, UK
baroncini@gmx.com

Abstract— **Subjective assessment in the context of video compression performance evaluation measures the impairment or the quality of the observed video by the mean opinion score provided by naïve viewers. In a recent update of Recommendation ITU-R BT.500-14, this concept has been extended to expert viewers. Experts, who often have strong opinions on specific artifacts, provide wider confidence interval values than naïve viewers; this implies, therefore, a reduced ability to rank the results. This paper presents a new training procedure for viewers, aiming to reduce the spread of ratings either due to strong personal opinions (i.e., those most likely from expert viewers) or to hesitation in performing a new task (i.e., those most likely from naïve viewers). The new design of the training session includes suggesting of the expected range of scores, thereby helping the viewer to achieve a more consistent understanding of the scoring scale. In our experiment we apply the DSIS (Double Stimulus Impairment Scale) / DCR (Degradation Category Rating) test protocols. The same video clips were assessed by expert viewers both with and without the new training method, as well as by naïve viewers in a laboratory test. The analysis of the results shows a promising consistency in the results of the different tests.**

Keywords—Subjective assessment, viewer training, DSIS, DCR

I. INTRODUCTION

Subjective assessment is the established tool for determining the compression performance of video coding technology. Whereas objective metrics are frequently used in the development process due to their easy computation and reproducibility, the mean opinion score determined by formal subjective assessment provides the most reliable indication on the actual compression impact. Recommendation ITU-R BT.500-14 [1] includes the specification of subjective assessment with expert viewers; this approach has been demonstrated to be efficient and reliable. The design of compact and efficient tests with or without expert viewers is an active field of research, see. e.g. [2]. One characteristic of the expert tests is typically found in large confidence intervals (CIs). This can be partially attributed to the fact that experts show stronger opinions on specific artifacts. Since overlapping CIs for competing coding schemes imply an unclear interpretation of the results, achieving consistent test results with narrow confidence intervals in this process is desirable. The ability to achieve such improved results is of high importance especially when used in the context of tool evaluation in the development process of video coding standards.

In this paper, the impact of using different training methods for the subjective assessment is studied. The goal is to potentially develop enhanced training methods for the purpose of visual assessments with expert or naïve viewers. Expert viewing sessions are, e.g., frequently performed the context of video coding tool development by the Joint Video Experts Team (JVET) of ISO/IEC and ITU-T. Improved viewing methods providing results with small CIs would be of great benefit in this context.

To study the impact of the training method, three different visual assessments were performed on a set of compressed video clips: Two expert viewing tests using different training methods were conducted at a meeting with international experts. A formal test with naïve viewers was conducted in a laboratory setting. The details and context of this test as well as the tested coding schemes can be found in [4][5]. This paper builds upon [6] with an extended analysis, specifically including a study of the relation between the tests with naïve and expert viewers. The description of the modified training method is presented in [6] and is replicated in the context of this paper.

The remainder of this paper is organized as follows: In Section II the test setup of the different tests is described, including logistics, test set, and test methodology and design. Section III details the modified training method. In Section IV, the results of the tests are analyzed, and a discussion is provided. Section V concludes the paper.

II. TEST SETUP

A maximum number of 24 viewers participated in the tests at the meeting site as well as the laboratory. One on-site expert viewing was performed with 12 viewers, using a subset of video sequences of the laboratory test and applying the modified training procedure. In the remainder of this paper, the tests are denoted as follows:

- EXP1: Expert viewing test with normal training and 24 viewers;
- EXP2: Expert viewing test with modified training and 12 viewers;
- LAB: The laboratory test with 24 naïve viewers and modified training.

The LAB viewers were checked for visual acuity and normal color vision. Experts in EXP1 and EXP2 were requested to only volunteer under the condition of corresponding capability. It is noted that the number of viewers in EXP2 is lower than intended for the purpose of this study. Participation in the experiment was on a voluntary basis among the experts present on-site at the meeting and a higher number of participants could not be gained at the given point in time. 8 experts participated in both, EXP1 and EXP2. Despite the low number of participants, the results are considered to be worth publishing. Future work will include extended studies with higher numbers of viewers. Note that EXP1 and LAB include sessions for both, UHD and HD resolutions. The focus in this paper is on EXP2 which focuses on UHD.

TABLE II. TEST SETUP AT THE MEETING SITE (TOP). TEST SETUP AT THE LABORATORY SITE (BOTTOM)

Test Site	On-site
Display	2× LG 65" E9, HDMI (3840×2160)
Viewing distance	3 viewers at 1.5H
Viewing angle	±75°, 90° (at screen center)
Viewers EXP1	24 (4 female, 20 male)
Viewers EXP2	12 (3 female, 9 male)

Test Site	Laboratory
Display	LG 65" CX6LA, HDMI (3840×2160)
Viewing distance	1.5H
Viewing angle	90° (at screen center)
Viewers	24 (18 females, 6 males)

A. Logistics

1) On-site setup for expert viewing

At the meeting site, two identical setups were employed with three viewers placed in front of each screen. These included a PC with a Decklink video board for an HDMI connection and SSD drives capable of stable playout of the raw YUV data at the required frame rate. The test setup is summarized in the upper part of Table II. All volunteering experts confirmed visual acuity and normal color vision.

2) Laboratory setup

The laboratory setup is summarized in the lower part of Table II. The viewers, aged between 19 and 24, were checked for acuity and color blindness (18 females, 6 males).

B. Test sequences and rate points

The full tests included both UHD and HD test sequences which were assessed in the experiments EXP1 and LAB. Since EXP2 only used the UHD sequences, the focus of this paper is on this test set.

The test sequences used in the experiments are reported in Table I. All test sequences are of UHD resolution (3840×2160 pixels) and of 10 s length. The uncompressed YUV files were encoded using the random-access encoder configuration for the employed JVET reference software packages [8][9]. For each test sequence, a set of four rate points was determined in pre-experiments to cover a wide range of visual quality. Since the JVET focus for the tests was on the comparison of the VVC reference software (VTM) and the emerging Enhanced Compression model (ECM), the ECM performance was used for determining the rate points. These experiments are reported in [4]. The corresponding VTM bitstreams were generated allowing for a one-time QP switch (QP+=1) during encoding of the test sequence to match the VTM bitrate to the ECM. The switching points were chosen such that the ECM rate should never be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate. This approach corresponds to the method, e.g., used in the context of the call for proposals for VVC [7]. As an additional comparison point, VVenC [11] bitstreams were added to the tests, using rate matching with 2-pass encoding and rate control, thus enabling very close matching of the rate points. For details of the chosen quantizers and the QP switching points applied in the VTM simulations, the reader is referred to [5][6].

C. Test method and test design

The Degradation Category Rating (DCR) method was applied for the subjective evaluation [12]. The test sequences were evaluated using the 11-grade scale as specified in Rec. ITU-R BT.500-14 [1], shown in Fig. 1. Each basic test cell (BTC) is structured as follows: Text "Original" (1sec) - [uncompressed sequence] (10sec) - Text "A" (1sec) - [PVS] (10sec) - Text "Vote <N>" (5sec).

Here, PVS denotes the processed video sequence under evaluation.

TABLE I. TEST SEQUENCES.

Configuration	Sequence name	Frame rate
UHD RA	Campfire	30
UHD RA	CatRobot1	60
UHD RA	DaylightRoad2	60
UHD RA	DrivingPOV3	60
UHD RA	Marathon2	30
UHD RA	MountainBay2	30

Score	Impairment item	
10	Imperceptible	
9	Slightly perceptible	somewhere
8		everywhere
7	Perceptible	somewhere
6		everywhere
5	Clearly perceptible	somewhere
4		everywhere
3	Annoying	somewhere
2		everywhere
1	Severely annoying	somewhere
0		everywhere

Fig. 1. Meaning of the 11 grade numerical scale as specified in Rec. ITU-R BT.500-14 Table 2-4.

1) On-site tests

For EXP1, a total of 6 test sessions were designed: three for the UHD sequences and three for additional HD sequences which are not further considered in this paper. All test sessions included a stabilization phase of three BTCs. The scores of the stabilization phase were not regarded in the evaluation. The session duration was chosen to be no longer than 13 minutes (with a maximum of 24 votes) to avoid the impact of fatigue. Furthermore, the test sessions included trapping BTCs where the original uncompressed sequence was shown for evaluation. For EXP1, the participating experts were trained with one training session for UHD resolution (8 votes) and one session for HD resolution (7 votes). All test sequences under evaluation occurred at least once in the training sessions, and a selection of rate points representing the expected impairment range was presented for both resolutions. Before the presentation of the training sessions, the experts were instructed on the meaning of the impairment scale. Any occurring requests or questions on the test procedure or the scale were answered. The experts were advised to calibrate their personal voting scale during the training sessions and apply it in the actual test sessions.

The UHD and HD test sessions were both presented with a viewing distance of 1.5H from the UHD display for the center seat. The HD sequences were displayed without scaling in the center of the UHD area with a mid-gray padding around them. Thus, an effective 3H viewing distance for the HD content was achieved.

Experiment EXP2 used the same session design with a maximum duration of 13 minutes and used similar trapping

sequences as in EXP1. The viewing sessions were redesigned compared to EXP1 including a different randomization.

2) Laboratory tests

In the laboratory test, two test sessions for HD and three test sessions for UHD three were designed. Each test session included a three-BTC stabilization phase, with the scores being discarded before evaluation. The five test sessions were all made of 27 BTCs, for a test length of less than 12 minutes. One test point comparing reference vs. reference was included for each session. This allowed a check of the consistency of the test results and viewers' behavior.

Two training sessions were run, one for each resolution. In the training sessions all test sequences occurring in the actual test were shown at different compression rates to provide an overall indication of the impairments. After a general description of the experiment, the training was performed according to the method described below. The test was done using the DSIS test protocol, as described in Recommendation ITU-R BT.500-14. An degradation scale with 11 levels was used; the meaning of each impairment level was explained as found in Fig. 1.

III. MODIFIED TRAINING METHOD

A. Context

When a visual assessment of video is required, the procedure is a key aspect in the achievement of reliable and usable results. An important part of the procedure is the training of the volunteers participating in the experiment. Traditional training strategies were based on reading of an explanatory text followed by a short training test session; the training sessions were not permitted to include any test sequences used in the actual test; questions were allowed at the end of the training sessions. This approach showed strong limitations as the standard deviation of the data produced by the participants in the experiment is typically large and thereby, led to a reduction of the ability to rank the test points according to the resulting MOS scores.

More recent training strategies permitted the use of test sequences used in the actual test, and a deep and accurate description of the possible impairments when the training session was run. The content of the training session included as many visual quality cases as possible, which were able to cover the whole quality range foreseen in the experiment. This approach produced a significant improvement in the results by lowering the standard deviation (and, consequently, the confidence interval, CI), and, thereby, providing a higher discrimination capability for the visual assessment.

When expert viewing (i.e., a visual assessment performed by experts in the field of video coding instead of naïve viewers) was introduced, the training phase was initially considered to be less important. However, the results revealed a high level of standard deviation values. Based on this observation, the training for the expert viewers was refined, mainly devoted to providing an accurate description of the organization and structure of the experiment (i.e., what is shown and when and how to express the scores). Nonetheless, the values of the standard deviation in expert viewing experiments remained higher, when compared with those obtained from the same test with naïve viewers.

These observations motivate the redesign of the training procedures. This matter has not been resolved so far and remains a matter of further study. One possible approach to improve on the above-mentioned problems is described below.

B. Description

First, a set of "ground truth" MOS (mean opinion score) values is prepared for each test case to be used in the training sessions; the "ground truth" MOS values are prepared by the test administrators. The set of "ground truth" MOS values delineates an "ideal" visual assessment as a guideline for the viewers.

The training session is edited by modifying the BTCs shown during the training session: At the end of the voting period, the "ground truth" MOS value is disclosed. The test administrator comments on the presented value by the explaining that the shown score is the one suggested for a correct evaluation of that PVS.

This approach was applied for the training in the LAB tests. For the tests in EXP2, the corresponding "ground truth" MOS values taken from the expert viewing sessions of the tests in EXP1 which was conducted and evaluated before the beginning of the EXP2 tests. A training session with 9 BTCs was designed with examples of VTM and ECM bitstreams covering all test sequences occurring in the viewing sessions. It is important to note that in any case, no guidance on specific sequence details or what to look for is provided to the viewers.

IV. RESULTS AND ANALYSIS

In this section, a comparison of the MOS results achieved in the three different experiments is presented. The actual MOS corresponding CI values are available in [5][6]. MOS and CI are computed according to ITU-R BT.500 [1].

A. Data processing for EXP1

In one test session, a playout problem with one of the two PCs occurred for one test sequence. The affected experts were presented the missing BTCs in a separate session to complete their votes.

In a first evaluation step, the participants votes were screened with respect to the trapping BTCs. In a total of 6 cases, viewers voted below score of 8 for the original. In each case, the results of the session including this trapping BTC were not regarded for the affected viewer. As a second step, the outlier screening, according to ITU-R BT.500-14 A1-2.3.1 [1], was applied. Based on this, the scores of one participant were removed from the set. In a final processing step, isolated outliers, which were considered to be obvious errors, were removed.

B. Data processing for EXP2

The scores were first screened for the trapping BTCs. The viewers consistently scored the originals with a MOS score of 9 or 10, with only two votes giving a score of 8. The Pearson correlation coefficient of the viewers was in the range of 0.89 to 0.97 with an average value of 0.94. Based on these findings, no further outlier processing was applied.

C. Data processing for LAB

Due to issues with the bitstreams at the highest and lowest rate point of the DrivingPOV3 sequence which were available for the LAB tests, the results for these points were discarded.

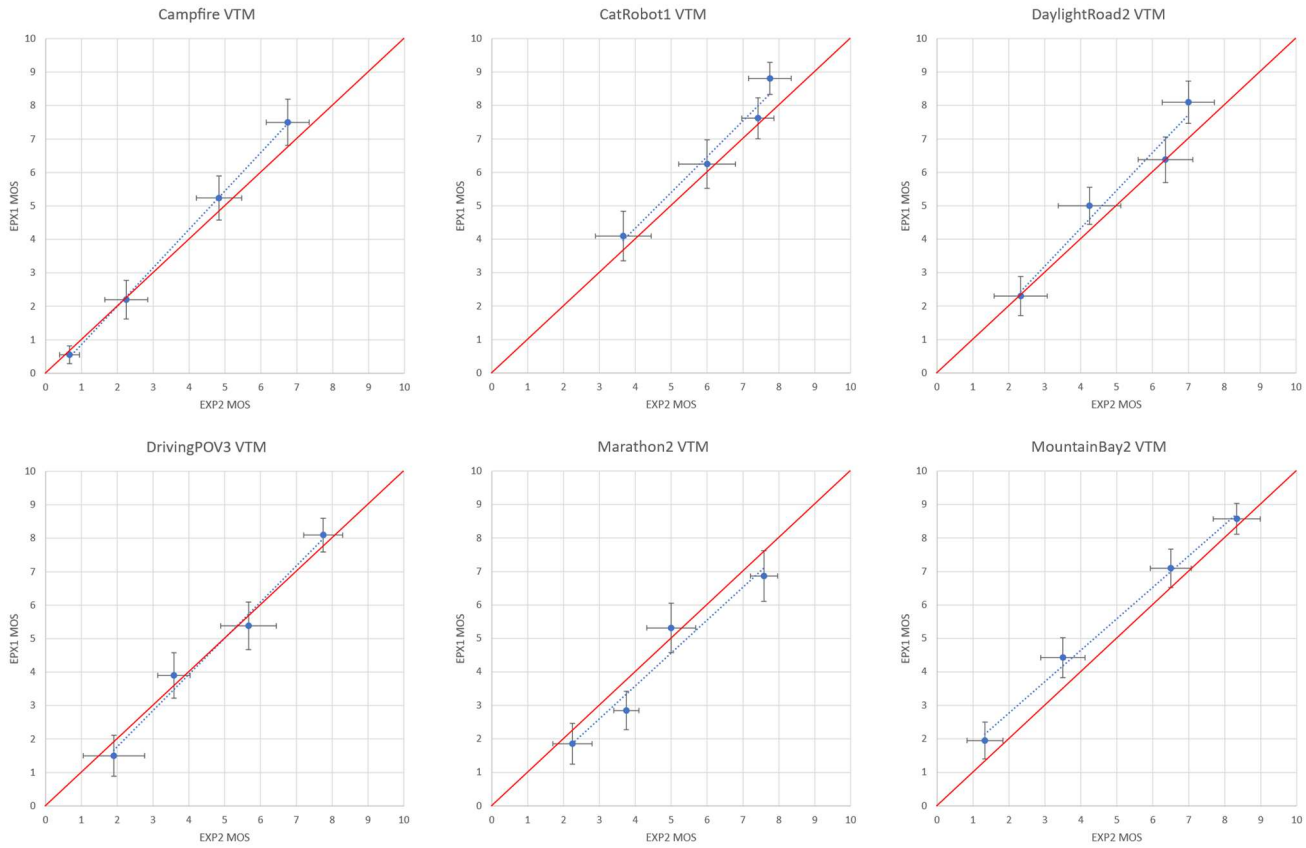


Fig. 2. Scatter plots comparing the MOS results of EXP1 and EXP2.

TABLE III. CORRELATION COEFFICIENTS BETWEEN THE TESTS

Correlation coefficient	PCC	SRCC
EXP1-EXP2	0.98	0.98
LAB-EXP1	0.97	0.96
LAB-EXP2	0.97	0.95

A statistical analysis (Pearson correlation) was done on the raw scores to verify the viewers' behavior. No viewer had to be excluded from the evaluation based on trapping BTCs. Only a few (6 out of 1920 scores) outliers in the raw data were discarded.

D. Comparison of MOS results

In order to study the relation of the MOS values resulting from the three tests, EXP1, EXP2, and LAB, scatter plots are provided where the MOS values of two of the tests are arranged along the abscissa and ordinate, respectively. The plots comparing the two expert viewing tests are presented in Fig. 2. The plots comparing EXP1 and EXP2 to LAB are provided in Fig. 3 and Fig. 4, respectively. Since MOS results for the VTM bitstreams have been evaluated in all tests, these are reported for all sequences. The MOS results for the VVenC bitstreams are reported for EXP2 and LAB since these were not evaluated in EXP1. The linear trendline for the VTM points is plotted as a blue dashed line. The trendline for the VVenC points is plotted in orange for the applicable cases. Further, the confidence interval on the corresponding axis is indicated for each data point. The line of slope 1 is added to the diagrams in red to mark the location of a theoretically ideal match between the two tests.

The Pearson correlation coefficient (PCC) and the Spearman rank order coefficient (SRCC) between the MOS scores of the three tests are provided in Table III. For computation of the coefficients, the full set of comparable data is used, i.e., both VTM and VVenC MOS scores for computing the correlation of LAB and EXP2, and only the VTM scores when computing the correlation of EXP1 and either EXP2 or LAB.

E. Discussion

The scatter plots comparing EXP1 and EXP2 reveal a consistent pattern of the MOS scores resulting from the two expert viewing tests. The PCC and the SRCC between EXP1 and EXP2 have a value of 0.98. For all test sequences, the trendline of the data points is close to the main diagonal. Out of a total of 28 data points, 19 points have an overlap of the CIs for both axes with the main diagonal. In three cases, only one of the CIs overlaps, and in six cases, the CIs do not overlap with the main diagonal. Based on these observations we conclude that the results are consistent for both tests. When comparing the confidence intervals by size, in many cases a very similar CI is observed. When computing the ratio between the CIs of data points from EXP1 and EXP2, a ratio range of 0.5 to 1.59 is observed, with an average of 1.04 and a standard deviation of 0.25. These results suggest that a comparable confidence has been achieved with only half of the number of expert viewers. Taking into consideration the effort of conducting expert viewing tests, this result is taken as an encouraging indication to further study and potentially refine the modified training method.

The scatter plots comparing the results of the expert viewing tests to the laboratory tests show a similar pattern. The PCC for both, LAB vs. EXP1 and LAB vs. EXP2, is equal to 0.97. The SRCC for LAB vs. EXP1 is 0.96 and the SRCC for LAB vs. EXP2 is 0.95. Here, the correlation coefficients for LAB vs. EXP2 are computed using the full set of VTM and VVenC data. When studying the scatter plots, it is observed that the expert viewers scored more critically than the naïve subjects especially in the lower quality range. Still, a quite consistent scoring behavior is observed for both the LAB-

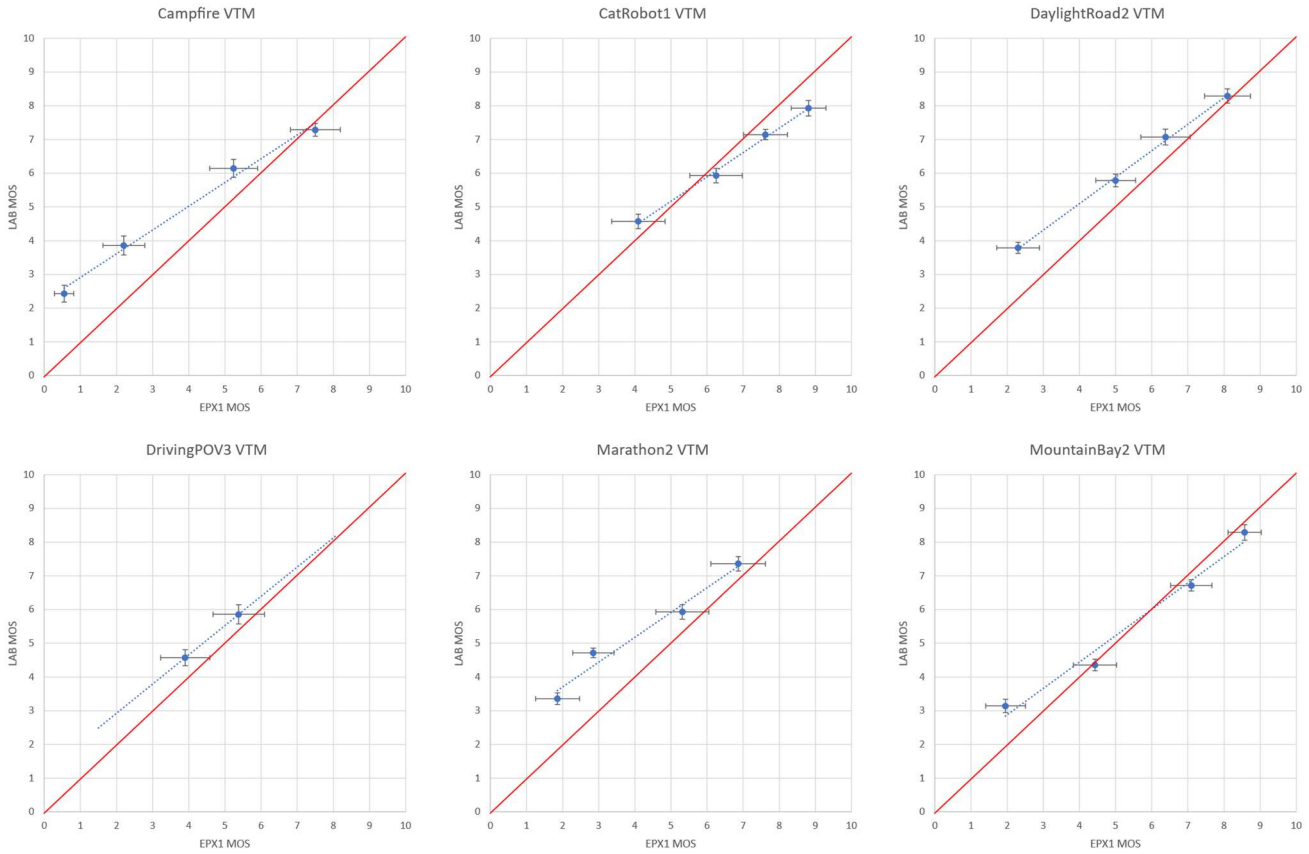


Fig. 3. Scatter plots comparing the MOS results of LAB and EXP1.

EXP1 and the LAB-EXP2 comparison. For the VVenC curve of the Campfire sequence in the LAB-EXP2 comparison, a saturation effect is noted for the expert viewers while the naïve viewers continue to show differentiating MOS values. At the same time, the results for the lower VVenC rate points are quite consistent between the expert and the naïve viewers. These observations are considered to potentially stem from differences in the setup, especially since test room for the on-site tests was not at the same level as the laboratory room. For the test sequence DrivingPOV3, only the two middle rate points are evaluated for the VTM bitstreams, as discussed earlier. In general, the CIs for the experiments with the naïve viewers are considerably smaller than the CIs of the expert viewing tests. The observed ratio range for the CIs in the LAB-EXP1 comparison is between 1.05 and 4.06 with an average of 2.98 and a standard deviation of 0.74. For the LAB-EXP2 comparison, the range is 1.11 to 5.34 with an average of 3.05 and a standard deviation of 0.99.

V. Conclusions

This paper presents a comparison of MOS scores resulting from three tests using DCR conducted on the same data set: An on-site expert viewing, an on-site expert viewing using a modified training procedure, and a test with naïve viewers using the modified training procedure in the controlled environment of the laboratory. The results of the expert viewing tests suggest that the new proposed training procedure is functioning. Comparable CI values with half the number of viewers are observed. The results of both expert viewing tests are found to be consistent with the results of the formal test done with naïve viewers, further confirming the validity of this protocol. We suggest to further study the modified training method as a potential improvement for the established DSIS/DCR test protocols. This includes conducting more tests with the new protocol in a laboratory environment and with a higher number of expert viewers.

References

[1] Recommendation ITU-R BT.500-14 (2019), *Methodologies for the subjective assessment of the quality of television images*.

[2] P. Perez, L. Janowski, N. Garcia, M. Pinson, "Subjective Assessment Experiments That Recruit Few Observers with Repetitions (FOWR)," IEEE Transactions on Multimedia, vol. 24, pp 3442-3454, 2022.

[3] J.-R. Ohm, "Meeting Report of the 28th JVET Meeting," Doc. JVET-AB1000, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 28th meeting, Mainz, DE, Oct. 2022.

[4] M. Wien, "AHG4, 7, 12: Report on AHG meetings on ECM performance evaluation preparation," Doc. JVET-AB0041, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 28th meeting, Mainz, DE, Oct. 2022.

[5] M. Wien, J.-R. Ohm, V. Baroncini "Visual quality comparison of ECM/VTM encoding," Doc. JVET-AB2029, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 28th meeting, Mainz, DE, Oct. 2022.

[6] M. Wien, V. Baroncini, "Training Methods in Visual Assessment: Potential Improvements for Expert Viewing Tests," Doc. JVET-AC0267, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2023.

[7] A. Segall, V. Baroncini, J. Boyce, J. Chen, T. Suzuki (editors), "Joint Call for Proposals on Video Compression with Capability beyond HEVC," Doc. JVET-H1002, Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Oct. 2017.

[8] F. Bossen, X. Li, V. Seregin, K. Sharman, K. Sühring, "VTM and HM common test conditions and software reference configurations for SDR 4:2:0 10 bit video," Doc. JVET-Y2010, Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2022.

[9] M. Karczewicz, Y. Ye, "Common Test Conditions and evaluation procedures for enhanced compression tool testing," Doc. JVET-Y2017, Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2022.

[10] JVET, VTM software package, https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-11.0.

[11] VVenC software repository. https://github.com/fraunhoferhhi/vvenc.

[12] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*

MAiVAR-T: Multimodal Audio-image and Video Action Recognizer using Transformers

Muhammad Bilal Shaikh*, Douglas Chai†, Syed Mohammed Shamsul Islam‡ and Naveed Akhtar§
{*†School of Engineering, , ‡School of Science}, Edith Cowan University
§Department of Computer Science & Software Engineering, The University of Western Australia
Email: {*mbshaikh@our.,†d.chai,‡syed.islam}@ecu.edu.au, §naveed.akhtar@uwa.edu.au

***Abstract*—In line with the human capacity to perceive the world by simultaneously processing and integrating high-dimensional inputs from multiple modalities like vision and audio, we propose a novel model, MAiVAR-T (Multimodal Audio-Image to Video Action Recognition Transformer). This model employs an intuitive approach for the combination of audio-image and video modalities, with a primary aim to escalate the effectiveness of multimodal human action recognition (MHAR). At the core of MAiVAR-T lies the significance of distilling substantial representations from the audio modality and transmuting these into the image domain. Subsequently, this audio-image depiction is fused with the video modality to formulate a unified representation. This concerted approach strives to exploit the contextual richness inherent in both audio and video modalities, thereby promoting action recognition. In contrast to existing state-of-the-art strategies that focus solely on audio or video modalities, MAiVAR-T demonstrates superior performance. Our extensive empirical evaluations conducted on a benchmark action recognition dataset corroborate the model's remarkable performance. This underscores the potential enhancements derived from integrating audio and video modalities for action recognition purposes. To ensure transparency and reproducibility of our work, the source code is made publicly available at https://bit.ly/43do8DH.**

***Index Terms*—Multimodal Fusion, Transformers, Human Action Recognition, Deep Learning.**

I. Introduction

Human action recognition has become a critical task in various fields such as surveillance [1], robotics [2], interactive gaming [3], and health care [4]. Traditionally, most approaches have focused on visual cues [5]. However, human actions are not limited to visual manifestations; they also consist of rich auditory information [6]. Accordingly, Multimodal human action recognition (MHAR) that incorporates both visual and audio cues can provide more comprehensive and accurate recognition results [7].

Despite these promising prospects, the performance of current MHAR models is hampered by challenges of multimodal data fusion. Existing methods, including Convolutional Neural Networks (CNNs) [8]–[10] require significantly more computation than their image counterparts, some architectures factorise convolutions across spatiotemporal dimensions. Contrastingly, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) [11] have demonstrated constraints in processing large sequences, memory efficiency and parallelism.

In this paper, we propose a novel transformer-based model, Multimodal Audio-image and Video Action Recognizer using Transformers (MAiVAR-T). Our approach capitalizes on the self-attention mechanism inherent in transformers [12] to extract relevant features from both modalities and fuse them effectively. The proposed MAiVAR-T model outperforms state-of-the-art MHAR models on benchmark datasets [13], demonstrating the potential of transformer-based architectures in improving multimodal fusion and recognition accuracy.

To summarize, the contributions made in this paper are:

- A new feature representation strategy is proposed to select the most informative candidate representations for audio-visual fusion;
- Collection of effective audio-image-based representations that complement video modality for better action recognition are included;
- We apply a novel MAiVAR-T framework (see Fig. 1) for audio-visual fusion that supports different audio-image representations and can be applied to different tasks; and
- State-of-the-art results for action recognition on the audio-visual dataset have been reported.

The remainder of the paper is organized as follows: we begin with a review of related works on MHAR (Section II), followed by a detailed discussion of the proposed methodology (Section III). We then present the experimental setup (Section IV) and report the results (Section V). Finally, we concludes the paper with future directions (Section VI).

II. Related Work

A. Deep Learning for MHAR

Recently, deep learning models have shown remarkable results in MHAR [14]. They are capable of automatically learning a hierarchy of intricate features from raw multimodal data, which are beneficial for action recognition tasks.

CNNs have been widely adopted for MHAR to automatically extract spatial features from input data [15], and LSTMs are typically used for modelling the temporal dynamics of actions [11]. However, the traditional combination of CNNs and LSTMs for MHAR faces challenges such as ineffective multimodal fusion and difficulty handling long temporal sequences.

Transformers, introduced by Vaswani et al. [12], have demonstrated their superiority in many fields like natural

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

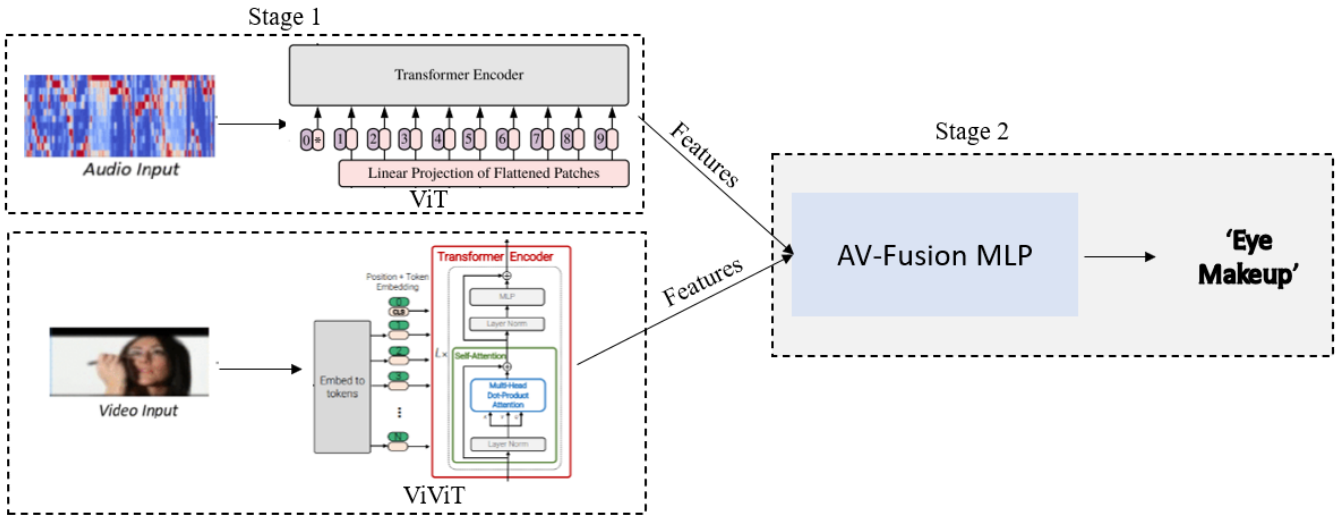


Fig. 1: The proposed framework contains two stages. The first stage extracts the features influencing the recognition while the second stage performs classification on the fused features. The input sequence consists of image and audio-image patches. These are then projected into tokens and appended to special CLS (classification). Our transformer encoder then uses self attention to model unimodal information, and send cross-modal information flow through to fusion network.

language processing [16], image classification [17], and video understanding [18]. The self-attention mechanism through its optimal complexity (see Table I) in transformers could potentially enhance the capability of feature extraction and multimodal fusion in MHAR tasks. However, the utilization of transformers in MHAR is relatively unexplored and demands further investigation.

B. Audiovisual Learning and Fusion

The field of audiovisual multimodal learning has a long and diverse history, both preceding and during the deep learning era [19]. Early research focused on simpler approaches, utilizing hand-designed features and late-stage processing, due to limitations in available data and computational resources [20]. However, with the advent of deep learning, more sophisticated strategies have emerged, enabling the implicit learning of modality-specific or joint latents to facilitate fusion. As a result, significant advancements have been achieved in various supervised audiovisual tasks [21].

It is common to jointly train multiple modality-specific convolution networks, where the intermediate activations are combined either through summation [22]. On the other hand, in transformer-based architectures, the incorporation of Vision Transformers (ViT) [17] and Video Vision Transformers (ViViT) [18] has brought about significant advancements in multimodal human action recognition. Initially, ViT proved instrumental in dissecting images into smaller segments, to interpret these patches as a sequence for more accurate image understanding. This ability greatly improved the recognition and classification of human actions within still images. The introduction of ViViT further extended this capacity, applying transformer techniques to analyze video data. By processing sequences of video frames, ViViT effectively interprets

TABLE I: Complexity comparison for different types of layer. Notations: n : sequence length, d: representation dimension, k kernel size.

Layer Type	Complexity per layer	Sequential Operations	Maximum Path Length
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Self-Attention	$O(n^2 \cdot d)$	$O(1$	$O(1)$

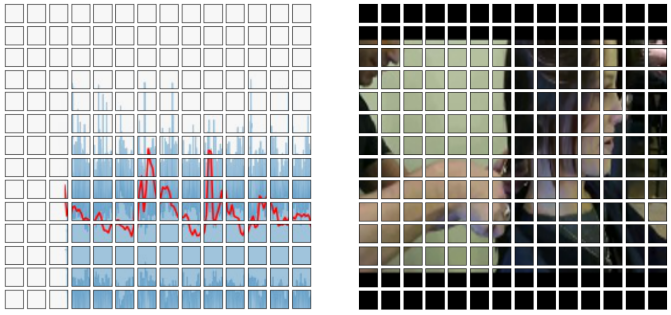
TABLE II: Hyper-parameters of the network.

Parameter	Value
Batch size	256
Initial learning rate	0.001
lr decay (every 4 epochs)	0.10
Learning rate patience	10
Epochs	100

the spatio-temporal dynamics involved in human movements. Together, the use of Vision Transformers and Video Vision Transformers can produce a shift in multimodal human action recognition, enhancing the capability of systems to accurately classify and understand complex human activities across visual and audio domains.

III. Proposed Methodology

Data Collection: We collected human actions from a benchmark dataset called UCF101 [13], with each instance containing video clips and their corresponding audio streams. UCF-101 contains an average length of 180 frames per video. We observed that half of the videos in the dataset contained no audio. Thus, in order to focus on the effect of audio features, we used only those videos that contained audio. This resulted in 6837 videos across 51 categories. Whilst this led the dataset



(a) (b)

Fig. 2: Image patches (a) Audio-image representation, (b) RGB video frame.

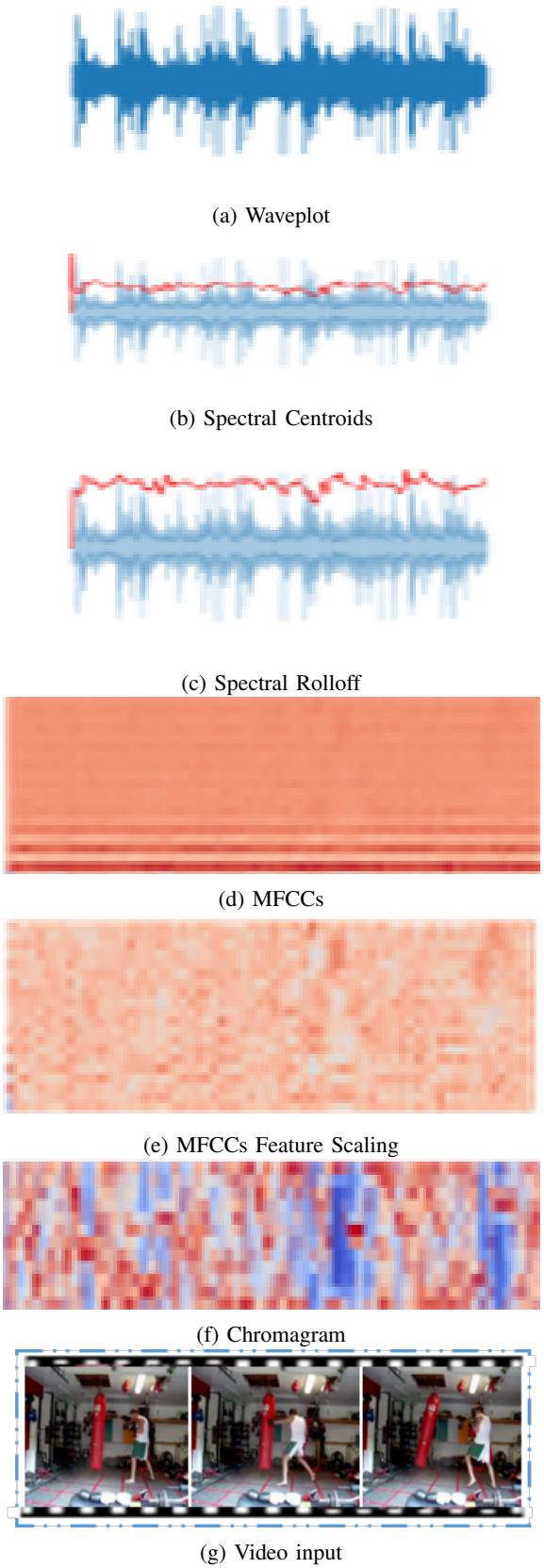
to be significantly reduced, the distribution of the audio dataset was similar to the video dataset. We used the first train-test split setting provided with this dataset, which resulted in 4893 training and 1944 testing samples. We reported the top 1 accuracies obtained by training on split 1.

Data Preprocessing: The video and audio data were preprocessed separately, as described in the following subsections. The video data was transformed into frames, while the audio data was converted into six audio-image representations following [14], [23]. Standard normalization techniques were applied to both modalities.

Audio image representations: Following are some of the key characteristics of audio-image representations (shown in Figure 3).

- Audio image representations provide a significant reduction in dimensionality. For example, spectral centroid images represent the frequency content of the audio signal over time, which is a lower-dimensional representation of the original video dataset. This can make it easier and faster to process the data and extract meaningful features.
- Audio images are based on the audio signal, which is less affected by visual changes, such as changes in lighting conditions or camera angles. This makes these representations more robust to visual changes and can improve the accuracy of human action analysis.
- Standardization as audio images can be standardized to a fixed size and format, which can make it easier to compare and combine data from diverse sources. This can be useful for tasks such as cross-dataset validation and transfer learning. Hence, this dataset can serve as a standard benchmark for evaluating the performance of different machine-learning algorithms for human action analysis based on audio signals.
- Suitable for privacy-oriented applications such as surveillance or healthcare monitoring, which may require the analysis of human actions without capturing the original visual information.

Architecture: The MAiVAR-T model comprises an audio transformer, a video transformer, and a cross-modal attention layer. The transformers process the audio and video inputs



(a) Waveplot

(b) Spectral Centroids

(c) Spectral Rolloff

(d) MFCCs

(e) MFCCs Feature Scaling

(f) Chromagram

(g) Video input

Fig. 3: Segmented video input and six different audio-image representations of the same action.

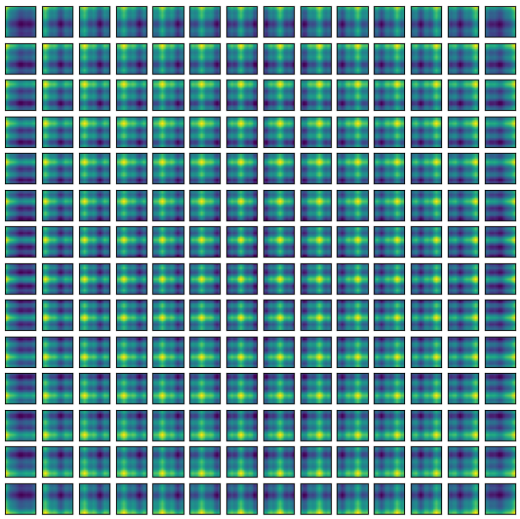


Fig. 4: Positional embeddings.

separately, after which the cross-modal attention layer fuses the outputs. Finally, a classification layer predicts the action present in the input data.

Audio Stream: The audio stream uses Vision Transformer (ViT) [24] to process 2D images with minimal changes. In particular, ViT extracts N non-overlapping image patches, $x_i \in \mathbb{R}^{h \times w}$, performs a linear projection and then rasterises them into 1D tokens $z_i \in \mathbb{R}^d$. The sequence of tokens input to the following transformer encoder is

$$\mathbf{z} = [z_{cls}, \mathbf{E}x_1, \mathbf{E}x_2, ..., Ex_N] + \mathbf{p}, \tag{1}$$

where the projection by E is equivalent to a 2D convolution. In addition, a learned positional embedding, $p \in \mathbb{R}^{N \times d}$, is added to the tokens to retain positional information, as the subsequent self-attention operations in the transformer are permutation invariant. The tokens are then passed through an encoder consisting of a sequence of L transformer layers. The MLP consists of two linear projections separated by a GELU non-linearity and the token-dimensionality, d, remains fixed throughout all layers. Finally, a linear classifier is used to classify the encoded input based on $z^L_{cls} \in \mathbb{R}^d$, if it was prepended to the input, or a global average pooling of all the tokens, z^L, otherwise. As the transformer [12], which forms the basis of ViT [17], is a flexible architecture that can operate on any sequence of input tokens $z \in \mathbb{R}^{N \times d}$, we describe strategies for tokenising videos next.

Video Feature Stream: We consider mapping a video $\mathbb{V} \in \mathbb{R}^{T \times H \times W \times C}$ to a sequence of tokens $z' \in \mathbb{R}^{n_t \times n_h \times n_w \times d}$. We then add the positional embedding and reshape into $\mathbb{R}^{N \times d}$ to obtain z, the input to the transformer.

IV. Experiments

A. Audio preprocessing

Each audio image representation was broken into patches as illustrated in the examples shown in Figure 2. For spatial context, positional embeddings for each input were projected into the architecture (see Figure 4). An internal schematic of the transformer model has been illustrated in Figure 5. Training data was batched into mini-batches of 16 instances each. Augmentation techniques like random cropping and time-stretching were applied to increase model robustness.

B. Video preprocessing

Following [18], the features extracted are then fed to the multimodal fusion module (AV-Fusion MLP) which later performs the classification for each action class.

C. Training

We utilized a multimodal cross-entropy loss function for training, balancing both audio and video modalities. The network hyperparameters are reported in Table II.

Hardware and Schedule: The training was performed on a high-performance computing cluster, equipped with GeForce GTX 1080 Ti GPUs. We trained the transformer-based model for 100 epochs, with a learning rate (α) schedule that decreased the rate by 10% every 4 epochs. **Optimizer:** The Adam optimizer [25] was used due to its effectiveness in training deep networks. **Regularization:** Dropout techniques [26] were applied to prevent overfitting during training.

V. Results

To assess the contribution of each component in our model, we performed an ablation study. Results demonstrate that both the audio and video transformers, as well as the cross-modal attention layer, contribute significantly to the final action recognition performance. The process of attention mechanism in the extraction of features through robust audio-image representations could be visualized in Figures 6 and 7. We have used an accuracy metric that measures the proportion of correct predictions made by the model out of all the predictions and defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

where TP are the correctly predicted positive values. TN are the correctly predicted negative values. FP, also known as Type I errors, are the negative values incorrectly predicted as positive. FN, also known as Type II errors, are the positive values incorrectly predicted as negative.

Table III compares the performance of transformer-based feature extractors with CNN-based counterparts. Proposed MAiVAR-T outperforms prior methods by a +3% as presented in Table IV.

VI. Conclusion

Over the past decade, Convolutional Neural Networks (CNNs) with video-based modalities have been a staple in the field of action video classification. However, in this paper, we challenge the indispensability of video modalities and propose a transformer-based multi-modal audio-image to video action recognition framework called Multi-modal Audioimage-Video Action Recognizer using Transformers (MAiVAR-T). This

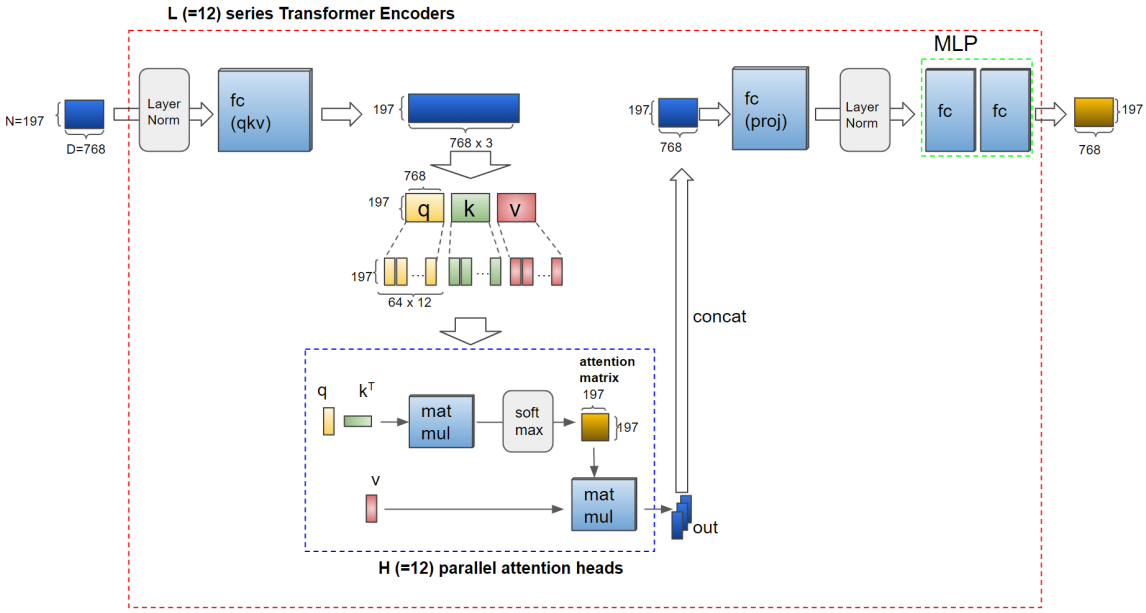


Fig. 5: Schematic of Vision Transformer Encoder.

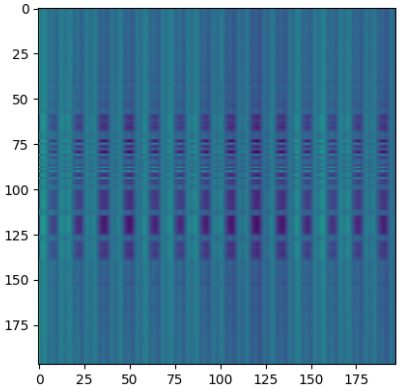


Fig. 6: Attention matrix for an audio-image representation.

TABLE III: Test accuracy of different audio representations with CNN and transformer-based backbones (InceptionResNet-v4(IRV4) and Vision Transformer (ViT) respectively)

Representation	IRV4	ViT
Waveplot	12.08	**19.7 (+7)**
Spectral Centroids	13.22	**28.65 (+15)**
Spectral Rolloff	16.46	**26.85 (+10)**
MFCCs	12.96	**18.26 (+6)**
MFCCs Feature Scaling	17.43	**17.44 (+0.01)**
Chromagram	15.48	**19.08 (+3)**

fusion-based, end-to-end model for audio-video classification features a transformer-based architecture that not only simplifies the model but also enhances its performance.

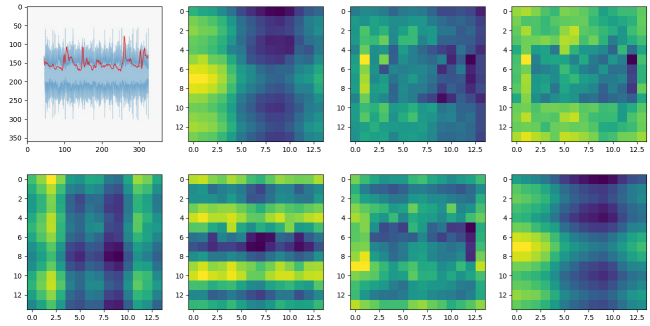


Fig. 7: Visualization of attention.

Experimental results demonstrate that our transformer-based audio-image to video fusion methods hold their own against traditional image-only methods, as corroborated by previous research. Given the significant improvements observed with pre-training on larger video datasets, there is considerable potential for further enhancing our model's performance. In future work, we aim to validate the efficacy of integrating text modality with audio and visual modalities. Furthermore, the scalability of MAiVAR-T on large-scale audio-video action recognition datasets, such as Kinetics $400/600/700$ will be explored. Additionally, we plan to explore better architectural designs to integrate our proposed approach with more innovative ideas, such as integrating generative AI-based transformer architectures, into our network could provide valuable insights into the impact of transformers on MHAR.

ACKNOWLEDGMENT

This work is jointly supported by Edith Cowan University (ECU) and the Higher Education Commission (HEC)

TABLE IV: Classification accuracy of MAiVAR compared to the state-of-the-art methods on UCF51 dataset after fusion of audio and video features.

YEAR	METHOD	ACCURACY [%]
2015	C3D [27]	82.23
2016	TSN (RGB) [28]	60.77
2017	C3D+AENet [29]	85.33
2018	DMRN [30]	81.04
2018	DMRN [30] + [31] features	82.93
2020	Attention Cluster [32]	84.79
2020	IMGAUD2VID [6]	81.10
2022	STA-TSN (RGB) [33]	82.1
2022	MAFnet [31]	86.72
2022	MAiVAR-WP [14]	86.21
2022	MAiVAR-SC [14]	86.26
2022	MAiVAR-SR [14]	86.00
2022	MAiVAR-MFCC [14]	83.95
2022	MAiVAR-MFS [14]	86.11
2022	MAiVAR-CH [14]	87.91
Ours	**MAiVAR-T**	**91.2**

of Pakistan under Project #PM/HRDI-UESTPs/UETs-I/Phase-1/Batch-VI/2018. Dr. Akhtar is a recipient of Office of National Intelligence National Intelligence Postdoctoral Grant # NIPG-2021–001 funded by the Australian Government.

REFERENCES

[1] H. Park, Z. J. Wang, N. Das, A. S. Paul, P. Perumalla, Z. Zhou, and D. H. Chau, "SkeletonVis: Interactive visualization for understanding adversarial attacks on human action recognition models," in *Proc. of AAAI*, vol. 35, no. 18, 2021, pp. 16 094–16 096. 1

[2] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, B. J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. Veloso, Z. L. Wang, and R. Wood, "The grand challenges of science robotics," *Science Robotics*, vol. 3, no. 14, p. eaar7650, 2018. 1

[3] H. Oinas-Kukkonen and M. Harjumaa, "Persuasive systems design: key issues, process model and system features 1," in *Routledge handbook of policy design*. Routledge, 2018, pp. 87–105. 1

[4] R. Liu, A. A. Ramli, H. Zhang, E. Henricson, and X. Liu, "An overview of human activity recognition using wearable sensors: Healthcare and artificial intelligence," in *Proc. of Internet of Things–ICIOT*. Springer, 2022, pp. 1–14. 1

[5] C. Li, Q. Zhong, D. Xie, and S. Pu, "Collaborative spatiotemporal feature learning for video action recognition," in *Proc. of CVPR*, 2019, pp. 7872–7881. 1

[6] R. Gao *et al.*, "Listen to look: Action recognition by previewing audio," in *Proc. of CVPR*. IEEE, 2020, pp. 10 457–10 467. 1, 6

[7] M. B. Shaikh and D. Chai, "RGB-D data-based action recognition: a review," *Sensors*, vol. 21, no. 12, p. 4246, 2021. 1

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. 1

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. 1

[10] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proc. of CVPR*. IEEE, 2016, pp. 770–778. 1

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 1

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 1, 4

[13] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012, doi:10.48550/arXiv.1212.0402. 1, 2

[14] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "MAiVAR: Multimodal audio-image and video action recognizer," in *Proc. of VCIP*. IEEE, 2022, pp. 1–5, doi:10.1109/VCIP56404.2022.10008833. 1, 3, 6

[15] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. of ICCV*, 2019, pp. 7083–7093. 1

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 2

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2, 4

[18] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proc. of ICCV*, October 2021, pp. 6836–6846. 2, 4

[19] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017. 2

[20] T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *Proc. of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998. 2

[21] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. of ICASSP*. IEEE, 2013, pp. 3687–3691. 2

[22] E. Kazakos *et al.*, "EPIC-Fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. of ICCV*, 2019, pp. 5492–5501. 2

[23] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "PyMAiVAR: An open-source python suite for audio-image representation in human action recognition," *Software Impacts*, p. 100544, 2023. 3

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021. 4

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014, doi:10.48550/arXiv.1412.6980. 4

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014. 4

[27] D. Tran, L. Bourdev, R. Fergu *et al.*, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. of ICCV*, 2015, pp. 4489–4497, doi:10.1109/ICCV.2015.510. 6

[28] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. of the ECCV*, 2016, pp. 20–36, doi:10.1007/978-3-319-46484-8_2. 6

[29] N. Takahashi, M. Gygli, and L. Van Gool, "AENet: Learning deep audio features for video analysis," *IEEE TMM*, vol. 20, no. 3, pp. 513–524, 2017. 6

[30] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. of ECCV*, 2018, pp. 247–263. 6

[31] M. Brousmiche, J. Rouat, and S. Dupont, "Multimodal attentive fusion network for audio-visual event recognition," *Information Fusion*, vol. 85, pp. 52–59, 2022. 6

[32] X. Long, G. De Melo, D. He, F. Li, Z. Chi, S. Wen, and C. Gan, "Purely attention based local feature integration for video classification," *IEEE TPAMI*, pp. 2140 – 2154, 2020. 6

[33] G. Yang *et al.*, "STA-TSN: Spatial-temporal attention temporal segment network for action recognition in video," *PloS one*, vol. 17, no. 3, pp. 1–19, 2022. 6

Semi-Supervised Anomaly Detection in Electronic-Exam Proctoring Based on Skeleton Similarity

1st Habibollah Agh Atabay
Image Processing & Data Mining Lab
Shahrood University of Technology
Shahrood, Iran
habib.atabay@shahroodut.ac.ir

2rd Hamid Hassanpour
Image Processing & Data Mining Lab
Shahrood University of Technology
Shahrood, Iran
h.hassanpour@shahroodut.ac.ir

***Abstract*—Anomaly detection in surveillance videos, particularly related to human behavior, is crucial for various applications. In electronic exams (e-exams), cheating can be detected using surveillance videos, but previous research mainly focused on predefined patterns, with less attention given to unsupervised methods. This study proposes a semi-supervised skeleton-based approach for abnormal behavior detection in e-exam proctoring videos. The proposed method segments the skeleton-based feature vectors of consecutive frames based on their similarity. The similarity is calculated based on the Euclidean distance between the mean feature vector and the standard deviation of segments. Similar segments of the training set, which is anomaly-free, are then combined to form distinct frame segments used to recognize the normal samples. Then, in the testing phase, a frame is recognized anomaly if it is not similar to any training segment, otherwise, it is considered normal. In addition, to get a better ranking of anomalous frames, we examine assigning a soft anomaly score considering the segment size and the maximum distance between the comparative features. In label assignment, the algorithm correctly categorizes the testing frame based on their similarity to the training segments. In comparison to the state-of-the-art reconstruction-based anomaly detection algorithms, the proposed method outperforms using the area under the ROC curve (AUC) metric. In addition to correctly detecting similar frames and segmenting videos, the advantage of the presented method is to determine the data labels without needing to process all inputs, making it possible for use in online applications.**

***Index Terms*—Unsupervised Anomaly Detection, Skeleton-based Features, E-Exam Cheating Detection, Skeleton Similarity Measurement**

I. Introduction

Anomaly Detection (AD) involves identifying instances and events that occur scarcely in the available training data. In other words, AD is the procedure of searching for concepts that are not yet visited. AD in videos is a challenging issue in computer vision, as identifying unusual events relies typically on contextual information and the surrounding environment, adding to the complexity of the task.

Unsupervised and semi-supervised learning methods are best suited to anomaly detection tasks. Unsupervised methods do not have any prior knowledge about data labels, and semi-supervised approaches learn from only limited normal samples of data. Their basis is the principle that normal events happen several times while abnormal events occur infrequently.

In surveillance videos related to human behavior, extracting body skeleton features provides a suitable solution for privacy protection, and it reduces the complexity of methods, especially when the sole purpose is to detect abnormal behavior in human actions. Skeleton-based methods [1] only focus on body joints and ignore facial identity, full body scan, or background information. So, these methods are insensitive to noise resulting from illumination, viewing direction, and background clutter, and they are released from the redundant burden of modeling the changes in those areas of the scene.

One of the applications of behavioral anomaly detection is in video proctoring of remote and electronic exams (e-exams) to detect unauthorized behaviors and cheating events, which is a challenging issue. Monitoring online examinations through human proctors is a common methodology to prevent cheating. However, the disadvantage of this method is the cost borne to employ individuals to monitor the exams. There is a high bandwidth requirement for communication in such cases, and there is no metric to evaluate the proctor's efficiency in cheating detection. Semi-automated proctoring has also been proposed in several research studies. In the recent review paper [2], the authors categorize the automated proctoring methods into six major groups, including Face Tracking, Face Expression Detection, Head Posture Analysis, Eye Gaze Tracking, Network Data Analysis and Traffic Classification, and IP Spoofing Detection, which there is no room for body-gesture-based methods. In [3] the authors incorporated the hand position along with head roll and yaw angles extracted by Microsoft Kinect camera. Similarly, other works in the literature have mainly relied on predefined rules and thresholding on extracted features such as head rotation angles [4]–[8]. Regrettably, when dealing with a large group of students having varying seating postures, this approach may not be as successful [2]. In Addition, in past research, abnormal behaviors have mostly been limited to specific predefined patterns, and unsupervised or semi-supervised methods have rarely been considered. Thus, in our viewpoint, a combination

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

of head and hand poses must be considered to recognize cheating in addition to unsupervised detection methods .

This article examines the detection of behavioral abnormalities among candidates in e-exam videos. Behavioral abnormalities refer to unusual behaviors exhibited by examinees that may indicate cheating when answering exam questions. The specific goal here is to detect behaviors that indicate cheating behind the scenes of the video. Since the type of abnormality is related to the candidate's body postures and movements, feature extraction based on the body skeleton has been emphasized. This approach not only facilitates the detection process, but also helps preserve participant's privacy. Additionally, the proposed method is a semi-supervised approach, since individuals' abnormal behaviors during exams vary greatly, while samples of normal behaviors can easily be collected.

The proposed method works by first dividing the frame sequence of the training videos into sequential segments during the training phase. Then, similar segments are merged to produce training clusters. To specify abnormality, we assign either a distinct label (zero or one) or soft score (from zero to one) to each frame. In the evaluation phase, the similarity between skeleton-based features of each frame and the segments recognized in the training set is measured, and the maximum distance of each frame features, and the training segments size are used to calculate the anomaly score. For assigning an anomaly label, we use standard deviation criteria such that if the distance of all skeleton features in the frame under observation to the corresponding features in the mean feature vector of a training cluster fall within the standard deviation range, that frame is considered normal (labeled zero).

The proposed method is evaluated on a dataset consisting of 91 mock e-exam videos. The initial part of each video is used as training data. Two approaches are considered for training. The Global approach uses the initial parts of all videos in the dataset as a training set for cluster extraction. The other is called Video-Specific, in which each video is processed separately. While the initial part of each video is used for training, the remaining part of the same video is used for testing. Finally, the results are compared with several semi-supervised anomaly detection algorithms. In label assignment, the detection precision of the normal samples and the recall of detecting abnormal samples are about 99%, that is the algorithm correctly detects the similar testing samples to the training clusters, but the precision of detecting ground-truth abnormal samples is still low (about 20%). In terms of soft scoring, the performance of the algorithm outperforms the compared reconstruction-based anomaly detection algorithms using the area under the ROC (AUC) metric. The algorithm can be used to summarize human-oriented videos based on the skeleton pose similarity. Another advantage of the presented method over other semi-supervised reconstruction-based anomaly detection methods is that it determines data labels without the need to process all inputs, which makes it possible to use it in online applications.

Section II, describes the collected dataset and extracted skeleton-based features. Section III proposes the skeleton-based video anomaly detection algorithm. IV presents the results of the experiments and Section V concludes the paper.

II. Dataset and Feature Extraction

For the experiments of this paper, a dataset was collected for the ultimate goal of detecting cheating in electronic exam videos. It includes 91 videos from participants in mock remote exams held for a maximum of 30 minutes. The videos of this dataset are taken from the side camera view and show the perfect situation of participants behind the computer desk. Since most similar works have used the laptop viewing angle (facing angle), the collected dataset is distinct because it includes the hand and upper body posture. The video frames were captured at a rate of 1 frame per second by the software implemented for conducting remote exams. The video frames were labeled based on the individual's condition in the scene concerning the presence of cheating. Sample frames from the collected dataset are shown in Fig. 1. The skeleton features extracted from the collected dataset and the implementation codes are available on GitHub[1]. Participants were asked to behave normally at the beginning of the exam so that it can be used as the training set for semi-supervised methods. The dataset can be expressed as follows.

$$Dataset = \{X^i | i \in 1..N^d\}, \tag{1}$$

where X^i is the frame features of the video i, and $N^d = 91$ is the number of dataset videos. X_i contains the initial part X_t^i and the remaining part X_v^i.

$$\begin{gathered} X^i = [X^{it}, X^{iv}], \\ X^{it} = \{F_k^i | k \in 0..N^{ti}\},\ X^{iv} = \{F_k^i | k \in N^{ti}..N^i\} \end{gathered} \tag{2}$$

where F_k^i is the skeleton-based feature vector of the frame k of the video i, and $N^{ti} \simeq 0.3 \times N^i$ denoating that approximately the initial 30% of each video frames is considered without any abnormalities. The MoveNet framework, implemented in Tensorflow [9], was used to extract the 2D skeleton features of the body. MoveNet extracts the 2D coordinates of each skeleton joint within the normalized range of [0, 1]. This study uses the 2D coordinates of upper body joints, including 11 positions of the nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, and right wrist. Therefore, the basic features for each frame form a vector of length 22. In addition to the 2D joint coordinates, this paper also experiments with simplified feature vectors, which consist of 6 features, including three significant points of the body: nose, right wrist, and left wrist. The idea behind this simplification is that perhaps for a human proctor, the location of the wrists may be more important than the hand muscle configuration. In other words, changing the hand position without moving the wrist cannot have a specific meaning. However, confirming the validity of this point requires further investigation and experience.

[1] https://github.com/habibatabay/skeleton_clustering_anomaly

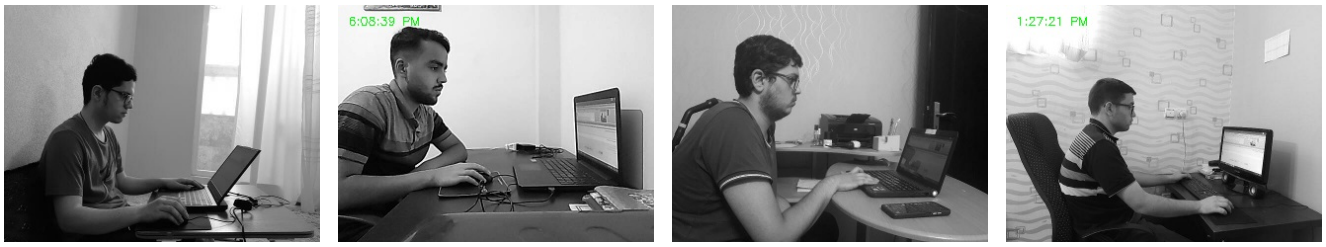
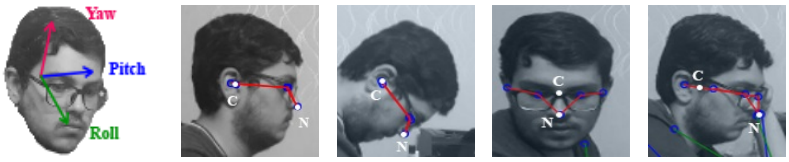


Fig. 1. Sample frames of the videos of the collected dataset



(a) (b) (c) (d) (e)

Fig. 2. (a) The rotation axes of the head, (b) to (e) show the position of the nose (N) and the center of gravity of the head (C). b) is a normal head state, and (c) to (e) are considered abnormal head poses. In (c) compared to (b), the position of N has changed both on the x-axis and in the y-axis (in 2D projected space). (d) The biggest change is observable in the x-axis, and in (e) a small difference is observable in the x-coordinate. (b) to (c) show that the changes in the coordinates of point N relative to C can indicate the change in the head pose.

Regarding the head, since the head posture or gaze direction is of particular importance, raw nose coordinates are not sufficient to show the head posture. For this reason, the position of the nose with respect to the head's center of gravity is considered a feature. With this modification, the nose position along the x-axis can indicate changes in the yaw angle, and changes in the pitch angle can be indicated by the nose position in height (y-axis) with respect to the center of gravity. The roll angle is less computable due to the camera's placement, and of course, this angle is less important in detecting cheating. Fig. 2 shows the head angles and postures in three different positions. The average position between the two ear joints is used to calculate the head's center of gravity.

III. The Proposed Method

Fig. 3 illustrates the overall process of the proposed anomaly detection method. After extracting skeleton features from each video frame, consecutive frames with similar skeleton poses are grouped together (Fig. 3.a). The obtained segments from the training video are merged based on their similarities (Fig. 3.b), and distinct poses are identified (Fig. 3.c). To detect anomalous segments in the test video, its frames are first segmented like the training video (Fig. 3.d), and segments that do not have similarities to the identified segments in the training phase are considered as anomalous segments. The core idea of the proposed method is the similarity measurement between skeleton poses, which is calculated in two modes: Label and Score. In the Label mode, we determine the similarity with a binary label of either zero or one. However, in the Score mode, we assign a score between zero and one. The similarity label, SL_{ij}, between two segments, i and j, can be expressed as follows:

$$SL^{ij} = \begin{cases} 1 & : D_k^{ij} < (SD_k^i + sd) \; or \; D_k^{ij} < (SD_k^j + sd) \\ & \qquad \forall k \in 1..N^f \\ 0 & : otherwise \end{cases}, \tag{3}$$

in which $D^{ij} = |M^i - M^j|$ is the distance vector of two mean feature vectors (M^i and M^j), N^f is the number of features, SD_k^i is the standard deviation of the k_{th} feature of the segment i, and sd is an added value of standard deviation. The added value of standard deviation is considered to prevent the standard deviation of a segment from being zero (when it contains just one frame) and also increase the probability of being similar.

In the scoring mode, we use the maximum difference between the feature values of two average skeletons as the difference score. We also affect the size of the anomalous cluster to alleviate the score. The similarity score can be formulated as follows.

$$SS^{ij} = w^i(1 - Max_{k \in 1..f}(|M_k^i - M_k^j|)), \tag{4}$$

where w_i is the size of the cluster i with respect to the maximum size of the detected clusters in the video, and M_k^i is the k_{th} feature of the mean feature vector of the segment.

Based on the above similarity measurement, the steps of the proposed anomaly detection method are as follows: Let X^t be the training set of frames.

1) The first step is segmenting the training set. Then we have a set of training segments S^t as:

$$S^t = \{S_k^t \mid k \in 1..N^{ts}\} \tag{5}$$

where N^{ts} is the number of training segments, each contains similar frames:

$$S_k^t = \{F_a^t, ...F_b^t \mid a, b \in 0..N^{f^{tk}} \; and \; a \leq b\} \tag{6}$$

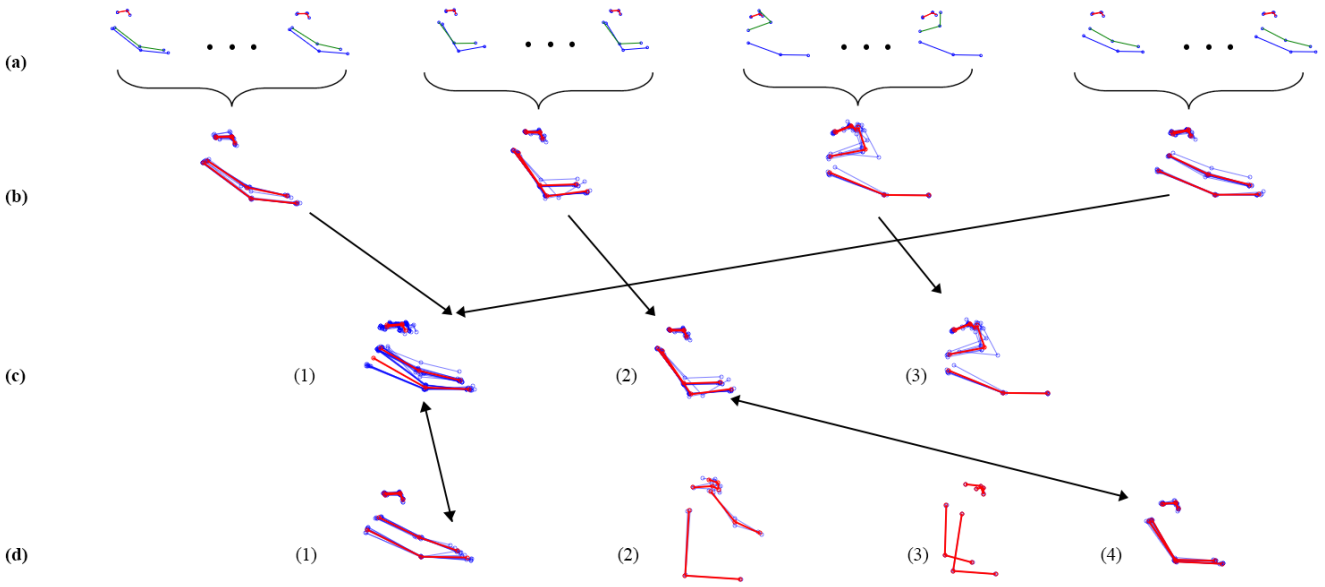


Fig. 3. The overall process of the proposed anomaly detection method: (a) segmentation of a training video based on the similarity of skeleton poses, (b) average skeleton for each segment, (c) merging similar segments, and (d) comparing the detected segments in the test video with the detected segments in the training phase. Segments (d.1) and (d.3) are recognized as normal since they correspond to similar segments (c.1) and (c.2), respectively, while segments (d.2) and (d.3) can be considered as anomalous segments.

where

$$SL^{G(i)G(i+1)} = 1 \mid a \leq i < c, \qquad G(i) = \{F_i\}, \tag{7}$$

where $G(i)$ is a single-frame segment consists of the frame number i. We want to arrange the frames in a segment so that the body state is not changed much, and whenever the body state changes, another frame segment is formed.

2) The similarity between non-sequential segments is determined in the training dataset, and similar segments are merged to avoid multiple comparisions in the testing phase.
3) Finally, the similarity between the testing frames and the training segments is determined similar to (7), and the similarity label or score is calculated for each one.

IV. Results

Given the collected dataset of e-exam videos, we try to detect human behavioral anomalies in a semi-supervised manner. We first separate the initial 30% of each video of the dataset as the training set and the remaining part as the testing set. For training, we examine two schemes of training: Global and Video-Specific. In the first training scheme, we consider recognized clusters of all training videos as a single training set of patterns. But in the Video-Specific method, we use the detected normal patterns of the initial part of a video to test the remaining part of the same video. In the Global mode, the goal is to use the normal patterns observed from all participants to detect abnormal patterns. On the other hand, in the second approach, each video is processed separately considering that the angle of each record is likely to be different from each other and also the normal behavioral patterns of different individuals may differ from each other.

In Global training mode, it is important to normalize the range of features of the skeletons because two skeletons can have the same pose but with different scales. Although the camera angle is also an important factor in shaping a specific pose, which cannot be avoided in 2D skeleton features, normalizing the skeletons can reduce the differences in poses with similar angles of view. To normalize the skeleton features, we calculate the position of the box surrounding the skeleton and rescale each joint position according to the bottom-left and top-right positions of the bounding box. Normalization is also important in video-specific mode when the candidate changes their position towards the camera or the camera's position changes during video capture.

The performance of the proposed algorithm is compared to several semi-supervised anomaly detection methods based on deep learning and data reconstruction, including Univariate and multivariate Auto-Encoders [10], LSTM-ED [11], TCN-ED [12], and VAE-LSTM [13]. The main parameters of these methods are the sequence length and the size of the hidden space, both of which were set to 15 and 10, respectively, for all of them. The proposed algorithm also has hyperparameters such as the added value of standard deviation, which is required to form a cluster. Based on observations and comparisons of the created pieces and examination of the results, a value of 0.05 was chosen for this parameter. AUC and AP metrics were used to compare the methods.

TABLE I
GLOBAL RESULTS USING THE SCORE SIMILARITY MEASUREMENT

Anomaly Detection Methods	Original		Simplified	
	AP	***AUC***	***AP***	***AUC***
UAE [10]	**0.57**	0.77	**0.57**	**0.77**
AE [10]	0.47	0.63	0.46	0.66
LSTMED [11]	0.49	0.67	0.44	0.60
TcnED [12]	0.36	0.52	0.31	0.45
VAE-LSTM [13]	0.34	0.52	0.30	0.51
Ours	0.43	**0.81**	0.35	0.71

TABLE II
VIDEO-SPECIFIC RESULTS USING THE SCORE SIMILARITY MEASUREMENT

Anomaly Detection Methods	Original		Simplified	
	AP	***AUC***	***AP***	***AUC***
UAE [10]	**0.51**	0.68	0.42	0.59
AE [10]	0.49	0.64	**0.47**	0.62
LSTMED [11]	0.43	0.57	0.32	0.52
TcnED [12]	0.51	0.71	0.40	0.58
VAE-LSTM [13]	0.37	0.55	0.29	0.50
Ours	0.44	**0.83**	0.39	**0.79**

Tables I and II show the results of the anomaly detection algorithms using the Global and Video-Specific training approaches and the soft scoring method. Table I shows that in the Global training methods and the soft scoring, our methods could not outperform the best previous method, which is UAE considering all metrics except AUC for the original feature type. The quality of frame-level scoring is degraded when the original coordinate features are replaced with simplified features. The score of our method in terms of the AUC metric is much better than AP. This could be because the proposed algorithm does not score individual frames well. The AP metric is sensitive to the final data ranking, indicating this issue. However, the proposed algorithm performs well compared to others concerning the AUC metric, which is less sensitive to data ranking. On the other hand, in Table II, the results of our algorithm are much better than the compared methods in terms of AUC metric. The same trends between the original and simplified features can be seen in Table I shows. In both tables, the scoring worked best for the original features. Another point is that our methods could not be improved when we use the global training approach, which can be attributed to the variations in camera views of dataset videos. It probably prevents the extension of the pool of known normal patterns for each video. The last notable point is that the results of our algorithm are not fully compatible with the results of the compared methods, because those methods use a fixed window size over the stream of frames but in our method, we do not have this.

To compare the performance of our method assigning the strict anomaly labels, we extract the best threshold from the scores generated by each compared algorithm and determine either zero or one label for each test sample. Specifically using the AUC curve, we find the threshold t so that:

$$t \;=\; Max_{i\in[0..T)}(TPR_i - FPR_i) \tag{8}$$

where T is the number of thresholds. Then we calculate metrics of classification accuracy, precision and recall of normal samples, and precision and recall of abnormal samples. The comparative results of those metrics are shown in Tables III and IV. Table III shows binary classification performance using our method in video-specific mode. If we consider the balance between precision and recall of the anomalous samples, the quality of our method seems lower than other methods. But it reveals interesting results. First, the classification quality by simplified feature seems higher than the original features. Next, in our method, the precision of detecting normal samples and recall of abnormal samples are very high It shows that the samples labeled normal in our method are 99% normal. On the other hand, using our method, most abnormal samples are labeled anomaly. Table IV shows a similar performance to Table III with a few improvements in results using simplified features.

V. CONCLUSION

This paper presented a semi-supervised method for detecting behavioral anomalies in e-exam videos. It provides a similarity criterion for grouping similar skeletons and segmenting videos. By segmenting the training video frames and calculating the similarity of the test frames to them, it can determine the abnormality of the test frames. The experimental results show that the proposed algorithm can easily identify the states of the body skeleton that do not exist among the patterns of the learning set. The proposed algorithm is used in two general and video-specific educational modes. The algorithm's quality in the video-specific mode compared to the Global mode shows that if enough normal data samples are available to analyze a person's behavior, using this method in the Video-Specific mode will produce good results. Another advantage of this method is that it can be used to summarize videos based on the similarity of individuals' skeletons because some sample frames from each segment can be included in the summary video. Also, because this method does not need to see other testing samples to label a sample, it can be used in online applications for anomaly detection.

REFERENCES

[1] P. K. Mishra, A. Mihailidis, and S. S. Khan, "Skeletal video anomaly detection using deep learning: Survey, challenges and future directions," *arXiv preprint arXiv:2301.00114*, 2022.

[2] S. Kaddoura, S. Vincent, and D. J. Hemanth, "Computational intelligence and soft computing paradigm for cheating detection in online examinations," *Applied Computational Intelligence and Soft Computing*, vol. 2023, p. 3739975, May 2023.

[3] Z. Fan, J. Xu, W. Liu, and W. Cheng, "Gesture based misbehavior detection in online examination," in *2016 11th International Conference on Computer Science & Education (ICCSE)*, pp. 234–238, 2016.

TABLE III
THE PERFORMANCE OF METHODS ASSIGNING STRICT ANOMALY LABELS IN THE VIDEO-SPECIFIC MODE

Anomaly Detection	**Original**					**Simplified**				
Methods	***ACC***	***PreN***	***RecN***	***PreA***	***RecA***	***ACC***	***PreN***	***RecN***	***PreA***	***RecA***
UAE [10]	**0.51**	0.68	0.70	**0.83**	0.73	0.52	0.63	0.42	0.59	0.69
AE [10]	0.49	0.64	0.69	**0.83**	0.72	0.53	0.61	0.47	**0.62**	0.70
LSTMED [11]	0.43	0.57	0.69	0.80	0.76	0.51	0.50	0.32	0.52	0.60
TcnED [12]	**0.51**	0.71	**0.72**	**0.83**	0.74	0.49	0.64	0.40	0.58	0.66
VAE_LSTM [13]	0.37	0.55	0.62	0.77	0.71	0.41	0.45	0.29	0.50	0.51
Ours	0.30	**0.99**	0.20	0.14	**0.99**	**0.61**	**0.97**	**0.57**	0.21	**0.87**

TABLE IV
THE PERFORMANCE OF METHODS ASSIGNING STRICT ANOMALY LABELS IN THE GLOBAL MODE

Anomaly Detection	**Original**					**Simplified**				
Methods	***ACC***	***PreN***	***RecN***	***PreA***	***RecA***	***ACC***	***PreN***	***RecN***	***PreA***	***RecA***
UnivarAutoEncoder	**0.72**	0.86	0.71	0.50	0.75	0.72	0.86	0.71	**0.49**	**0.73**
AutoEncoder	0.71	0.81	**0.77**	**0.53**	0.55	0.70	0.81	0.74	0.46	0.57
LSTMED	0.70	0.82	0.73	0.50	0.61	0.70	0.80	0.78	**0.49**	0.49
TcnED	0.64	0.79	0.73	0.52	0.45	0.63	0.77	0.73	0.44	0.39
VAE_LSTM	0.61	0.77	0.69	0.35	0.43	0.56	0.75	0.61	0.32	0.48
Ours	0.31	**0.99**	0.22	0.14	**0.99**	**0.77**	**0.92**	**0.80**	0.30	0.48

[4] M. Labayen, R. Vea, J. Flórez, N. Aginako, and B. Sierra, "Online student authentication and proctoring system based on multimodal biometrics technology," *IEEE Access*, vol. 9, pp. 72398–72411, 2021.

[5] S. Hu, X. Jia, and Y. Fu, "Research on abnormal behavior detection of online examination based on image information," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 02, pp. 88–91, 2018.

[6] B. Yang, H. Li, H. Xie, J. Zhao, R. Zhu, and L. Zhao, "Abnormal state recognition method for online intelligent examination based on improved genetic algorithm," *International Journal of Information and Communication Technology*, vol. 18, no. 3, pp. 334–350, 2021.

[7] H. Li, M. Xu, Y. Wang, H. Wei, and H. Qu, "A visual analytics approach to facilitate the proctoring of online exams," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2021.

[8] A. Tweissi, W. Al Etaiwi, and D. Al Eisawi, "The accuracy of ai-based automatic proctoring in online exams," *Electronic Journal of e-Learning*, vol. 20, no. 4, pp. 419–435, 2022.

[9] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: a system for large-scale machine learning.," in *Osdi*, vol. 16, pp. 265–283, Savannah, GA, USA, 2016.

[10] A. Garg, W. Zhang, J. Samaran, R. Savitha, and C.-S. Foo, "An evaluation of anomaly detection and diagnosis in multivariate time series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2508–2517, 2021.

[11] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.

[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[13] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.

CD-COCO: A Versatile Complex Distorted COCO Database for Scene-Context-Aware Computer Vision

Ayman Beghdadi
IBISC Lab, Paris-Saclay University
Evry, France
aymanaymar.beghdadi@univ-evry.fr

Azeddine Beghdadi
L2TI Lab, Sorbonne Paris Nord University
Villetaneuse, France
azeddine.beghdadi@univ-paris13.fr

Malik Mallem
IBISC Lab, Paris-Saclay University
Evry, France
malik.mallem@univ-evry.fr

Lotfi Beji
IBISC Lab, Paris-Saclay University
Evry, France
lotfi.beji@univ-evry.fr

Faouzi Alaya Cheikh
Norwegian University of Science and Technology (NTNU)
Gjovik, Norway
faouzi.cheikh@ntnu.no

***Abstract*—The recent development of deep learning methods applied to vision has enabled their increasing integration into real-world applications to perform complex Computer Vision (CV) tasks. However, image acquisition conditions have a major impact on the performance of high-level image processing. A possible solution to overcome these limitations is to artificially augment the training databases or to design deep learning models that are robust to signal distortions. We opt here for the first solution by enriching the database with complex and realistic distortions which were ignored until now in the existing databases. To this end, we built a new versatile database derived from the well-known MS-COCO database to which we applied local and global photo-realistic distortions. These new local distortions are generated by considering the scene context of the images that guarantees a high level of photo-realism. Distortions are generated by exploiting the depth information of the objects in the scene as well as their semantics. This guarantees a high level of photo-realism and allows to explore real scenarios ignored in conventional databases dedicated to various CV applications. Our versatile database offers an efficient solution to improve the robustness of various CV tasks such as Object Detection (OD), scene segmentation, and distortion-type classification methods. The image database, scene classification index, and distortion generation codes are publicly available [1].**

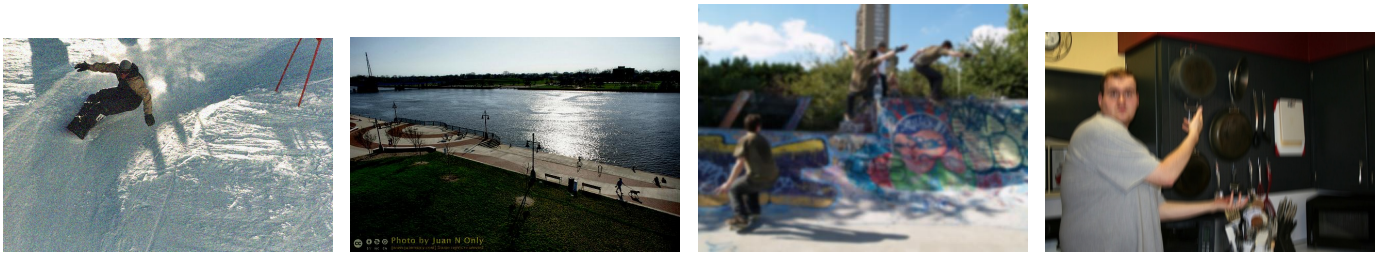
***Index Terms*—Dataset, Deep learning, Depth, Distortion, Object detection, Scene analysis, Segmentation**

I. Introduction

The interest in making databases available to the scientific community is becoming more and more important with the development of data-driven approaches, and in particular those based on deep neural network architectures. Few studies have been conducted to analyse the relevance and reliability of databases in the field of CV. However, we can point out some interesting studies where some attributes and descriptors have been introduced to measure the representativeness and the richness of the databases dedicated to the evaluation of image and video quality metrics [1], [2]. To the best of our knowledge, there have been no similar efforts to design realistic databases dedicated to improve methods developed for solving problems in the field of CV. Here, we are interested in the detection or segmentation of objects in an uncontrolled environment and under various constraints related to the image acquisition conditions. OD is still a hot topic and many methods have been proposed during these last two decades [3], [4]. However, the impact of the distortions on the performance of the proposed OD solutions was often neglected apart a few studies limited to object recognition and image classification under specific distortions (noise and blur) [5] and OD under photometric and geometric distortions [6]. A previous study [7] highlighted the distortion impact on the OD performance through global and local distortions without any scene context consideration have been achieved, which proved the usefulness of data augmentation by using a distorted database to improve OD models robustness. Consequently, we propose a novel distorted image database with complex and photorealistic distortions. This database offers the diversity and quality of distortions necessary for designing robust deep-learning models, in particular OD models. For this, we introduced the local and realist atmospheric distortions in our database. Unlike the classic so-called global distortions applied to the entire image, local distortions apply to defined areas. Local distortions correspond to the local representation of distortions resulting from scene conditions due to object motion or position in the scene, such as motion blur from moving objects, defocus blur and backlight phenomena. The proposed atmospheric distortions attempt to better replicate the natural rain and fog phenomena by applying these distortions in a non-homogeneous manner. These new distortions consider scene context through scene depth and object annotation from MS-COCO's ground truth for better photorealism. Furthermore, a manual annotation of the original COCO database was done to guide the choice of the distortion to be applied automatically to each image. In addition, a scene classification (indoor/outdoor) was performed to automatically manage the distortion intensity according to the type of scene. The main contributions of our study are summarized as follows:

[1] https://github.com/Aymanbegh/CD-COCO

979-8-3503-4218-5/23/31.00 2023 European Union



(a) Gaussian noise (b) Contrast changing (c) Global defocus blur (d) Global motion blur

Fig. 1: Some examples of global distortions.

- New distortions with improved realism are introduced, describing common phenomena in computer vision through complex local and atmospheric distortions.
- This paper proposes efficient algorithms to generate local and global photorealistic distortions that are not included in any existing database.
- A novel dataset is built from the MS-COCO dataset, dedicated to the improvement of the robustness of the OD and object segmentation models against a broad type of distortion.
- The image database, the proposed database scene classification index, and distortion generation codes are publicly available.

The remainder of the paper is organized as follows. Section II summarizes previous related literature. Section III is devoted to detail the methods of generating complex distorted images. Then, section IV is dedicated to show dataset details. Finally, conclusions and perspectives are provided in section V.

II. Related Work

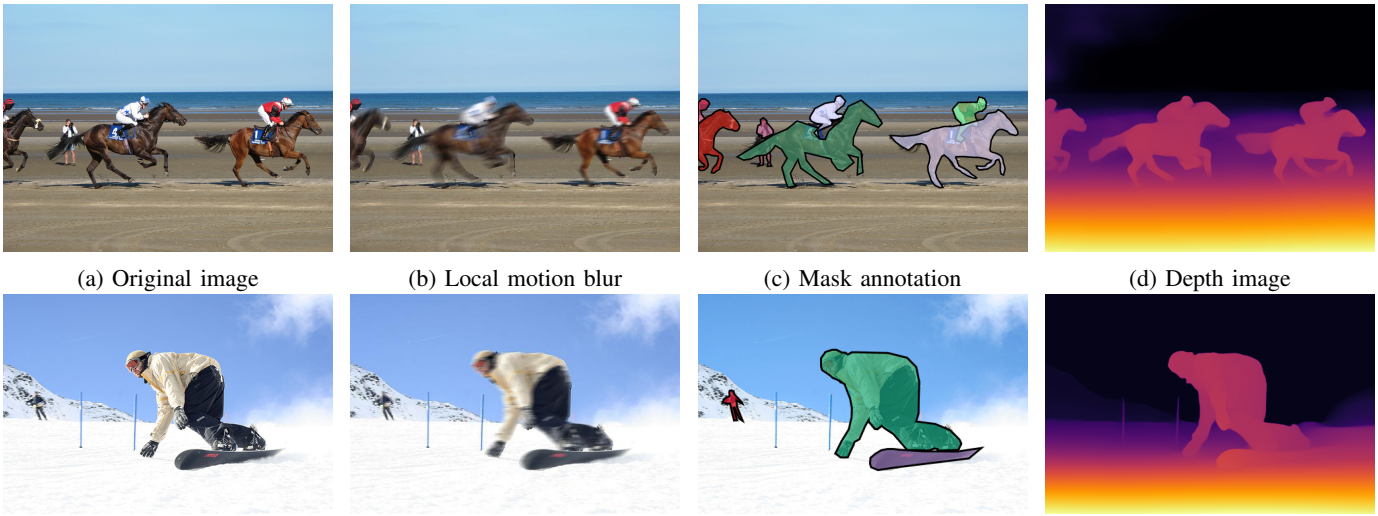
Object detection in video sequences or still images is a research topic of great interest given the numerous applications in the computer vision field [8], [9] and especially in video surveillance [10]. With the development of deep learning methods and the availability of many databases dedicated to this problem, this field of research has seen a real progress. A comprehensive survey on deep learning based OD approaches is provided in [11]. However, most of the available databases do not consider real-world scenarios, especially images and videos captured in uncontrolled environments, which are affected by various types of distortions. In fact, many studies have shown that OD performance is strongly influenced by the quality of the images [5], [6], [12], [13], [7]. It is worth noticing that the number and types of distortions considered in these studies and the existing dedicated dataset are limited. Furthermore, the case of multiple distortions appearing simultaneously has not been taken into account in OD performance evaluation studies. Multiple distortion scenarios have been considered in a few studies on video quality assessment but in limited contexts [14], [15], [16]. Some interesting studies investigated the impact of various distortions on the performance of CNN-based OD architectures [17], [18]. However, all these studies are limited to a few distortions and do not consider local distortions that really correspond to real scenarios. Indeed, if we take, for example, the blur caused by movement, it is usually simulated in a global way in the existing databases. While we know that in an observed scene there can be objects moving at different speeds and in different directions and therefore affected by blurs of different amplitudes and directions. The same applies to the defocusing blur, which depends on the depth of the objects in the filmed scene. In our database we have taken into account these aspects and others such as lighting effects that vary with the depth and geometry of objects. We have adopted the same approach concerning the distortions due to atmospheric phenomena such as rain and fog. Taking into account these aspects is not simple and it is one of the main originalities of our contribution.

III. Complex distortion generation algorithms

Well-known global distortions have been applied to our database through classical distortions methods. In our case, global distortion refers to the classic distortions that apply more or less homogeneously to the entire image, regardless of the context of the scene. Thus, we applied global distortions for some images resulting from image acquisition (noise, compression, contrast changing) or camera (motion and defocus blur) conditions without considering the scene context (see fig.1). However, some images have specific scene contexts that require the application of local or atmospheric distortions using more sophisticated approaches. Our generated complex distortions use scene depth information, ground truth information from COCO annotations (object masks), and object and scene type to produce complex and photorealistic distortions. Scene depth information is obtained using the MiDaS depth estimation model [19].

A. Local motion blur

Local motion blur is a local application of motion blur phenomena to the annotated objects. It represents the cases of image acquisition where objects move rapidly in front of the camera. This local distortion requires the ground truth masks to define the pixel area where the blur motion is to be applied. Furthermore, the object mask is also used to determine the distortion orientation through a strategy specific to the nature of the object (object classes). Another dual strategy allows us to compute the motion magnitude applied to each object in the images. First, an interval of motion magnitude



(a) Original image (b) Local motion blur (c) Mask annotation (d) Depth image

(e) Original image (f) Local motion blur (g) Mask annotation (h) Depth image

Fig. 2: Illustration of the local motion blur.

is derived from the nature of the object and prior knowledge about the speed of the object type. Then, a magnitude value is computed by considering the object's depth and the others. this value into the global scene context. Thus, each magnitude value is obtained by correlating the nature and depth of objects, ensuring the global consistency of each local blur motion distortions relative to each other. Object orientation is obtained by computing the angle between the X-axis and the ellipse's major axis containing the object. Then, a checking strategy of the orientation is adapted to apply a motion blur according to the object's nature. Furthermore, a checking of object interaction is achieved to prioritize the magnitude and orientation of higher-level objects on lower ones as shown in fig.3. This is done by correlating their depth proximity and

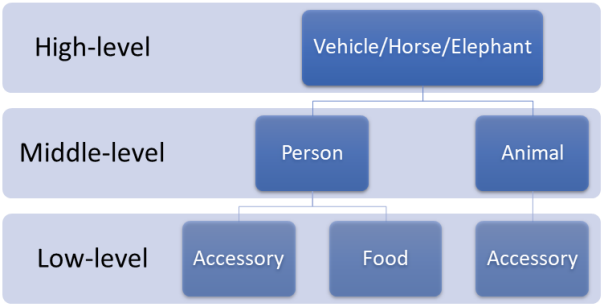


Fig. 3: Interaction hierarchy related to the object type

their bounding box overlap to ensure distortion consistency for linked objects. Thus, the magnitude and orientation of higher-level objects are applied to lower-level objects with which they are interacting. The complete algorithm follows the following steps:

1) Find the scene context: ski, riding, sport, skate or surf depending on present objects in the image.
2) Object classification: create object superclasses by grouping objects together to think globally (vehicle, person, animal, food, etc...).
3) Compute the average depth of each annotated objects.
4) Calculate amplitude and orientation using depth and object type to distort each object individually.
5) Find interactions between objects by correlating their depth proximity and their overlapping bounding boxes to apply the same distortion to interacted objects.
6) Sort the objects according to their depth to adjust their motion amplitude for a global consistency of the scene and distortions.

B. Local Defocus blur

Local defocusing blur results from focusing only on only the background orforeground. To create a realistic defocus blur, we used successive smooth thresholding to create three distinct areas related to scene depth. This thresholding process is performed using a nonlinear smooth function Ω expressed as:

$$\Omega(x) = 1 - \frac{1}{1 + \exp -15(x - 0.5)} \quad (1)$$

Where x represents the keypoint depth normalized by the average depth of the closest object as follows:

$$x = \frac{threshold - p_i}{threshold} \quad (2)$$

Figure 5b illustrates the image splitting through the smooth thresholding of the scene depth to get the three different grounds. Foreground corresponds to depths with threshold coefficients higher than the high threshold, middle-ground to coefficients between the high and low thresholds, and the background for coefficients lower (see fig.5a). Then, the average depths δ, δ_m, and δ_b of the three grounds are computed to perform a proportional defocus blur related to the depth. We applied a cumulative defocus blur magnitudes λ, λ_m, and λ_b

(a) Original image (b) Local defocus blur (c) Mask annotation (d) Depth image

(e) Original image (f) Local defocus blur (g) Mask annotation (h) Depth image

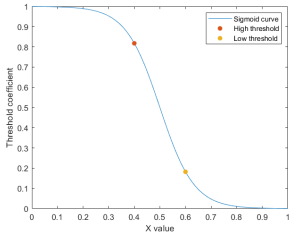
Fig. 4: Illustration of the local defocus blur.

going from foreground to background on each zone, expressed as follows:

$$\lambda = 0.5 + \frac{\delta_f - threshold}{threshold} \cdot 1.5 \tag{3}$$

$$\lambda_m = \lambda + \frac{\delta_m - threshold}{threshold} \cdot 1.2 \tag{4}$$

$$\lambda_b = \lambda_m + \frac{\delta_b - threshold}{threshold} \cdot 1.2 \tag{5}$$



(a) Sigmoid curve for smooth thresholding

(b) Threshold related to the depth

Fig. 5: Smooth thresholding related to the depth.

The aread bounded by the masks are distorted according to their corresponding defocus blur magnitude (λ, λ_m and λ_b), then fused to get the complete distorted image I_d. as shown in Fig. 4. The proposed local defocus blur algorithm is described in the algorithm 1.

C. Atmospheric distortion: the rain

Synthesizing the rain homogeneously, without any scene depth consideration, lacks realism. Indeed, the size and density of the rain depends on the distance from which it falls. Thereby, our rain generation algorithm used the method from algorithm 1 for performing a scene depth classification into foreground, middle-ground, and background. Each ground is assigned a rain intensity level that replicates the rain density. Note that the rain masks are obtained from images of flowing water like rain produced under experimental conditions.

Algorithm 1 Locale defocus blur algorithm

Input: Image I, Keypoints depth p_i
Output: Distorted Image I_d

Find the closest object depth: threshold
High threshold th_f = 0.8176
Low threshold th_b = 0.182
for each $p_i \in I$ **do**
  $\sigma = \frac{threshold - p_i}{threshold}$
  $\Delta(p_i) = 1 - \frac{1}{1+\exp(-15(\sigma - 0.5))}$
  if $threshold > p_i$ **then**
    $Foreground \leftarrow p_i$
  else if $th_f <= \Delta(p_i)$ **then**
    $Foreground \leftarrow p_i$
  else if $th_b <= \Delta(p_i)$ **then**
    $Middleground \leftarrow p_i$
  else if $th_f >= \Delta(p_i)$ **and** $th_b > \Delta(p_i)$ **then**
    $Background \leftarrow p_i$
  end if
  $\delta_f \leftarrow$ Foreground average depth
  $\delta_m \leftarrow$ Middleground average depth
  $\delta_b \leftarrow$Background average depth
end for

We extract three rain densities from these rain masks by performing some erosion and dilation processes. These three rain sub-masks are applied for each ground according to a random constant α of blending, achieving image blending as shown in fig.6. It is worth noticing that the rain sub-masks are applied cumulatively from the foreground to the background as follows:

$$I_f = 1 - ((1 - I) \cdot (1 - (\alpha \cdot R_f))) \tag{6}$$

$$I_m = 1 - ((1 - I_f) \cdot (1 - (\alpha \cdot R_m))) \tag{7}$$

$$I_d = 1 - ((1 - I_m) \cdot (1 - (\alpha \cdot R_b))) \tag{8}$$

Where I, I_f, I_m and I_d are the original, foreground, middle-ground and final distorted images respectively. Likewise, R_f, R_m, and R_b are the three rain sub-masks. Thus, this approach

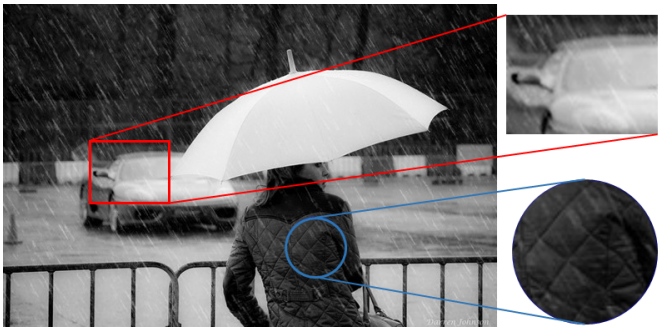
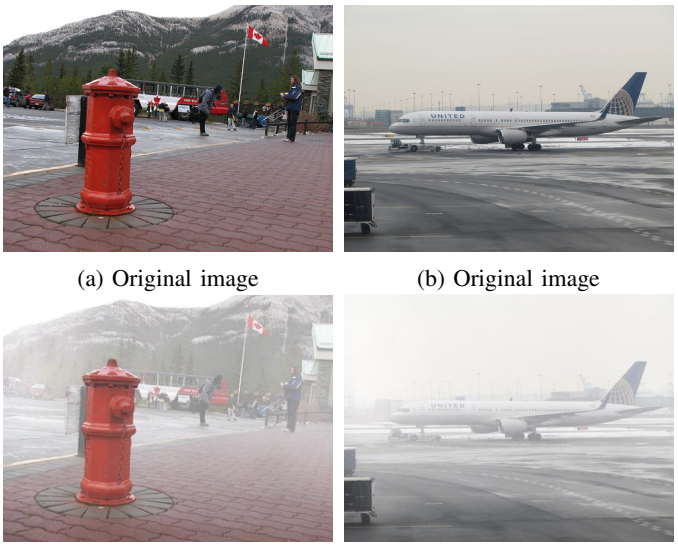


Fig. 6: Rain distortion example

applies only the fine rain corresponding to the distant rain stream to the distant parts of the scene for better realism, as shown in Fig.6. Our complex atmospheric rain improves the global consistency of distortion by considering the spatial relationship between scene depth and rain phenomena.

D. Atmospheric distortion: the fog

Generating synthetic fog is a complex task. Indeed, no mathematical fog model could have been used for generating synthetic fog. Thus, we opted to use fog masks extracted from images of experimental creations of fog with black background.



(a) Original image (b) Original image

(c) Fog distortion (d) Fog distortion

Fig. 7: Examples of atmospheric distortion: the fog.

Many masks have been extracted to provide a large fog sample with diverse densities and forms. These masks are applied to the original images, seamlessly blending through a mask adjustment. However, applying a mask homogeneously produce a non-realistic fog. Considering the scene depth for applying the mask to match the thick fog effect in real cases seems crucial. Thereby, we carry out this mask H with a variable factor $\kappa(i,j)$ proportional to the normalized depth $Depth_n(i,j)$ of each image pixel $I(i,j)$ and a constant value α as summarized in algorithm 2.

Algorithm 2 Fog generation algorithm

Input: Image I, fog mask H
Output: Distorted Image I_d

α = 0.95
for each pixel $i,j \in I$ **do**
$Depth_n(i,j) = \frac{Depth(I(i,j))}{Depth_{max}}$
$\kappa(i,j) = \alpha \cdot Depth_n(i,j)$
$I_d(i,j) = (1-(1-I(i,j)) \cdot (1-(\kappa(i,j) \cdot H(i,j))) \cdot 255$
end for

To give the images generated more photo-realism, the thickness of the fog is adapted to the depth of the observed scene, reproducing the effect of fog accumulation, as shown in figure 7.

E. Local Backlight

Local backlight distortion is generated by applying a local contrast enhancement process to the luminance component by using the object segmentation mask. This pixel-wise intensity transformation takes into account the position of the light source and that of the illuminated object on which the effect is to be brought out. This operation, which is nothing more than tone mapping, is applied to three preselected intensity intervals, semi-automatically and randomly. Figure 8 illustrates this type of photometric distortion.

(a) Original image (b) Distorted image

Fig. 8: Local backlight distortion

IV. Dataset

The generated dataset consists of more than 123K images with 80 object classes organized in three sets: 95K, 5K and 23K images for train, validation and test sets respectively. The ground truth annotation provides the objects' classes, bounding boxes, and masks for each image, which can be used for training object detection, and segmentation models.

A. Distortions

Our distorted dataset is composed of ten distortion types, 5 global distortions, 2 global atmospheric distortions and 3 local distortions. In order to generate the different distortions in a coherent and relevant way, a first scan of all the images is performed to prepare the distortion assignment protocol according to the semantic content of the scene and the context. The different distortions are automatically applied to the images previously annotated during the first process. Images annotated as global distortions are then distorted by one of the global distortion types chosen randomly (see table I).

TABLE I: Distribution of distortions.

Distortion type	Number of images	Ratio
Compression artefact	17989	15.3%
Contrast changing	18038	15.4%
Gaussian noise	18055	15.4%
Global motion blur	18018	15.3%
Global defocus blur	17792	15.1%
Fog	787	0.7%
Rain	845	0.7%
Local Backlight	296	0.3%
Local defocus blur	7061	6.0%
Local motion blur	18625	15.9%

B. Scene classification

The observed scenes are classified into indoor and outdoor scenes based on the context of the images. Indoor scenes were attributed to scenes where most information is included in indoor environments (room, building, hall, vehicle interior, etc.). Conversely, outdoor scenes correspond to open environments. The table II summarises the scene classification of our dataset.

TABLE II: Scene classification.

Scene type	Number of images
Indoor scene	45884
Outdoor scene	72404
Skiing scene	4434
Surfing scene	3635
Skating scene	3603
Sport scene	11965

V. Conclusion

In this study, we presented novel local and global complex distortions generated by reliable algorithms considering the scene context to achieve a high level of photo-realism. The proposed database will improve not only OD algorithms but also many scene analysis, classification and image segmentation methods, providing a more complete and beneficial framework for deep learning-based methods. As a perspective, it would be interesting to enrich this database with other distortions and in particular those related to atmospheric perturbations such as the heat diffusion effect and pollution. Another aspect that could be considered in the future is to incorporate pose object estimation when applying distortion.

References

[1] Stefan Winkler. Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):616–625, 2012.

[2] Muhammad Ali Qureshi, Azeddine Beghdadi, and Mohamed Deriche. Towards the design of a consistent image contrast enhancement evaluation measure. *Signal Processing: Image Communication*, 58:212–227, 2017.

[3] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.

[4] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[5] Tejas S Borkar and Lina J Karam. Deepcorrect: Correcting dnn models against image distortions. *IEEE Transactions on Image Processing*, 28(12):6022–6034, 2019.

[6] Zhongqi Lin, Zengwei Zheng, Jingdun Jia, Wanlin Gao, and Feng Huang. Ml-capsnet meets vb-di-d: A novel distortion-tolerant baseline for perturbed object recognition. *Engineering Applications of Artificial Intelligence*, 120:105937, 2023.

[7] Ayman Beghdadi, Malik Mallem, and Lotfi Beji. Benchmarking performance of object detection under image distortions in an uncontrolled environment. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2071–2075. IEEE, 2022.

[8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.

[9] Constantine P Papageorgiou and Tomaso Poggio. A trainable object detection system: Car detection in static images. 1999.

[10] Azeddine Beghdadi, Muhammad Asim, Noor Almaadeed, and Muhammad Ali Qureshi. Towards the design of smart video-surveillance system. In *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, pages 162–167. IEEE, 2018.

[11] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE access*, 7:128837–128868, 2019.

[12] Sebastian Cygert and Andrzej Czyżewski. Toward robust pedestrian detection with data augmentation. *IEEE Access*, 8:136674–136683, 2020.

[13] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16622–16631, 2021.

[14] Zohaib Amjad Khan, Azeddine Beghdadi, Mounir Kaaniche, and Faouzi Alaya Cheikh. Residual networks based distortion classification and ranking for laparoscopic image quality assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 176–180. IEEE, 2020.

[15] Roger Gomez-Nieto, José Franciso Ruiz-Muñoz, Juan Beron, César A Ardila Franco, Hernán Darío Benítez-Restrepo, and Alan C Bovik. Quality aware features for performance prediction and time reduction in video object tracking. *IEEE Access*, 10:13290–13310, 2022.

[16] Roger Gomez Nieto, Hernan Dario Benitez Restrepo, and Ivan Cabezas. How video object tracking is affected by in-capture distortions? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE, 2019.

[17] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.*, 20:184:1–184:25, 2019.

[18] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[19] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

Skeleton-based Hand Gesture Recognition using Geometric Features and Spatio-Temporal Deep Learning Approach

*Note: Sub-titles are not captured in Xplore and should not be used

1st Abu Saleh Musa Miah
School of Computer Science and Engineering,
The University of Aizu,
Aizuwakamatsu 965-8580,
Fukushima, Japan,
d8231105@u-aizu.ac.jp

2nd Jungpil Shin
School of Computer Science and Engineering,
The University of Aizu,
Aizuwakamatsu 965-8580,
Fukushima, Japan,
jpshin@u-aizu.ac.jp

3rd Md. Al Mehedi Hasan
Department of Computer Science and Engineering,
*Rajshahi University of Eng.& and Tech. (RUET),*Rajshahi-6204, Bangladesh
mehedi_ru@yahoo.com

4th Yusuke Fujimoto
5th Asai Nobuyoshi
School of Computer Science and Engineering,
The University of Aizu,
Aizuwakamatsu 965-8580,
Fukushima, Japan, fujimoto@u-aizu.ac.jp; nasai@u-aizu.ac.jp;

***Abstract*—Dynamic hand gesture recognition using a 3D skeleton dataset has become the most attractive research domain because of the multipurpose application. Although many researchers have been working to develop hand gesture systems, they are still facing challenges in achieving satisfactory performance and more generalizable properties because of the various complexities such as lacking effective features, computational complexity and slow execution speed etc. In the study, we proposed a selected joint skeleton feature selection approach along with CNN-based spatial and Multi-head attention network (MHAN) based temporal feature extraction to alleviate the problems. In the procedure, we selected the most effective skeleton point based on the visualized capability to extract geometric features and motion speed considering slow and faster motion speed. After enhancing the feature with CNN based spatial model, we produced a final joint skeleton-independent spatial feature vector. After that, we enhanced temporal contextual information by feeding them into MHAN; we applied a classification module to refine the feature with classification. We used two benchmark datasets to evaluate the model: DHGD and SHREC'17. The high performance of the proposed model proved the superiority of the proposed model.**

Keywords*—*Hand Gesture Recognition, Hand Pose Recognition, Multi-head Attention Network, Geometric Feature, Convolutional Neural Network (CNN).

I. INTRODUCTION

Hand gesture recognition has attracted researchers for the last few years because of the emerging ripple impact in society. Currently, most people feel the flexibility to use hand gestures to control many devices such as television channel, the speed of the fan, the temperature of the air-conditioning, CCTV camera, open close door, driving a car, computer, the sound of the system, operation room and various kinds of real-life application. In addition, human-computer interaction controls wheelchairs, sign language recognition, nonverbal communication, medical assistive application and human behaviour understanding [1-2,3]. There are many researchers have been working to develop a hand gesture recognition system using images [4-9] and hand skeleton datasets. They still face difficulties producing good performance because of the lack of effective features [10-11].

To solve the lacking of effective feature problems, many researchers employed statistics, mathematics and geometrical formula to extract from the skeleton information [12-13]. Some of the researchers used cartesian coordinate-based features [14], but this feature varies from point to point or location to location and viewpoint to point. On the other hand, while there may be drawbacks to the geometric-based features of the skeleton, it is important to note that these features remain consistent across different locations and viewpoints.

The main drawback of the existing geometric feature-based research work is dataset dependent [12-13] and huge redundant elements, which cause the heavy meaningless computational cost [14]. To overcome the lack of an effective feature and reduce the redundant feature to faster the model, we selected some specific key points among 22 hand skeleton points that carry the most effective hand gesture information based on the study [11]. In the work, we extracted four different features, 2 of which are considered geometric features from the selected key point and the rest two were extracted from the coordinate point and finally concatenated them. In the first case, we extracted distance features from 10 selected joints: 5 tips, four bases and one palm centre point. Although four base points carry effective information for the gesture, these key points are not much effective for the angle-based feature. By considering this issue, we consider only 6 points to calculate the angle feature using five tips and one palms centre point. In 2nd case, we calculated slower and faster motion from the coordinate point to produce the identical value for the same gesture and high difference among inter-gesture information. Based on the above information, we extracted (i) the distance feature from the selected skeleton key point, (ii) the angle feature from the selected skeleton key point human, (iii) the slower motion feature, and (iv) the faster motion feature. Then we enhanced the feature using a spatial CNN architecture, concatenated four features to make joint invariant embedding features and fed them into attention-based architecture to enhance the temporal feature, which is demonstrated in Fig. 1.

II. RELATED WORK

Researchers have recently extracted skeleton points from the body using excellent deep-learning skeleton acquisition

techniques instead of high price devices such as depth cameras [10] or motion capture devices [15]. Researchers can use general cameras such as the web, CCTV, or mobile to collect the RGB dataset and infer the 3D skeletons [4-9] or 3D skeleton [**3**] information. Sometimes different signals, such as WiFi signals, can produce the skeleton data [11]. This portable and less expensive device and system helps create huge datasets related to resources and increases researchers' interest in developing skeleton-based action and hand gesture recognition systems [7-11]. Usually, we can distinguish two different sections based on the existing skeleton-based hand gesture recognition research work. One type can generate new feature extraction from the skeleton sequence [10,15], then used machine learning algorithm [16-18]. 2nd type can design and develop a deep neural network model [4-9,**19-21**]to enhance spatial and temporal contextual information and refine the produced feature and classification [3]. The skeleton-based dataset is usually considered a good representation containing the viewpoint invariant global motion features [9-10,22-23]. The main drawback of the single feature-based research work is too inefficient of the effective feature. Many other researchers focused on the viewpoint invariant frame-based feature without considering global motion features [12,13, 15]. Recently, yang et. proposed to combine global motion and without global motion features with the DD-Net CNN network to recognize gesture recognition [10]. The main drawback of their model is the less potential and redundant features with unsatisfactory performance. Some researchers used CNN [22-23], RNN [24-28], Graph attention model [3], and 1D CNN to recognize hand gestures. We proposed here with selected skeleton point-based feature and attention-based temporal feature extraction technique to faster the model speed with height performance accuracy of the model.

III. DATASET

We studied ten skeleton-based hand-gesture datasets to evaluate the model, such as SHREC'17 [15], DHGD [14], JHMBD [10], MSRA [3], ICVL, NV Gesture, NYU, NTU, UCF-Kinetic, UTKinetic, Florence 3-D action [3] dataset which is mainly considered as a benchmark dataset of the hand gesture recognition. We used two datasets which are most similar to our target. We considered the Skeleton dataset as a 3D dataset that can be expressed according to the following equations (1).

$$D= (Q_1, Q_2, Q_3, \dots, Q_n)^T \quad (1)$$

Where the dataset sequence is denoted by D and the multivariate time sequence of the frame is denoted by Q_j. In addition, T is denoted by the transpose operation of a matrix, and the component of each frame can be written as Equation (2).

$$Q_j(t) = \left(X^{(i)}, Y^{(i)}, Z^{(i)}\right) \quad (2)$$

Here, $Q_j(t)$ is denoted the joint skeleton position for i-th skeletal and axis j_i. In our cases, the DhGD and SHREC'17 have 22 key points collected from the intel creative camera visualized in Fig. 2. The position of each skeleton point of the camera can be expressed as $Q_i = (X_i, Y_i, Z_i) \in \mathbb{R}^3$, $\forall_i \in [\![1; N]\!]$, where N=22. Fig. 2 visualizes the 22 points' location and the number of hand key points.

A. SHREC'17 Dataset

SHREC'17 is one of the most usable benchmark datasets for skeleton-based hand gestures [15]. This dataset contains 14 right-hand gestures, including finger information and configuration. This dataset was recorded with Intel RealSense Camera from 27 people by considering two different ways of the class label setting. They collected 2800 video sequences datasets, each containing 20 to 70 frames. In the work, we considered 32 frames from every single video, and there are 22 skeleton points in each frame. To configure the class label setting, they follow two ways: coarse gesture and fine gesture, which mainly depend on finger spelling. Fig. 2 demonstrates the 22-hand skeleton points of this dataset, and the dataset is available in the following link:
https://projet.liris.cnrs.fr/eg3dor17/#shrec

B. DHGD Dataset

This publicly available dataset uses a skeleton-based dynamic hand gesture dataset with 14 rig hand gestures and finger spelling styles [14,29]. This dataset is recorded from 20 people, and are collected 2800 videos. In our study, we considered 32 frames for each video, where each franc consists of 22 hand key points, including 3D coordinates. This dataset is also coarse and fine based on finger spelling. The label configuration of the dataset also followed the coarse and fine procedure. Fig. 2 shows the 22 key points of this dataset.

IV. PROSED METHODOLOGY

Fig. 1 demonstrates the proposed method architecture, where we composed the hand-crafted feature extraction technique with a spatial-temporal deep neural network. We extracted four kinds of geometrical and coordinated (G&C) features, including Selected Joint-Coordinate Distance (SJCD), Selected Joint-Coordinate Angle (SJCA) [10-11], fast motion and slow-motion features [4, 10]. The main contribution of the work is given below:

- We calculated the geometric distance feature from 10 selected joints and the angles from 6 selected joints.
- Extracted cartesian coordinate feature as slower and faster motion.
- Applied a spatial CNN to frame wise spatial information
- Applied multi-head attention model on the concatenated feature to enhance the temporal feature.

- Finally, a new classification module was applied for predicting the new hand gesture images.

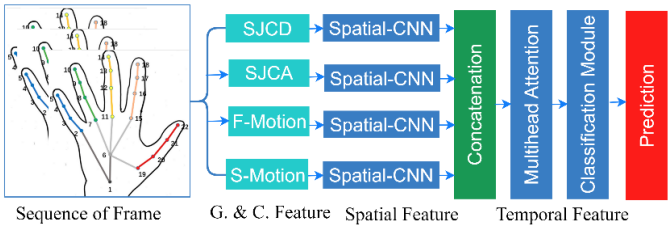


Fig. 1. Proposed working flow architecture

A. *Feature extraction*

Feature extraction in machine learning is the process of extracting the x-axis, y-axis, and z-axis, which were chosen based on their high correlation, for more details, please refer to the subsequent subjection 1) Selected Joint-Coordinate Distance (SJCD), 2) Selected Joint-Coordinate Angle (SJCA), 3) Motion Features-Fast and Slow below.

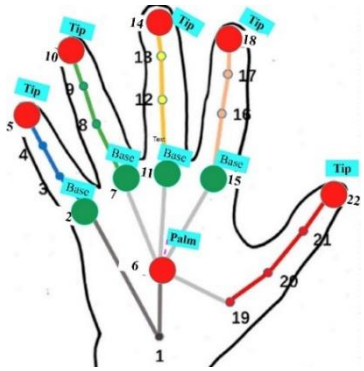


Fig. 2. 22-hand skeleton and selected joint

1) *Selected Joint-Coordinate Distance (SJCD)*

To calculate the distance and angle feature, we are considering the tip and palm points of the hand. In more explanation, 32 frames in each video can be defined as a T, and each frame has N number of joints and, in our case, N=22. For a specific frame, 3D coordinates can be expressed as $J_i^T = (x, y, z)$. Among the 22 joints, we selected ten joints for calculating the feature based on the study [11], where they considered five tip joints, four Based joints, and one from palm point sequence number is 2, 5,7,10,11,14,15,18,22 and pulm point 6 visualized in Fig. 2 red and green color.

TABLE I. DISTANCE-BASED FEATURES

Starting Joint Number	Distance to the Joint Numbers	No. of Distance-Based Features
6	{2, 5,7,10,11,14,15,18,22}	9
2	{5,7,10,11,14,15,18,22}	8
5	{7,10,11,14,15,18,22}	7
7	{10,11,14,15,18,22}	6
10	{11,14,15,18,22}	5
11	{14,15,18,22}	4
14	{15,18,22}	3
15	{18,22}	2
18	{22}	
Average	Palm to all Joint	1

We put all selected joints together, which can be written as $S^T = \{J_1^T, J_2^T, \dots, J_3^T\}$.

$$SJCD^K = \begin{bmatrix} \left\|\overrightarrow{J_6^T J_2^T}\right\| & \cdots & \left\|\overrightarrow{J_6^T J_{22}^T}\right\| \\ \vdots & \vdots & \\ \left\|\overrightarrow{J_{18}^T J_{22}^T}\right\| & & \end{bmatrix} \quad (1)$$

Here $\left\|\overrightarrow{J_6^T J_2^T}\right\|$ $(j \neq j)$ represents the distance between J_j^T and J_i^T. The final feature vector of the SJCD is generated as a dimension vector, and its size is small because of the selection procedure. Table I demonstrates the distance feature, which we calculated using Equation 1. In the first raw, we demonstrated the distance from the palm or the centre of the hand palm point to each point and got the nine distances. Then we sequentially calculated the distance from all other points to other joints.

2) *Selected Joint-Coordinate Angle (SJCA)*

We are considering only five tips and one palm point of the hand to calculate the angle feature. In more explanation, 32 frames in each video can be defined as a K, and each frame has N number of joints and, in our case, N=22. For a specific frame, 3D coordinates can be expressed as $J_i^k = (x, y, z)$. Our study selected only six joints to calculate the angle feature. Fig. 2 demonstrates the six points in red color. Table II demonstrates the calculated angle from each selected joint [10].

TABLE II. ANGLE-BASED FEATURES

Joint Index	Vector sets Starting Joint (P) and Another Joint Variable (V)	Other Joints (V)	Number of Angle-Based Features
6	$\{\overrightarrow{P_6 V}\}$	$\{P_5, P_{10}, P_{14}, P_{18}, P_{22}\}$	$5 \times 3 = 15$
5	$\{\overrightarrow{P_5 V}\}$	$\{P_{10}, P_{14}, P_{18}, P_{22}\}$	$4 \times 3 = 12$
10	$\{\overrightarrow{P_{10} V}\}$	$\{P_{14}, P_{18}, P_{22}\}$	$3 \times 3 = 9$
14	$\{\overrightarrow{P_{14} V}\}$	$\{P_{18}, P_{22}\}$	$2 \times 3 = 6$
18	$\{\overrightarrow{P_{18} V}\}$	$\{P_{22}\}$	$1 \times 3 = 3$

3) *Motion Features-Fast and Slow*

Motion information is considered as one of the most important features of the skeleton-based datasets, which we calculated by subtracting the Joint to Joint from the adjacent frame using the following formula: $V_T^M = (x_{V,T} - x_{V,T+1}, y_{V,T} - y_{V,T+1})$. Here V_T^M the video or sequence of frames that contain motion value is the frame, and t is an index of the frame. However, the motion is mainly calculated from the temporal differences of the cartesian coordinate, which can be classified into fast motion and slow motion based on the location of the joint in various frames. The scale of the global motions may vary from gesture to gesture but may not be identical for the same gesture. The scale of the temporal different can be faster or slower [12]. So exploring the actual global motion features of faster and slower motion can be more helpful. To account for this issue, we focused here on both slow global motion and fast global motion, which will be formed as two scaled global motion features. Two scales of the global motion can be expressed using the following Equation.

$$GM^T = \begin{cases} M_{Slow}^T = x_{V,T} - x_{V,T+1}, y_{V,T} - y_{V,T+1} \\ M_{Fast}^T = x_{V,T} - x_{V,T+2}, y_{V,T} - y_{V,T+2} \end{cases} \quad (2)$$

Here, GM^T, M^T_{Fast}, M^T_{Slow} are represented by the global motion, faster global motion and slower motion, respectively. Here, $V, T+1$ and $V, T+2$ are the future frame of the V, T single frame and double frame sequentially. In the same way, for all of the T frames in video V [1,…, T], we got the slow and fast motion like $M^{[1,\ldots,T-1]}_{Slow}$, $M^{[1,\ldots,T/2]}_{Fast}$

B. Spatial-CNN Model

The position of the joint can be dynamically changed based on the gesture to gesture, but sometimes it may be almost similar. Most of cases, deep neural networks consider that all joints are correlated with each other, and it is challenging to use uncorrelated data because of the various complexity. We applied a spatial CNN architecture to solve the issues, enhance the frame-wise information, and make a latent vector. We applied a new CNN architecture demonstrated in Fig. 3. It will learn the joint correlation automatically through convolution. It is also used to reduce the skeleton effect noise. Let's assume that the spatial representation of the $SJCD^T, SJCA^T, M^T_{slow}$ and M^T_{Fast} are

$$\mathcal{S}^T_{SJCD} = Spatial1(SJCD^T);$$

$$\mathcal{S}^T_{SJCA} = Spatial1(SJCA^T); \quad (3)$$

$$\mathcal{S}^T_{M_{Slow}} = Spatial1(M^T_{slow});$$

$$\mathcal{S}^T_{M_{Fast}} = Spatial2(M^T_{Fast});$$

Also, in the figure, we demonstrated the same feature for Saptial1 and Spatial2 because of the max pooling operation in the 3rd of the CNN because of the D/2 dimension, which contained about T/2 frames of the M^T_{Fast} compared to the others.

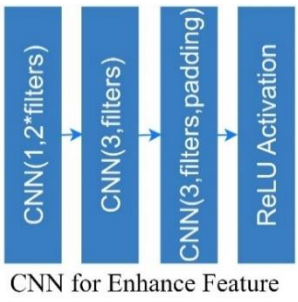


Fig. 3. Temporal Context Enhancing CNN Module

After generating the spatial feature individually, we concatenated them using the following Equation,

$$\mathcal{S}^T = \mathcal{S}^T_{SJCD} \oplus \mathcal{S}^T_{SJCA} \oplus \mathcal{S}^T_{M_{Slow}} \oplus \mathcal{S}^T_{M_{Fast}} \quad (4)$$

Where the concatenation operator is denoted by $\oplus$ and the dimension of the final feature vector can be written as $\mathcal{S}^T \in \mathbb{R}^{(\frac{T}{2})\times filers}$. We next applied the attention model based on the concatenated features to learn temporal features.

C. Multi-Head Attention Architecture

Fig. 4 demonstrates the proposed attention model we redesigned from [10,28,30].

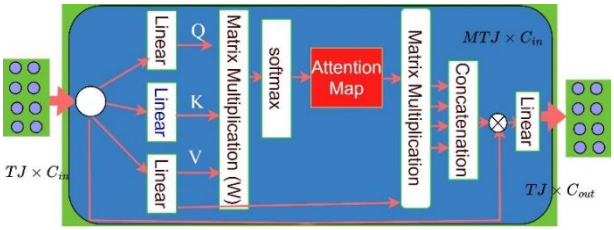


Fig. 4. Proposed Attention Module

Let our concatenated feature $f_{(T,i)}$ for the specific joint $J_{(t,i)}$ which is considered an initial feature input feature of the attention model. We are considering multi-head attention here and assume that one head name is m-th attention. Three fully connected layers are used parallelly to produce the mapping information, namely query, key and value of the input feature $f_{(T,i)}$. These mapping functions can be expressed as the following equations.

$$Q^m_{(T,i)} = W^m_Q f_{(T,i)}$$

$$K^m_{(T,i)} = W^m_K f_{(T,i)} \quad (5)$$

$$V^m_{(T,i)} = W^m_V f_{(T,i)}$$

Here, $Q^m_{(t,i)}$, $K^m_{(T,i)}$, and $V^m_{(T,i)}$ is represented by query, key and value, respectively. In addition, the weight matrix of these mapping is denoted by W^m_Q, W^m_K, W^m_V. Fig. 4 demonstrates each step operation in detail, where query and key are used to operate the dot product, then applied an activation function, producing an attention map and multiplying the attention map with the value matrix [3, 10,28,30,31]. The matrix multiplication can be expressed with the following equations.

$$u = \frac{<Q,K>}{\sqrt{d}} \text{ and } \alpha = \frac{\exp(u)}{\sum_{n=1}^{N} exp\ (\mathrm{u})} \quad (6)$$

Where the dimension of the key vector is represented with d, the scale dot product between the query and key matrix is represented by u, where the inner matrix is denoted by $\langle .,. \rangle$. In the same way, we produced four output matrices with the four heads and finally concatenated the four features and produced the final feature. After applying linear activation, concatenated the output of the attention model with the original information through the skip connection to retrieve the missing information.

D. Classification Module

After producing the temporal feature with the attention network, we applied a classification model to refine the final feature and the classification. Then we applied an averaging filter and fully connected layer for classification. Fig. 5 shows the classification module that included several CNN layers with different filters and pooling layers to refine the final feature.

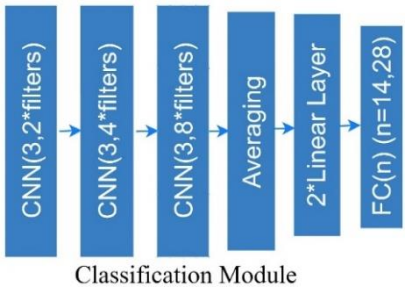


Fig. 5. Classification Module

V. Experimental Evaluation

To evaluate the model, we used the DHG and SHREC'17 datasets here. To compile the model, we used Adam optimizer with a 0.001 learning rate for the training model and 32 batch size with a 0.1 dropout rate [3-8]. We used a GPU RTX 3090 with a PyTorch platform with 24 GB GPU memory to run the code.

A. Performance Accuracy with DHGD Dataset

Table III shows the performance of the proposed model with the DHGD dataset and a state-of-the-art performance comparison with various models. Table III demonstrated that the proposed model achieved 92.21%, whereas the existing model reported a lower performance compared to ours. Some models used skeleton and depth information and achieved 8.00% average performance accuracy. Some models used joint similarities, which achieved 83.35% accuracy, and using RNN, they achieved 84.68% accuracy. They achieved 90.48% accuracy with the Hif3d model, and 91.00% accuracy was reported using the DG-STA model.

TABLE III. Performance accuracy with Shrec'17 Dataset

Method	SHREC'17 Dataset Performance (%)
SMTRM [32]	79.61
SoCJ + HoHD + HoWR [15]	88.24
Res-TCN [33]	91.10
STA-GCN [34]	92.27
MFA-Net [35]	91.31
DG-STA [31]	93.00
Yang [10]	94.60
Proposed Method	95.00

B. Performance Accuracy with SHREC'17 Dataset

Table IV demonstrates the performance of the proposed model with state-of-the-art performance for the SHREC'17 dataset. Table IV shows that our proposed model achieved 95.00% accuracy, whereas other models reported lower performance compared to our study. Among the Rcent model, STA-RES-TCN reported 93.60% accuracy, DG-STA showed 94.40% performance, and Jiang et al. reported 94.60% accuracy.

TABLE IV. Performance accuracy with DHGD Dataset

Methods	DHGD Dataset Performance (%)
GREN [22]	82.29%
ASJT [23]	82.50
SoCJ + HoHD + HoWR [14]	83.07
JAHOG [36]	83.85%
MARNN [24]	84.68
NIUKF-LSTM [25]	84.92%
CNN+RNN [26]	85.46
CNN+LSTM [27]	85.46%
Res-TCN [33]	86.90
STA-GCN [34]	91.20
MFA-Net [35]	91.31
DG-STA [31]	91.00
Multi-Branch-STA[3]	92.00
Proposed Method	**92.21**

VI. Conclusion

In the study, we proposed several feature extraction techniques with the selected hand joint key point instead of all the points of the hands. Then we applied the CNN-based spatial model to enhance the spatial feature and the attention model to enhance the temporal feature. Finally, we used a classification model to refine the final feature and predict the new input. The proposed study's high-performance accuracy proved the proposed model's effectiveness. In addition, we are focusing only on the selected six key points to extract the angle feature and ten key points for the distance feature instead of the 21 joint key points, reducing the proposed model's computational complexity. In addition, concatenated features can overcome the lacking of effective feature problems. In the future, we will evaluate the proposed model with other datasets and develop a hand gesture generalized system.

References

[1] Z. Ren, J. Meng, J. Yuan, and Z. Zhang, "Robust hand gesture recognition with kinect sensor," in Proceedings of the 19th ACM international conference on Multimedia. ACM, 28 November 2011- 1 December 2011 , Scottsdale Arizona USA, pp. 759–760.

[2] S.-E. Wei, N. C. Tang, Y.-Y. Lin, M.-F. Weng, and H.-Y. M. Liao, "Skeleton-augmented human action understanding by learning with progressively refined data," in Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia. ACM, 7 November 2014, Orlando Florida USA, pp. 7–10,

[3] Miah AS, Hasan MA, Shin J. Dynamic Hand Gesture Recognition using Multi-Branch Attention Based Graph and General Deep Learning Model. IEEE Access. 2023 Jan 9 (14)

[4] Miah, A. S. M., Hasan, M. A. M., Jang, S. W., Lee, H. S., & Shin, J. (2023). Multi-Stream General and Graph-Based Deep Neural Networks for Skeleton-Based Sign Language Recognition. Electronics, 12(13), 2841.

[5] Miah, Abu Saleh Musa, Md. Al Mehedi Hasan, Jungpil Shin, Yuichi Okuyama, and Yoichi Tomioka. 2023. "Multistage Spatial Attention-Based Neural Network for Hand Gesture Recognition" Computers 12, no. 1: 13. https://doi.org/10.3390/computers12010013

[6] Abu Saleh Musa Miah, Jungpil Shin, Md A.M. Hasan, and Md A. Rahim. "BenSignNet: Bengali Sign Language Alphabet Recognition Using Concatenated Segmentation and Convolutional Neural Network" Applied Sciences 12, no. 8: 3933. April 2022. [SCI indexed]

[7] A. S. Musa Miah, J. Shin, M. A. M. Hasan, M. A. Rahim and Y. Okuyama, "Rotation, translation and scale invariant sign word recognition using deep learning," Computer Systems Science and Engineering, vol. 44, no.3, pp. 2521–2536, 2023

[8] Shin, Jungpil, Abu Saleh Musa Miah, Md. Al Mehedi Hasan, Koki Hirooka, Kota Suzuki, Hyoun-Sup Lee, and Si-Woong Jang. 2023. "Korean Sign Language Recognition Using Transformer-Based Deep

Neural Network" Applied Sciences 13, no. 5: 3029. https://doi.org/10.3390/app13053029

[9] Md Abdur Rahim, Abu Saleh Musa Miah, Jungpil Shin, "Hand gesture recognition based on optimal segmentation techniques in human-computer interaction", 3rd IEEE International Conference on Knowledge Innovation and Invention (IEEE ICKII 2020), K200134, Aug. 21-23, 2020, Kaohsiung, Taiwan. (20-25) 4-9

[10] Yang, F., Wu, Y., Sakti, S. and Nakamura, S., 2019. Make skeleton-based action recognition model smaller, faster and better. In Proceedings of the MMAsia '19: ACM Multimedia Asia, 10 January 2020, Beijing China (pp. 1-6).

[11] Fan Yang, Sakriani Sakti, Yang Wu and Satoshi Nakamura ," Make Skeleton-based Action Recognition Model Smaller, Faster and Better", 2020, 1907.09658, archivePrefix,

[12] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, and J. Xiao, "Learning a 3d human pose distance metric from geometric pose descriptor," IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 11, pp. 1676–1689, 2011.

[13] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," IEEE Transactions on Multimedia, vol. 20, no. 9, pp. 2330–2343, 2018.

[14] Quentin De Smedt, Hazem Wannous and Jean-Philippe Vandeborre, Dynamic Hand Gesture Recognition using Skeleton-based Features , 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) June 26 2016 to July 1 2016, Las Vegas, NV, USA.

[15] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "SHREC'17 track: 3d hand gesture recognition using a depth and skeletal dataset," in 10th Eurographics Workshop on 3D Object Retrieval, , *Lyon, France, April 23-24*, 2017.

[16] Miah, A. S. M., Shin, J., Hasan, M. A. M., Molla, M. K. I., Okuyama, Y., & Tomioka, Y. (2022, December). Movie Oriented Positive Negative Emotion Classification from EEG Signal using Wavelet transformation and Machine learning Approaches. In 2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC) (pp. 26-31). IEEE.

[17] Miah, A. S. M., Shin, J., Islam, M. M., & Molla, M. K. I. (2022, February). Natural Human Emotion Recognition Based on Various Mixed Reality (MR) Games and Electroencephalography (EEG) Signals. In *2022 IEEE 5th Eurasian Conference on Educational Innovation (ECEI)* (pp. 408-411). IEEE.

[18] Miah, A. S. M., Mouly, M. A., Debnath, C., Shin, J., & Sadakatul Bari, S. M. (2021, February). Event-Related Potential Classification based on EEG data using xDWAN with MDM and KNN. In *International Conference on Computing Science, Communication and Security* (pp. 112-126). Cham: Springer International Publishing.

[19] Hossain, M. M., Hossain, M. A., Musa Miah, A. S., Okuyama, Y., Tomioka, Y., & Shin, J. (2023). Stochastic Neighbor Embedding Feature-Based Hyperspectral Image Classification Using 3D Convolutional Neural Network. *Electronics*, *12*(9), 2082.

[20] Kafi, H. M., Miah, A. S. M., Shin, J., & Siddique, M. N. (2022, February). A Lite-Weight Clinical Features Based Chronic Kidney Disease Diagnosis System Using 1D Convolutional Neural Network. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)* (pp. 1-5). IEEE.

[21] Miah, A. S. M., Mamunur Rashid, M., Redwanur Rahman, M., Tofayel Hossain, M., Shahidujjaman Sujon, M., Nawal, N., ... & Shin, J. (2021). Alzheimer's disease detection using CNN based on effective dimensionality reduction approach. In *Intelligent* Computing and Optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020) (pp. 801-811). Springer International Publishing.

[22] Ma, C.; Zhang, S.; Wang, A.; Qi, Y.; Chen, G. Skeleton-Based Dynamic Hand Gesture Recognition Using an Enhanced Network with One-Shot Learning. Appl. Sci. 2020, 10, 3680. Green

[23] De Smedt, Q.; Wannous, H.; Vandeborre, J.P. 3D Hand Gesture Recognition by Analysing Set-of-Joints Trajectories. In Proceedings of the International Conference on Pattern Recognition (ICPR)/UHA3DS 2016 Workshop, Cancun, Mexico, 4 December 2016

[24] Chen, X.; Guo, H.;Wang, G.; Zhang, L. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 2881–2885. [CrossRef]

[25] Ma, C.; Wang, A.; Chen, G.; Xu, C. Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network. Visual Comput. 2018, 34, 1053–1063.

[26] Lin, C.; Lin, X.; Xie, Y.; Liang, Y. Abnormal gesture recognition based on multi-model fusion strategy. Mach. Vision Appl. 2019, 30, 889–900. [CrossRef] 41

[27] Nunez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Velez, J.F. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. 2018, 76, 80–94. [CrossRef]

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv.Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.

[29] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using lstm and cnn," in Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on. IEEE, 2017, pp. 585–590.

[30] Shi, L., Zhang, Y., Cheng, J., Lu, H. (2021). Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition. Computer Vision – ACCV 2020, 15th Asian Conference on Computer Vision, Kyoto, Japan,

[31] Chen Y, Zhao L, Peng X, Yuan J, Metaxas DN. Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. arXiv preprint arXiv:1907.08871. 2019 Jul 20.

[32] Chunyu Xie, Ce Li, Baochang Zhang, Chen Chen, Jungong Han, Changqing Zou, and Jianzhuang Liu. Memory attention networks for skeleton-based action recognition. In IJCAI, 2018, 13-19 July 2018, Stockholm, Sweden.

[33] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatialtemporal attention res-tcn for skeleton-based dynamic hand gesture recognition," gesture, vol. 30, no. 5, p. 3, 2018.

[34] Hang, R., & Li, M. (2022). Spatial-temporal adaptive graph convolutional network for skeleton-based action recognition. In *Proceedings of the Asian Conference on Computer Vision* (pp. 1265-1281).

[35] Chen, X.; Wang, G.; Guo, H.; Zhang, C.; Wang, H.; Zhang, L. MFA-Net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data. Sensors 2019, 19, 239. [CrossRef]

[36] Ohn-Bar, E.; Trivedi, M. Joint angles similarities and HOG2 for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 465–470.

CTL-NET: Deep Learning Network for Cattle Teat Length Trait Analysis

Hina Afridi*‡, Mohib Ullah*, Øyvind Nordbø†, Anne Guro Larsgard‡ and Faouzi Alaya Cheikh*
* Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway
† Norsvin SA, Storhamargata 44, 2317, Hamar, Norway.
‡ Geno SA, Storhamargata 44, 2317, Hamar, Norway.

***Abstract*—We proposed a deep cattle teat length network (CTL-NET) for the analysis of cattle teat length trait. Our network consists of a major part and an extended part to perform regression. These parts consolidate our network to learn layer-wise teat length patterns and formulate a deep end-to-end architecture. Our network makes the feature maps less dimensional, preventing overfitting. We also use various augmentation techniques to encode variations in the collected data in terms of RGB only, depth only and fused RGB and depth images. By capturing non-linear relationships, conditioned on augmented and non-augmented data, these quantitative assessments provide good insights into how our network articulates around the teat length trait. We used mean absolute error as a performance metric. We compared the performance with four recent and state-of-the-art networks, including a network driven by spatial and channel attention.**

***Index Terms*—Cattle traits, Teat length trait, Deep learning for milk productivity, Udder conformation.**

I. Introduction

Deep learning has achieved great success in various domains, including energy [1], agriculture [2], healthcare [3], farming [4], robotics [5], security [6] and construction [7]. This technology can analyze vast data sets, recognize patterns, and make predictions with great accuracy. Nevertheless, unlike other fields, deep learning has not been explored extensively in the analysis of cattle traits. Cattle traits analysis involves examining cattle's genetic and physical features to improve breeding programs and enhance productivity. This area is still new, and deep learning has not been fully investigated yet. Additionally, expertise in both data analysis [8] and the relevant field is necessary for deep learning [9].

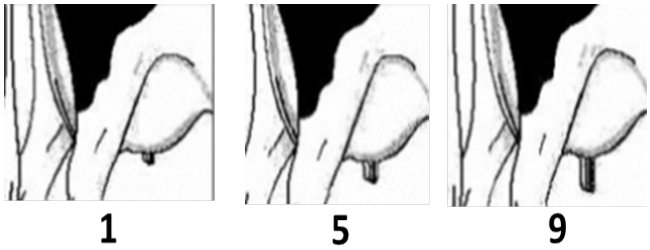


Fig. 1. Cattle teat length trait. Scores from 1 to 3 mean a short teat length. Scores from 4 to 6 represent intermediate teat length. Scores from 7 to 9 represent long teat length [10].

In the Norwegian breeding program, udder conformation traits are scored by breeding advisors using a linear scale [11]. For example, the teat length trait score ranges from 1 to 9. The score 1 is given to the shortest teat length, and score 9 is given to the longest teat length (as shown in Fig.1). This is the standard Norwegian red breading program system for scoring teat length traits. It is worth noticing that these early evaluations may not reliably predict the quality of the udder later in life. Monitoring changes in udder conformation requires repeated assessments over time, which can be a costly and time-consuming process. To alleviate this, sensors and cameras can be utilized to collect data on udder conformation, and deep learning models can be trained to automate the decision-making process. In this work, we consider cattle teat length trait. The physical characteristic known as "cattle teat length" relates to the length of the teats, which are the cows' projecting nipples on the udder. In analyzing cow qualities, the length of the teat is a crucial consideration because it might affect the effectiveness and simplicity of milking. While a cow with shorter teats may be more pleasant for the milker and produce milk more quickly, a cow with longer teats may be more difficult to milk, as shown in Fig. 1. Consequently, depending on the desired outcome, breeding programs can choose to favour or disfavour teat length. There is a potential for deep learning networks to make a considerable contribution to the sector. Therefore, we propose a deep cattle teat length network (CTL-NET) inspired by MobileNet V2 (MobN) [12]. Our CTL-NET learns representations that are appropriate for the cattle teat length attribute. Our model understands highly non-linear connections in data by interpreting probabilities or continuous values in terms of regression. Our contributions are listed as follows:

- We proposed the CTL-NET, which efficiently captures the variations in cattle data without overfitting.
- We have collected and used RGB only, depth only, and fused RGB and depth images for analysis.
- We performed analysis using non-augmented and augmented data considering four recent and state-of-the-art networks.

The rest of the paper is organized in the following Sections. In Section II, we present the related works. We present our proposed method in Section III. Experimental results on our collected dataset are presented in Section IV. The conclusion

979-8-3503-4218-5/23/$31.00 ©2023 IEEE

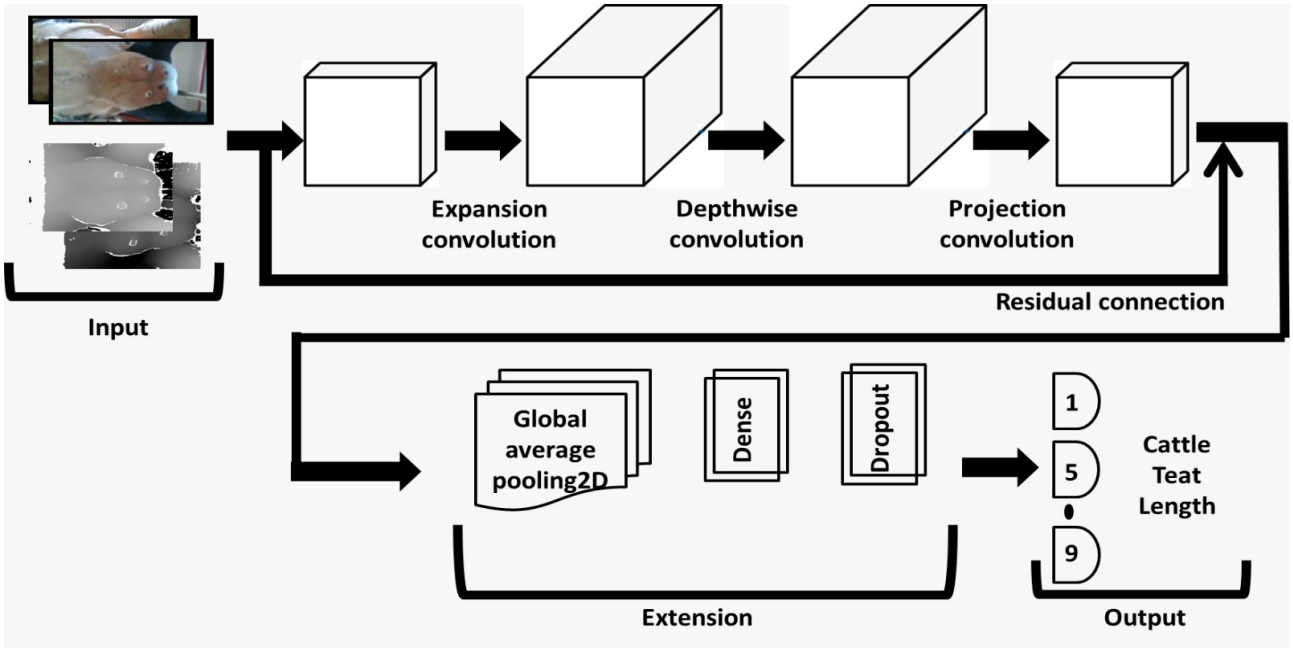


Fig. 2. Deep cattle teat length network (CTL-NET). The CTL-NET has an extended part that captures the teat length features more efficiently.

is presented in Section V.

II. Related Works

We divide the relevant works into two categories: traditional approaches and methods based on deep learning.

In the category related to the traditional approaches, De et al. [13] used Bayesian inference to evaluate breeding values through two-trait analyses to address this issue, and Carvalho et al. [14] estimated genetic parameters of several conformation and management qualities. In order to establish a connection between functional traits and milk production, Kappes et al. [15] used multivariate analysis to look at lameness score, udder cleanliness score, and udder depth. The genetic and phenotypic changes in udder shape that take place within and between parties were the focus of the study of Poppe et al. [16]. Vlahek et al. [17] took into account a variety of functional parameters, such as lameness, claw health, and female fertility. They stated that the lack of data, poor heritability of features, and limited genetic gain following selection oppose the identification based on functional qualities. According to the study [18], milk yield and udder characteristics were shown to be closely related. A probabilistic technique was used by Stefani et al. [19] to explore the genetic gains of udder, foot, and leg features. These traits' heritabilities were calculated, and the researchers discovered that choosing traits that are marginally associated, such as well-placed medium-length teats and a reasonable set of feet and legs, may help animals obtain longevity early in the genetic development process.

In the deep learning category, Fadul et al. [20] developed predictive and prescriptive decision support tools using a variety of machine learning algorithms to identify mastitis at an early stage. A difficult problem in managing dairy cow udder health, Porter et al. [21] investigated the viability of employing a deep learning system to monitor teat tissues. Their method was efficient in routinely and properly measuring teat-end status, but it ignored mastitis illness, a serious economic and health issue in the dairy industry. Xudong et al. [22] used a deep learning network based on bilateral filtering augmentation of thermal images to further speed up and automate the detection of mastitis. Ebrahimi et al. [23] studied a strategy for early diagnosis of sub-clinical mastitis utilizing deep learning-based algorithms to identify patterns of risk factors in order to overcome this difficulty. Nye et al. [24] used a composite deep learning-based system to evaluate conformational traits and derive phenotypic data from morphological aspects. The study used pedigree and picture data to estimate high heritabilities while accounting for relevant biological information.

Our study falls into the latter group and will help enhance deep learning-based techniques for examining features related to the teat length.

III. Proposed Method

We proposed a deep cattle teat length network (CTL-NET) which is a fully end-to-end trait regressor inspired by MobileNet V2 (MobN) [12]. We estimate the length trait by combining the MobN network with an extension part. The two parts are complementary to each other and are combined for efficient teat length trait estimation. Fig. 2 demonstrates a general-to-specific deep learning network CTL-NET which reduces the challenge of over-fitting on the cattle teat length trait data. Our CTL-NET network consists of a major part and an extended part. The major part is the base model of MobileNet V2. The extended part consists of a global average

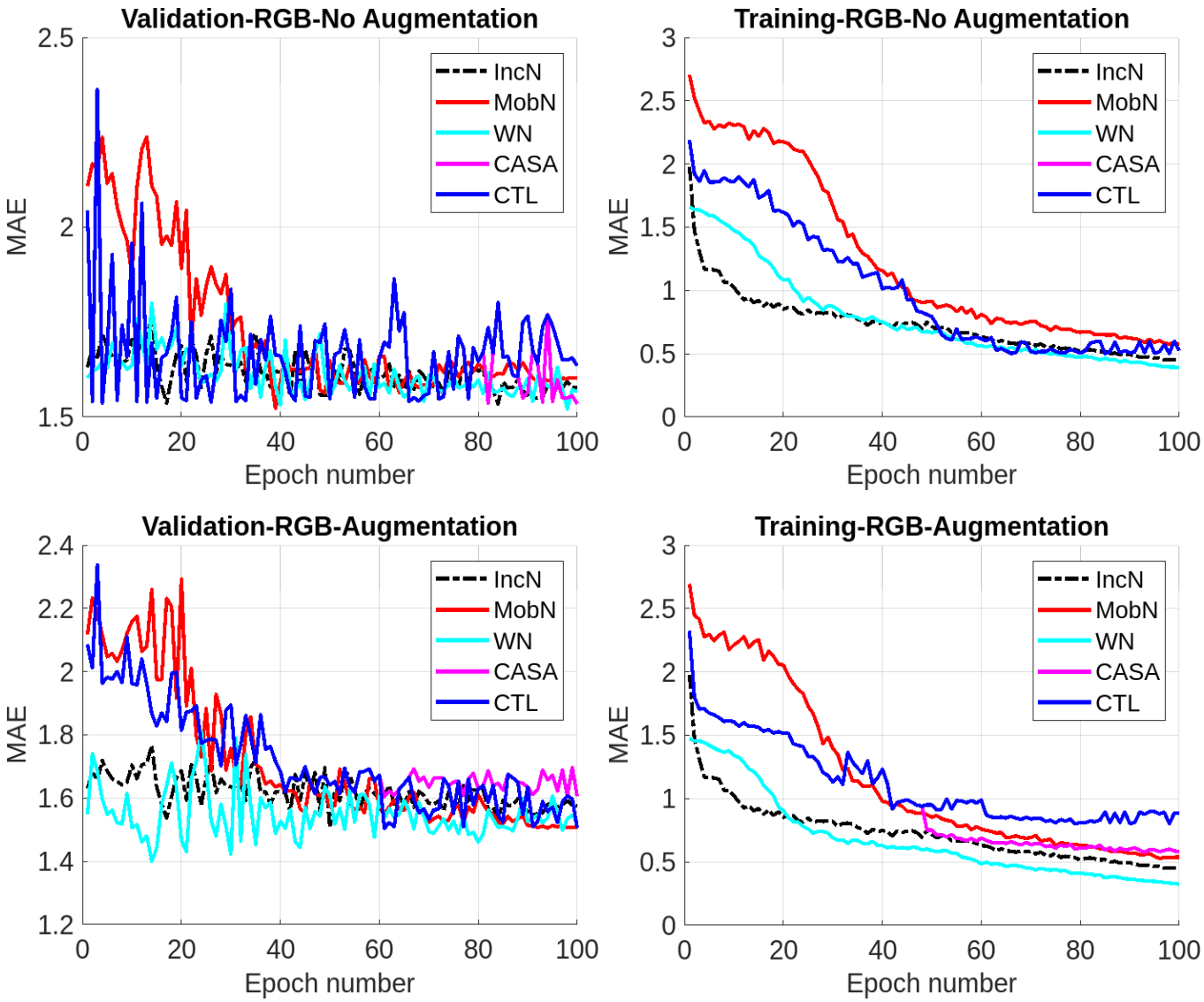


Fig. 3. Mean absolute error (MAE) for RGB-only images: We present validation and training MAE values for IncN, MobN, WN, CASA and our proposed CTL-NET without augmentation in the first row. We also present validation and training MAE values for the five networks with augmentation in the second row.

pooling layer, a dense layer with 128 nodes, a dropout layer with a rate set to 0.3, and a regression layer with a single output. The CTL-NET learns layer-wise teat length pattern representations and formulates an end-to-end deep architecture for teat length trait estimation.

The MobN network uses inverted residual blocks to increase the network's non-linearity while reducing the number of parameters. In these blocks, the input features are first extended to a larger number of channels with a 1x1 convolution, then processed by a lightweight depthwise convolution, and then projected back to a smaller number of channels with another 1x1 convolution. The MobN network also uses linear bottlenecks, which are developed to amplify the flow of features through the architecture and reduce the impact of the non-linear activations on performance. In these bottlenecks, the input features are initially processed through a linear bottleneck layer with a smaller number of channels before being transformed by the main convolutional layer. Moreover, the MobN network exploits width and resolution multipliers, which can be considered to tune the size of the architecture. The width multiplier overcomes the number of channels in the network, and the resolution multiplier adjusts the size of the input images.

Owe to the usage of the CTL-NET; we find it necessary to add an extension in terms of different components. For this purpose, we consider the global average pooling technique to minimize the spatial dimensions of the output by averaging all the data [25]. It makes the feature maps less dimensional, preventing overfitting and lowering the number of parameters in the network, improving its computational efficiency. It also

consolidates the network with a type of spatial regularization that guarantees the network learns features that are resilient to minute translations in the input image. For the image regression task, the global average pooling technique performs better than the flattened layer. The dense layer in the extension catches complicated patterns in the cattle trait data that can help the network understand intricate connections between the input and output. Sharing parameters across all neurons is possible with dense layers, which minimizes the number of parameters in the network and makes it easier to train and more computationally efficient. Our CTL-NET uses the dense layer in the extension part to learn representations of the cattle teat length trait data that is hierarchical in nature, with each layer capturing increasingly abstract characteristics of the data. The extension part also consists of dropout to avoid overfitting and enhance generalization performance. The dropout in our network prevents any individual neuron from being overly dependent on any other specific neuron by randomly dropping out (i.e., deactivating) a percentage of the neurons during each training iteration. This drives the neural network to acquire more resilient and universal characteristics. The dropout trains the CTL-NET with various neuron configurations, which essentially enables it to learn more varied and complementary representations of the input data. Additionally, the dropout in the extension part can enhance the network's overall performance on the tests by enhancing its capacity to generalize to new and unexplored data.

IV. Experimental Analysis

An Intel RealSense D415 camera and tablet were used to create a handheld gadget for taking pictures of cattle teats. Images were taken vertically up from the floor with the camera positioned beneath the cow's udder. The photos were taken in a variety of housing arrangements, including tied-stall, loose-housing, and milking parlour arrangements. The number of original images related to score 1, score 2, score 3, score 4, score 5, score 6, score 7, score 8, and score 9 are 126, 206, 195, 201, 201, 196, 118, 50, and 19, respectively. The total number of images is equal to 1312. These 1312 images are in RGB and depth pairs. To balance the number of images, we used augmentation techniques like rotation, shifting, zooming, flipping, changing brightness, and shearing. These augmentation techniques are applied to the last two scores to balance them with the rest of the data. Therefore, the number of augmented images for the last two scores is 196 and 190. The total number of images is equal to 1648 after augmentation. These images are in RGB and depth pairs, which means there are 1648 RGB images and 1648 depth images. We made sure that the same augmentation techniques were applied to the RGB and depth images in a pair to keep consistency in the data across the two data modalities. We have used the Adam optimizer, and the batch size equal to 10.

We compare the performance of the CTL-NET with four state-of-the-art networks, namely: Inception (IncN) [26], MobileNet V2 (MobN) [12], and Wide Residual Network (WN) [27]. All these models are pre-trained on the ImageNet dataset. We also compare the performance with a convolutional neural network with channel attention [28] and spatial attention (CASA). We present the experimental results in Fig 3 for all four deep learning models using only RGB images. The first row shows the results considering the dataset without augmentation. The second row shows the results considering the augmentation. We depict them in terms of validation and training performances considering the mean absolute error (MAE). As can be seen, it is challenging for all the networks to learn the variations in the dataset due to its imbalance nature and limited number of samples in the first row. In the second row, all four reference networks tend to face overfitting. The CTL-NET learns variations in the dataset efficiently.

We also present the results using only depth images in Fig. 4. As can be seen, the reference models tend again toward overfitting where our CTL-NET has the generalization capability and avoid overfitting. Last but not least, we present the results using fused depth and RGB images in Fig. 5. As can be observed, our CTL-NET learns the variations in the dataset properly. The performances of all four reference networks are not good. Our dataset is complex, consisting of nine different teat length trait scores. The reference networks have limited expressive power in terms of representing complex functions to capture the underlying patterns in the teat-length trait data. They struggle to automatically learn different features from the data due to a lack of scalability. Our CTL-NET captures more intricate patterns and relationships from the data.

TABLE I
WE PRESENT THE AVERAGE MEAN ABSOLUTE ERROR FOR THE REFERENCE NETWORKS AND OUR PROPOSED CTL-NET.

Networks	RGB	Depth	RGB and depth
IncN	1.58	1.28	1.38
MobN	1.43	1.25	1.45
WN	1.52	1.46	1.45
CASA	1.65	1.63	1.68
CTL-NET	1.59	1.25	1.21

We also present the average mean absolute error in Table I for the reference networks and our proposed CTL-NET. It is worth noticing that we consider only the values for the last 28 epochs where networks are expected to show stability. As can be seen, our CTL-NET shows smaller average mean absolute errors considering both the depth images and fused RGB and depth images. In the case of depth images, the error is 1.25, and the error is 1.21 in the case of fused information. The MobN network shows a smaller error equal to 1.43 for only RGB images. However, all these reference networks face overfitting. Therefore, having smaller errors only for RGB images does not show generalization capability. These results also show that our proposed CTL-NET achieves higher performance when the depth and RGB data are fused together. It is worth in the deep learning domain since fusing different modalities of data provides variations, and our proposed network is consolidated with these variations. Considering

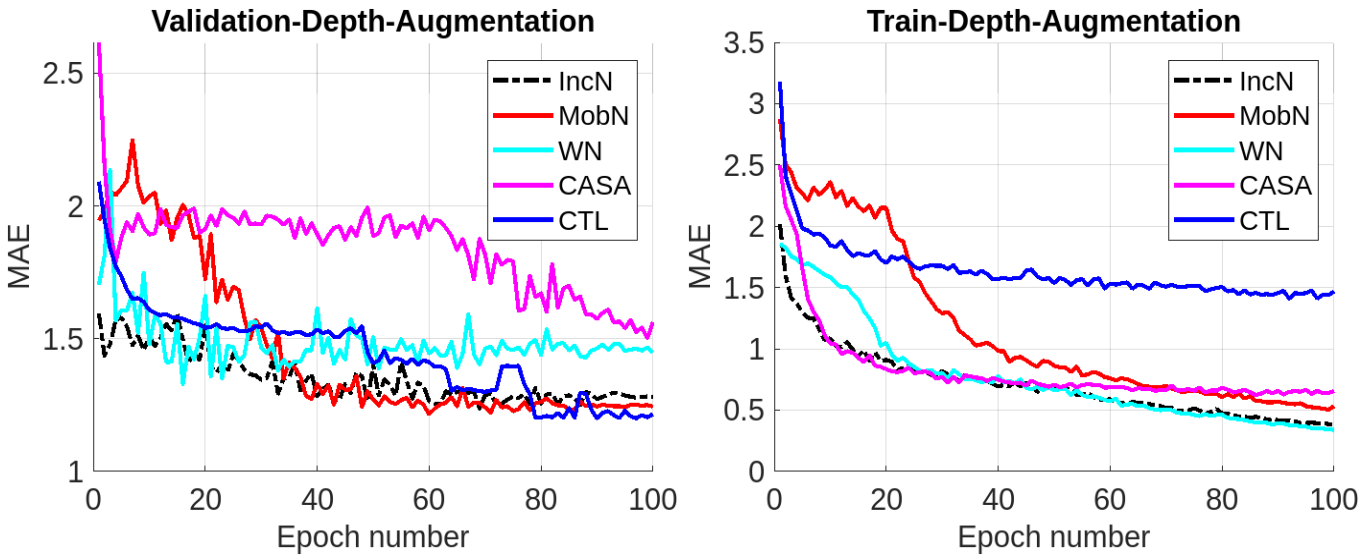


Fig. 4. Mean absolute error (MAE) for depth-only images: We present validation and training MAE values for IncN, MobN, WN, CASA and our proposed CTL-NET with augmentation.

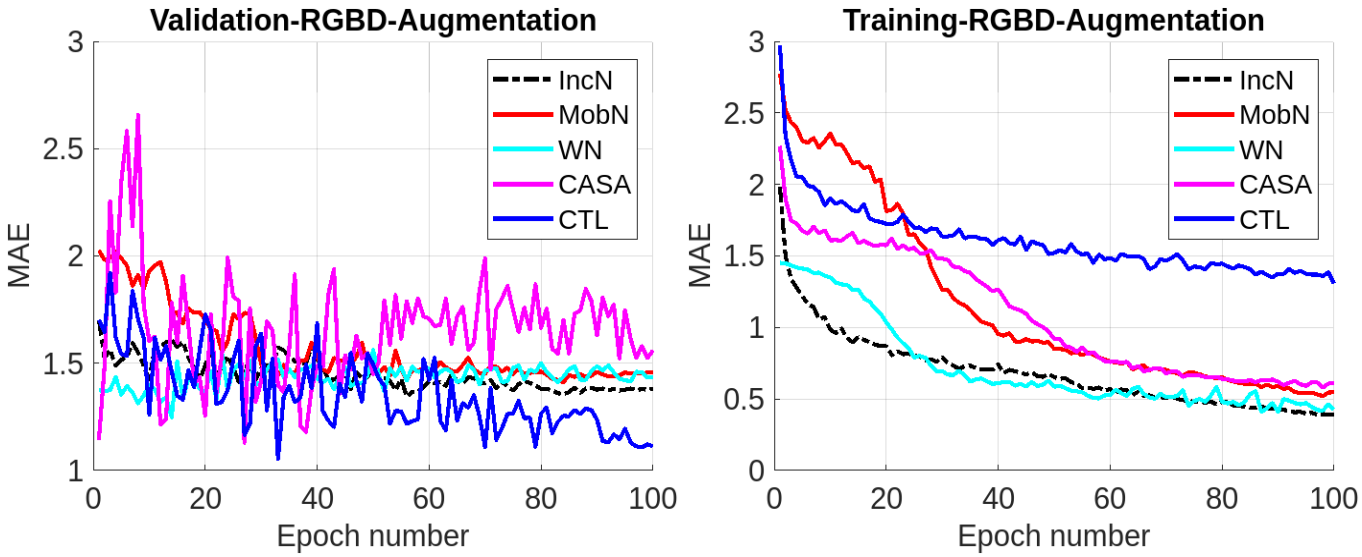


Fig. 5. Mean absolute error (MAE) for fused depth and RGB images (RGBD): We present validation and training MAE values for IncN, MobN, WN, CASA and our proposed CTL-NET with augmentation.

the limitations, we did not explore the same models without pretraining them. The models in their pre-trained status are not trained for a longer time [29]. Moreover, different approaches regarding the combination of RGB and depth information should be explored.

V. Conclusion

We proposed a CTL-NET model for teat length trait analysis. We presented the results in comparison with four state-of-the-art reference methods using the mean absolute error as the performance metric. For this purpose, we consider RGB only, depth only and fused RGB and depth images of teat length trait images. We also looked into how augmented and non-augmented datasets affected these networks' performances. Our research has shown that our CTL-NET model efficiently learns the variations in the data without facing the problem of overfitting in all the cases.

Acknowledgment

We would like to thank the Research Council of Norway for funding this study within the BIONÆR program, project number 282252 and the Industrial PhD program, project number 310239. We would also like to thank the Norwegian

University of Science and Technology for supporting the research activities.

We would like to thank Hans Snerting and Jan Atle Bakkenget hans.snerting@tine.no, jan.atle.bakkenget@tine.no (both advisors in Tine SA) for the image acquisition.

REFERENCES

[1] Elena Mocanu, Phuong H Nguyen, Madeleine Gibescu, and Wil L Kling, "Deep learning for estimating building energy consumption," *Sustainable Energy, Grids and Networks*, vol. 6, pp. 91–99, 2016.

[2] Ehsan Ullah, Mohib Ullah, Muhammad Sajjad, and Faouzi Alaya Cheikh, "Deep learning based wheat ears count in robot images for wheat phenotyping," *Electron. Imag*, vol. 34, pp. 1–6, 2022.

[3] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[4] Milan Kresovic, Thong Nguyen, Mohib Ullah, Hina Afridi, and Faouzi Alaya Cheikh, "Pigpose: A realtime framework for farm animal pose estimation and tracking," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2022, pp. 204–215.

[5] Fadi Al Machot, Mohib Ullah, and Habib Ullah, "Hfm: A hybrid feature model based on conditional auto encoders for zero-shot learning," *Journal of Imaging*, vol. 8, no. 6, pp. 171, 2022.

[6] Azeddine Beghdadi, Muhammad Ali Qureshi, Borhen-Eddine Dakkar, Hammad Hassan Gillani, Zohaib Amjad Khan, Mounir Kaaniche, Mohib Ullah, and Faouzi Alaya Cheikh, "A new video quality assessment dataset for video surveillance applications," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1521–1525.

[7] Y. Xu, Y. Zhou, P. Sekula, and L. Ding, "Machine learning in construction: From shallow to deep learning," *Developments in the built environment*, vol. 6, pp. 100045, 2021.

[8] H Ullah, M Ullah, S Daud K, and F. A. Cheikh, "Evaluating deep semi-supervised learning methods for computer vision applications," *Electronic Imaging*, , no. 6, pp. 313–1, 2021.

[9] Z. Shagdar, M. Ullah, H. Ullah, and F. A. Cheikh, "Geometric deep learning for multi-object tracking: A brief review," in *9th IEEE EUVIP*, 2021, pp. 1–6.

[10] S. Atasever and H. Erdem, "Relationships between somatic cell count and udder type scores in holstein cows," *International Journal of Agriculture and Biology*, vol. 15, no. 1, 2013.

[11] Hina Afridi, Mohib Ullah, Øyvind Nordbø, Faouzi Alaya Cheikh, and Anne Guro Larsgard, "Optimized deep-learning-based method for cattle udder traits classification," *Mathematics*, vol. 10, no. 17, pp. 3097, 2022.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L-C Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE CVPR*, 2018, pp. 4510–4520.

[13] Matheus Henrique Vargas de Oliveira, Josineudson Augusto II de Vasconcelos Silva, and da Silva Faria et al, "Genetic evaluation of weaning weight and udder score in nellore cattle," *Livestock Science*, vol. 244, pp. 104400, 2021.

[14] NS Carvalho, DS Daltro, JD Machado, EV Camargo, JC do C Panetto, and JA Cobuci, "Genetic parameters and genetic trends of conformation and management traits in dairy gir cattle," *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, vol. 73, pp. 938–948, 2021.

[15] R. Kappes, Deise A. Knob, and Thaler et al., "Cow's functional traits and physiological status and their relation with milk yield and milk quality in a compost bedded pack barn system," *Revista Brasileira de Zootecnia*, vol. 49, 2020.

[16] M Poppe, HA Mulder, BJ Ducro, and G De Jong, "Genetic analysis of udder conformation traits derived from automatic milking system recording in dairy cows," *Journal of dairy science*, vol. 102, no. 2, pp. 1386–1396, 2019.

[17] I. Vlahek, V. Sušić, A. Ekert Kabalin, S. Menčik, M. Maurić Maljković, A. Piplica, J. Šavorić, and Siniša Faraguna, "Functional traits in dairy cattle," *Hrvatski veterinarski vjesnik*, vol. 31, no. 1, pp. 48–56, 2023.

[18] PR Shorten, "Computer vision and weigh scale-based prediction of milk yield and udder traits for individual cows," *Computers and Electronics in Agriculture*, vol. 188, pp. 106364, 2021.

[19] G. Stefani, L. El Faro, Mário Luiz S. Júnior, and H. Tonhati, "Association of longevity with type traits, milk yield and udder health in holstein cows," *Livestock Science*, vol. 218, pp. 1–7, 2018.

[20] L. Fadul-Pacheco, H. Delgado, and V. E Cabrera, "Exploring machine learning algorithms for early prediction of clinical mastitis," *International Dairy Journal*, vol. 119, pp. 105051, 2021.

[21] IR Porter, M Wieland, and PS Basran, "Feasibility of the use of deep learning classification of teat-end condition in holstein cattle," *Journal of Dairy Science*, vol. 104, no. 4, pp. 4529–4536, 2021.

[22] Z. Xudong, K. Xi, F. Ningning, and L. Gang, "Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector," *Computers and Electronics in Agriculture*, vol. 178, pp. 105754, 2020.

[23] M. Ebrahimi and Mohammadi-D. et al., "Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep learning and gradient-boosted trees outperform other models," *Computers in biology and medicine*, vol. 114, pp. 103456, 2019.

[24] J. Nye, L. M Zingaretti, and M. Pérez-Enciso, "Estimating conformational traits in dairy cattle with deepaps: a two-step deep learning automated phenotyping and segmentation approach," *Frontiers in Genetics*, vol. 11, pp. 513, 2020.

[25] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation," in *IEEE CVPR Workshops*, 2015, pp. 16–24.

[26] C. Szegedy, W. Liu, and et al. Jia, "Going deeper with convolutions," in *IEEE CVPR*, 2015, pp. 1–9.

[27] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE CVPR*, 2018, pp. 7132–7141.

[29] S D Khan, A B Altamimi, M Ullah, H Ullah, and F A Cheikh, "Tcm: Temporal consistency model for head detection in complex videos," *Hindawi, Journal of Sensors*, pp. 1–13, 2020.

Underwater Object Detection using Image Enhancement and Deep Learning Models

Adane Nega Tarekegn
Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
adan.n.tarekegn@ntnu.no

Faouzi Alaya Cheikh
Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
faouzi.cheikh@ntnu.no

Mohib Ullah
Department of Computer Science
Norwegian University of Science and Technology
Gjovik, Norway
mohib.ullah@ntnu.no

Erik Tobias Sollesnes
USEA Ocean Data
Oslo, Norway
erik.sollesnes@useaoceandata.com

Cornelia Alexandru
Research & Development Department
BEIA Consult International
Bucharest, Romania
cornelia.alexandru@beia.ro

Saeed Nourizadeh Azar
R&D and AI Department
OBSS Technology
Istanbul, Turkey
saeed.nourizadehazar@obss.com.tr

Erdeniz Erol
Elkon Elektrik San. Tic. A.S
Istanbul, Turkey
eer@elkon-tr.com

George Suciu
Research & Development Department
BEIA Consult International
Bucharest, Romania
george@beia.eu

***Abstract*—**Autonomous underwater vehicles (AUVs) are efficient robotic tools, offering a wide range of applications in ocean exploration and research, such as oceanographic mapping, environmental monitoring, and archaeology. Incorporating an automatic object detection system with AUVs can substantially improve their ability to perceive and recognize objects in a complicated and often hazardous environment. Currently, detecting underwater objects relied on a man-in-the-loop approach, where AUVs captured vast amounts of data and saved them in memory for offline processing. This study investigates the use of deep learning for automatic image preprocessing and object detection, evaluating and comparing three state-of-the-art YOLO (You Only Look Once) models, including YOLOv8, YOLOv7, and YOLOv5. Extensive experiments were conducted using publicly available underwater image datasets, revealing that the pre-trained models attain superior performance on the Brackish dataset. YOLOv5 and YOLOv8 achieved the highest mean average precision (mAP) with a score of 99%, while YOLOv7 scored 89%. Furthermore, an underwater image enhancement algorithm is employed on the URPC2021 dataset, significantly improving the detection accuracy with a 3% increase in mAP across all three models. In terms of inference speed, YOLOv5 demonstrated the highest frames per second (FPS), while maintaining comparable performance in mAP and recall.

Keywords—Underwater robotics, AUV, underwater object detection, image enhancement, YOLOv8, YOLOv7, YOLOv5.

I. INTRODUCTION

The marine environment is a diverse and intricate part of the Earth's surface, serving a significant role in sustaining both the environment and human populations. It provides valuable minerals, oil, gas, and other aquatic resources, making it a target for marine exploration endeavours [1]. However, its harsh conditions hinder exploration through traditional means, rendering it the least explored environment. In recent years, the development of underwater robots, such as autonomous underwater vehicles (AUVs) provides a great opportunity to explore and protect the resources beneath the water. AUVs come with various sensing devices, including underwater cameras, sonars, depth sensors, and lighting. They are also equipped with other payload devices that enable them to monitor underwater environments and carry out intricate underwater operations. These operations include capturing marine organisms, creating oceanographic maps, inspecting pipes and cables, conducting environmental surveillance, and exploring wrecks and archaeological sites. The flourishing growth of artificial intelligence (AI) and intelligent systems are indispensable technologies to accomplish these tasks and play an important role in the development of AUVs. Integrating an intelligent object detection system on board can significantly enhance the perception and recognition capability of AUVs.

Underwater object detection methods rely on either acoustic images or optical images [2]. Sonars and vision cameras are key perception equipment used to identify and detect objects in underwater environments. In contrast to sonars, optical images captured by vision cameras offer higher resolution and a greater amount of detailed information [3]. Moreover, optical systems are more cost-effective in terms of acquisition methods. As a result, there is an increasing interest in using optical systems for underwater target detection.

Traditionally, the task of detecting underwater objects was performed by a man-in-the-loop approach where AUVs capture imaging data and store them in memory for offline processing by expert analysis [4]. However, there is an increasing demand for automatic underwater processing to enable on the fly decision-making and to extend mission times. Specifically, undersea exploration using automatic object detection has two advantages. Firstly, it allows AUVs to make real-time decisions based on the data it collects where accurate detection and recognition of objects undersea is imperative, thereby saving a lot of time and allowing longer surveys. Secondly, real-time underwater object detection can enable greater autonomy for AUVs, which perform preprogrammed missions. AI-powered AUVs are expected to perform not only to collect data but also to perceive and react to the data it collects immediately (e.g., reinspection of interesting objects).

In the last few years, deep learning (DL) techniques have revolutionized the field of computer vision and have fuelled the practical application of underwater object detection. Villon et al. [5] compared the traditional approach (histogram of oriented gradients + support vector machine) with the deep learning method in coral reef fish detection and their experimental analysis showed the superiority of the deep learning method for object detection underwater. In their

979-8-3503-4218-5/23/$31.00 ©2023 IEEE

work, Wang et al. [6] introduced a deep learning architecture that incorporates convolutional encoding and decoding features to recognize objects underwater. Their proposed framework utilizes a pre-trained convolutional model, AlexNet, which was initially trained for the ImageNet task. The authors transfer the knowledge of the first two layers of the model to facilitate the underwater detection task. Similarly, another study [7] utilized deep convolutional networks, transfer learning, and data augmentation to develop a real-time fish detection and tracking framework from video monitoring systems of AUVs. Recently, Michael et al. [8] created a method for detecting litter in underwater environments using visual deep learning to tackle the issue of plastic debris pollution. The researchers assessed the effectiveness and precision of different deep learning models, such as Faster RCNN, SSD, YOLOv2, and Tiny-YOLO. Faster RCNN was found to have the best performance, although with a weakened inference time. YOLOv2 achieved a good trade-off between speed and accuracy.

This paper aims to explore the use of the latest deep-learning techniques for automatic object detection in AUVs and to determine the most suitable algorithm for deployment using underwater images and videos. AUVs require real-time decision-making, where correct detection and classification of objects underwater is imperative. Detection based on classical computer vision is difficult and error-prone due to manually created feature extractions. The key highlights of our study are outlined in the following.

- An underwater image enhancement pipeline that incorporates colour correction, dehazing, and contrast enhancement is developed to enhance the quality of underwater images and improve detection accuracy.
- Evaluate the performance of three state-of-the-art YOLO models (YOLOv5, YOLOv7, and YOLOv8) for detecting marine objects in challenging underwater conditions.
- Conduct a comprehensive experimental study on three different underwater benchmark datasets to determine the most effective object detection approach for AUVs.

II. METHODS AND MATERIALS

A. Overall Framework

Fig.1 presents the overall framework of the underwater object detection network proposed in this study. Three publicly available underwater image datasets were employed to train the YOLO (You Only Look Once) models [9], providing a diverse range of images with varying lighting conditions, water depths, and underwater scenes. Initially, images undergo pre-processing via augmentation and an underwater image-enhancement pipeline. The resulting image, along with the original image, is then utilized as input data for the object-detection network. Three YOLO frameworks (namely, YOLOv5, YOLOv7, and YOLOv8) are used to detect and recognize objects. The performance of the trained models was evaluated on the test sets of the three publicly available underwater image datasets.

B. Underwater Datasets

This paper employs three publicly available datasets for underwater object detection. The underwater robot professional contest 2021 (URPC2021) dataset [10], the Brackish dataset [11], and the Aquarium dataset [12] were used for training underwater object detection algorithms. Fig.2 illustrates the statistical data that displays the number of targets in each dataset.

The URPC2021 dataset is an underwater robot professional contest dataset of 2021, which was created to evaluate the performance of underwater object detection algorithms. The dataset consists of 8200 underwater images that were extracted from videos captured by an underwater robot ROV in natural environments. The dataset includes box-level annotations for four categories of objects: holothurian, echinus, starfish, and scallops. The echinus category stands as the most prevalent class, followed by starfish, holothurian, and scallop, in terms of abundance, as shown in Fig.2 (a).

The Brackish dataset, created in 2019, is an underwater

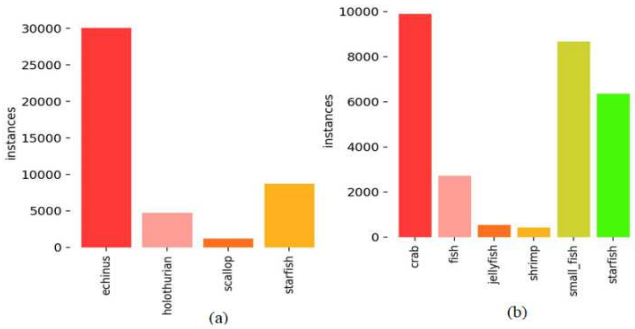


Fig.2. Statistical distribution of targets in each dataset. (a) URPC2021 dataset. and (b) Brackish dataset

dataset comprising more than 14,000 frames. It was created by annotating real filmed underwater videos and encompasses six distinct classes of underwater objects: Big fish, Crab, Jellyfish, Shrimp, Small fish, and Starfish. The dataset was collected using three mounted cameras positioned on the seabed, resulting in a diverse collection of images and viewpoints.

The Aquarium dataset is relatively smaller, consisting of only 638 images collected from two aquariums. However, it still contains multiple bounding boxes with seven different classes of underwater objects, which include fish, jellyfish, penguin, puffin, shark, starfish, and stingray. The dataset was labelled for object detection.

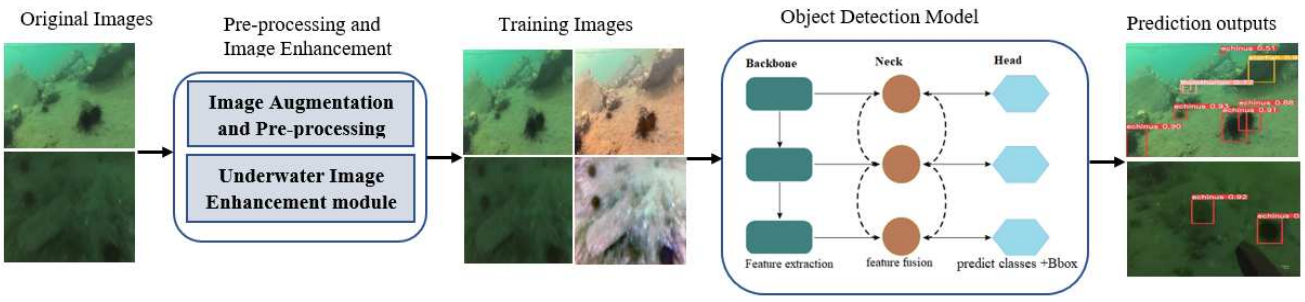


Fig.1. General framework of the object detection architecture

C. Underwater Image Preprocessing

Images captured underwater suffer from low visibility and colour distortions caused by light scattering by particles in the water and wavelength-dependent light absorption, unlike images taken on the surface. Light absorption results in significant colour distortion and loss of image information, while light scattering produces haze effects, suppresses image details, and reduces image contrast [13]. Detecting underwater objects using cameras is challenging due to these negative effects, as well as other complex background interferences such as camera shaking and non-uniform illumination, affecting real-time detection performance underwater. Fig. 3 shows some low-quality underwater images taken from the URPC2021 dataset. Image (a) depicts a low-resolution underwater image. In image (b), a noticeable colour bias is present, and the overall style appears to be dominated by green tones. Image (c) exhibits a haze effect caused by light scattering in underwater environments. The issue with image (d) lies in its low contrast and the presence of a colour cast. To improve the visual quality of such underwater images and to

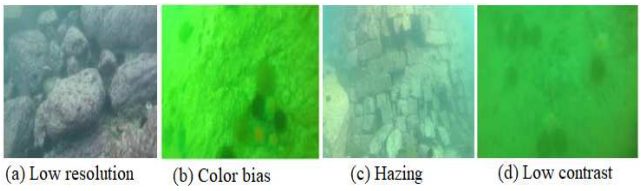


Fig.3 Example of an underwater images on the URPC2021 dataset.

enhance the detection accuracy, underwater image pre-processing, such as image enhancement or restoration, is an essential step [14].

In this study, underwater image enhancement techniques, such as dehazing, colour correction, and contrast enhancement have been applied to remove haze and the colour cast from images [15][16]. The enhancement technique used in this paper is based on a single-image approach that enhances underwater images without requiring prior knowledge of light properties or imaging models [17][18]. The image enhancement module consists of a series of independent processing steps. These steps are designed to effectively correct the degraded images and enhance their quality for improved object recognition.

D. Underwater Object Detectors

The objective of a contemporary object detector is to identify both the location and type of object in every input image. The state-of-the-art detectors consist of three primary components: a backbone that extracts features and generates a feature map representation of the input image through a reliable image classifier, a neck that is connected to the backbone and functions as a feature aggregator by assembling feature maps from various stages of the backbone and integrating these multi-level features, and a head that identifies bounding boxes and conducts classification predictions [19], as shown in Fig. 1.

The present study utilized YOLO architectures for underwater object detection. YOLO has gained widespread use as a real-time object detection system due to its exceptional speed and accuracy, resulting in its popularity in various fields like robotics, autonomous vehicles, and video surveillance. Specifically, three advanced YOLO detectors - YOLOv5, YOLOv7, and YOLOv8 - were compared in this study. YOLOv5 is a single-stage object detection algorithm, which is more efficient and versatile than its earlier iterations [20]. During evaluation on the MS COCO dataset test-dev 2017, YOLOv5 achieved an AP of 50.7% with an image size of 640 pixels. Additionally, YOLOv5 is known for its ease of use, training, and deployment. YOLOv7 is a modified version of YOLOv5, incorporating several enhancements, including the use of residual blocks, skip connections, and anchor boxes, to improve both accuracy and speed while reducing false positives [21]. YOLOv8 [22], which was recently introduced by Ultralytics, claims to be the current leader in real-time object detection. It offers faster processing speeds than previous versions of YOLO and supports state-of-the-art computer vision algorithms, such as instance segmentation and image classification.

E. Model Evaluation Measures

The standard metrics in object detection were used for evaluation and comparison of the models, including, precision, recall, precision-recall curve, average precision (AP), and mean average precision (mAP) with intersection over union (IoU).

Precision is the fraction of correct detections among all the detections made by the model, while Recall is the fraction of correct detections among all the true objects in the scene. Higher values of both metrics indicate better performance.

Intersection over Union (IoU) measures the overlap between the predicted bounding box and the ground truth bounding box. For example, how much of the picture does the predicted bounding box cover? An IoU value of 1.0 indicates a perfect overlap, while values closer to 0 indicate little to no overlap.

Average Precision (AP) measures the average precision across all recall values. A higher AP value indicates better performance. Mean Average Precision (mAP) is the average AP value across all object classes. It is commonly used in object detection competitions to evaluate the overall performance of a model.

- mAP@ 0.5: is the average of AP of all pictures in each category when IoU is set to 0.5.
- mAP@ 0.5:0.95: This is the average of mAP considering different IoU thresholds (from 0.5 to 0.95 in steps of 0.05)

III. EXPERIMENTS AND DISCUSSIONS

This section presents the implementation details and the experimental results of the proposed framework along with detailed discussions.

A. Experimental details

This study utilized the NTNU's IDUN computing cluster [23] for all experiments and implementations. The cluster comprises over 70 nodes and 90 general-purpose graphics processing units (GPGPUs), each of which is equipped with at least 128 GB of main memory and two Intel Xeon cores and is connected to an Infiniband network. Half of the nodes are fitted with two or more NVIDIA Tesla P100 or V100 GPGPUs. For training and testing YOLO models, the study employed CUDA 11.7, the PyTorch 2.0 framework, anaconda 3 with Jupyter Notebook, and Python 3.9.12.

The training epochs were set to 150 for the YOLO models. YOLOv5 and YOLOv8 models were trained on input images of size 640 × 640. In contrast, YOLOv7 was trained on size of 416 × 416 input images. All three models utilized stochastic gradient descent (SGD) as the default optimizer. The hyperparameters used for training and testing the models are summarized in Table 1. The image pre-processing method is applied to the training dataset that involves data preparation, noise reduction, augmentation, and image enhancement. In addition, we generated annotations in YOLO format, along

with configuring parameters within the pre-trained models. All three datasets (Aquarium, Brackish and URPC2021) used in this experiment were split into training (80%), validation (10%) and testing (10%).

Table 1. Parameters used in the experiment.

YOLO Model	Batch size	epochs	Learning rate	Weight decay	Input shape
YOLOv5s	32	150	0.01	0.0005	640 × 640
YOLOv7x	16	150	0.01	0.0005	416 × 416
YOLOv8	32	150	0.001	0.001	640 × 640

B. Experimental Results

Before implementing the underwater image enhancement algorithm, we conducted an experiment on the original datasets to identify the dataset that posed a great challenge for our object detection task. Our preliminary analysis of the Aquarium and Brackish datasets revealed that the YOLO models performed exceptionally well, exhibiting high precision, recall, and mean average precision (detailed performance scores can be found in Table 2). Conversely, the models yielded unsatisfactory results on the UPRC2021 dataset. The poor results obtained from the models can be attributed to the significant challenges posed by the complex environments present in the UPRC2021 dataset. These challenges include low resolution, haze due to motion blur, low contrast and colour cast as well as the frequency of smaller objects in a complex underwater environment. Fig.4 demonstrates an instance of the original distorted images (Fig.4a) from the URPC2021 dataset alongside its corresponding enhanced images (Fig.4b), obtained through the underwater image enhancement algorithm.

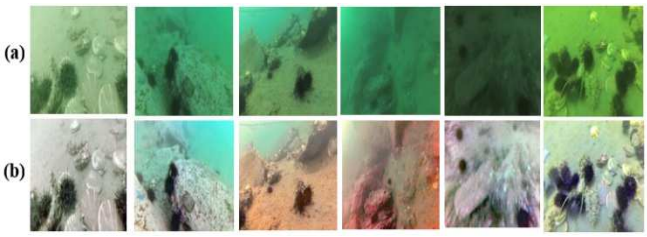


Fig. 4. Sample underwater images on the URPC2021 dataset: (a) Original images (b) Enhanced images.

In order to improve the recognition performance of pre-trained YOLO models, a combined approach is employed, utilizing both enhanced images and the original images as inputs for the object-detection network. Fig.5 illustrates the detection results in terms of mAP@0.5, comparing the performance with and without the image enhancement algorithm at 100 epochs. The model trained with image enhancement demonstrates superior performance compared to the model trained solely on the raw dataset. Notably, there is a significant increase of 3% in mAP for YOLOv5 and YOLOv7 and a 4% increase for YOLOv8. Fig. 6 shows the visualization of detection results on the URPC2021 dataset, where the different coloured squares in the figure represent the various targets in each YOLO model, such as echinus (red), holothurian (pink), starfish (yellow), and scallop (orange) in YOLOv5 and YOLOv8. YOLOv5 identified 3 bounding boxes for echinus, 2 for starfish, and 1 for holothurian, all with relatively high confidence scores. Despite the demonstrated improvement in performance using the image enhancement algorithm, YOLOv7 and YOLOv8 models still face challenges in accurately detecting Holothurian species, resulting in instances of missed detections. This is evident in Fig.6 where YOLOv5 successfully detects all classes, while YOLOv7 and YOLOv8 failed to detect Holothurian (i.e., missing detection) and exhibit occurrences of false detection in the URPC2021 dataset. The detection results on the brackish are visualized in Fig.7. It was observed that all three models successfully detected all the fish species (shrimp, starfish, and 2 crabs), with YOLOv5 and YOLOv8 having nearly equal recognition performance. Fig.8 illustrates an example of the results of using the testing dataset for the Aquarium dataset and shows the detection output of three different YOLO models, namely YOLOv5, YOLOv7, and YOLOv8. These models were able to detect all fishes and stingrays, demonstrating their flexibility to model and predict on various dimensions and scales. They were also able to handle the complexity of distinguishing between the underwater animals (such as fish and stingray) and the background, a significant challenge in underwater image analysis. Of the three models, YOLOv7 demonstrated the highest confidence value for detecting the stingray, with a value of 0.99. YOLOv8 came in second with a confidence value of 0.94, and YOLOv5 had the lowest confidence of 0.87.

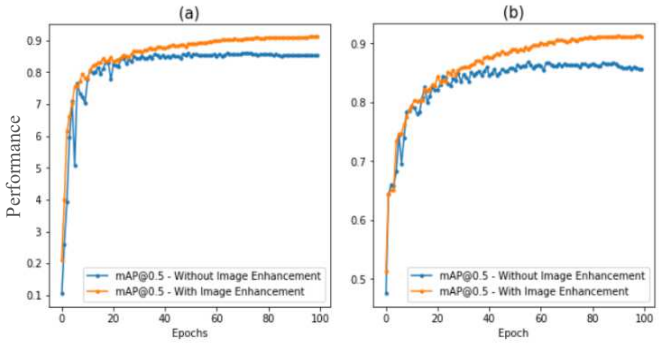


Fig.5. Comparison of mAP@0.5 for 100 epochs. Training with image enhancement shows a greater improvement in mAP compared to training without image enhancement for (a) YOLOv5 and (b) YOLOv8.

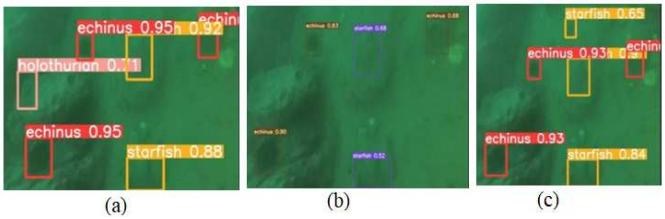


Fig.6. Detection outputs on UPRC2021 dataset: (a)YOLOv5, b) YOLOv7, (c) YOLOv8.

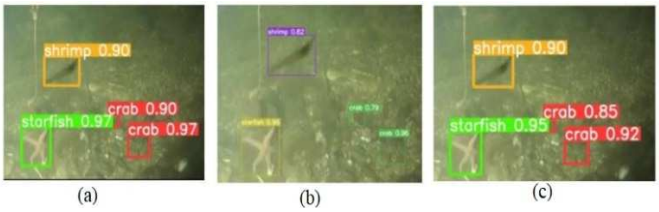


Fig.7. Detection outputs on Brackish dataset: (a)YOLOv5, b) YOLOv7, (c) YOLOv8.

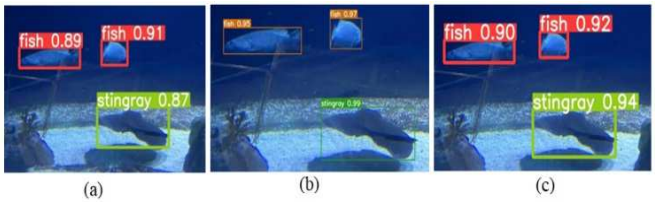


Fig.8. Detection outputs on Aquarium dataset: (a)YOLOv5, b) YOLOv7, (c) YOLOv8.

The PR curves of all classes for the YOLO models, along with the overall class curve, are depicted in Fig.9 to demonstrate their performance on the Brackish dataset. The average of all mAP classes was used to calculate the overall class curve. YOLOv5 and YOLOv8 exhibited nearly equal performance with the largest area under the curve in the figure, indicating better detection results for all target classes, especially for 'crabs', and 'starfish', with AP values of 99.5%.

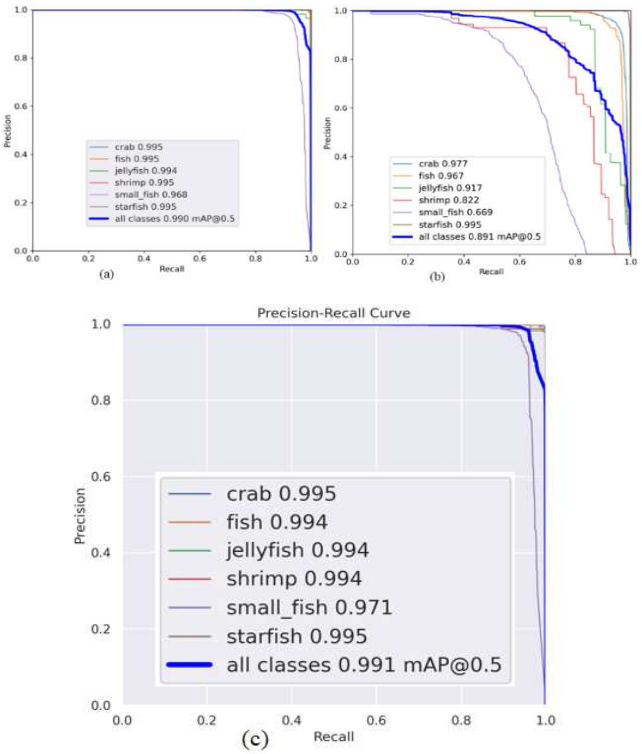


Fig.9. The precision-recall curve of (a)YOLOv5, (b)YOLOv7, and (c)YOLOv8 on Brackish dataset.

The mAP@0.5 value for YOLOv5 and YOLOv8 was 99% approximately. Conversely, YOLOv7 had the lowest area under the PR curve, resulting in a lower mAP value. Certain classes, like 'crab' and 'starfish,' are generally easier to detect due to their prevalence on the seafloor. In contrast, the 'small fish' and 'jellyfish' classes pose greater difficulty for models to learn because they can appear anywhere in the image with similar frequency and have relatively small sizes. Table 2 provides a detailed comparison of the three YOLO models on the original and enhanced dataset of URPC2021, as well as the Brackish and Aquarium datasets. The comparison encompasses various metrics such as accuracy, speed, and latency, providing a detailed analysis of each model's performance across these all the datasets. A closer look at the table, it is evident that the detection results on the URPC2021 enhanced dataset exhibit higher performance in terms of all the accuracy metrics compared to those on the original dataset. This highlights the effectiveness of image enhancement in improving the detection accuracy on the URPC2021 dataset.

On the URPC2021 enhanced dataset and Aquarium dataset, YOLOv7 achieved the highest performance values across all metrics measured. However, when applied to the Brackish dataset, YOLOv7 exhibited lower performance compared to other models. YOLOv5 and YOLOv8 exhibited excellent recognition performance on the Brackish dataset, outperforming YOLOv7 with the highest mean average precision (mAP@0.5) value of 99%. When considering recall, the YOLOv5 model exhibited superior performance compared to YOLOv7 and YOLOv8, achieving a value of 98.5%. On the other hand, YOLOv8 attained the highest performance with mAP@0.5:0.95 value of 85.6%, as shown in Table 2. Although YOLOv8 is claimed to be state-of-the-art and is expected to surpass previous versions of YOLO models in terms of performance, the detection outputs achieved on the three underwater datasets were slightly similar to those obtained with YOLOv5 across all evaluation metrics. However, since the YOLOv8 research is still in progress, it is challenging to exploit its full potential.

In addition to the detection accuracy metrics, the computational complexity of the YOLO models is provided in terms of FPS (frames per second), GFLOPS (giga floating point operations), and parameter size, as shown in Table 2. FLOPs and FPS are metrics that respectively gauge the computational complexity and detection speed of a detector, while parameter size determines the deployability of the detector. In terms of GFLOPS, the YOLOv5 architecture demonstrates an estimated computational cost of 16 GFLOPS, outperforming YOLOv7 and YOLOv8 across all three datasets (Aquarium, Brackish, and URPC2021). This indicates that YOLOv5 requires less computational power for object detection tasks compared to the latest versions. Additionally, YOLOv5 excels in terms of parameter size, making it more suitable for real-time detection and deployable on computing-constrained underwater vehicles such as AUVs (autonomous underwater vehicles). With FPS, the YOLOv5 model achieved the highest FPS on the Aquarium dataset at 135 and on the URPC2021 dataset at 169 when executed on a Tesla V100 GPU. In contrast, YOLOv7 had the lowest FPS on these two datasets. On the Brackish dataset, YOLOv8 outperformed the other YOLO models with the highest execution speed of 162 FPS. However, it had relatively a slower FPS on the UPRC2021 dataset.

In general, YOLOv5 can be optimized to deliver competitive detection performance while utilizing fewer FLOPS and achieving higher speeds. This can make it a good choice for AUVs that possess limited computing capability

Table 2: Performance results of the different YOLO models on the different datasets, including the enhanced URPC dataset.

Dataset	Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	GFLOPS	FPS	Parameters(M)
URPC Original Dataset	YOLOv5	0.869	0.780	0.854	0.664	15.8	169	7.021
	YOLOv7	0.737	0.660	0.724	0.486	188.5	122	70.835
	YOLOv8	0.863	0.784	0.857	0.685	28.4	130	11.138
URPC Enhanced Dataset	YOLOv5	0.903	0.841	0.911	0.700	15.8	169	7.021
	YOLOv7	**0.919**	**0.874**	**0.923**	**0.753**	188.5	122	70.835
	YOLOv8	0.890	0.851	0.912	*0.735*	28.4	130	11.127
Aquarium Dataset	YOLOv5	0.968	0.952	0.960	0.754	**15.8**	**135**	**7.029**
	YOLOv7	**0.982**	0.961	0.966	**0.867**	189.0	100	70.856
	YOLOv8	0.970	0.953	0.963	0.832	28.5	108	11.128
Brackish Dataset	YOLOv5	**0.984**	**0.985**	**0.990**	0.813	**15.8**	141	**7.029**
	YOLOv7	0.911	0.851	0.891	0.606	189.0	103	70.849
	YOLOv8	**0.985**	**0.982**	**0.990**	**0.856**	28.4	**162**	11.128

and memory, as it meets their requirements effectively. Despite their slightly lower inference speed compared to YOLOv5, both YOLOv7 and YOLOv8 can provide preferable options for AUVs in terms of robustness and detection performance in complex underwater environments.

IV. CONCLUSIONS

Achieving efficient recognition of objects underwater has been one of the main objectives of autonomous underwater vehicles (AUVs). This paper explores the use of vision-based deep learning algorithms for automatic object detection in (AUVs) using challenging underwater scenes. Three publicly available underwater datasets, Aquarium, Brackish, and UPRC2021, were used to compare the performance of three detection algorithms, namely YOLOv5, YOLOv7, and YOLOv8. An underwater image enhancement pipeline was developed to improve and support the object detection task using these algorithms. The objective was to select the best algorithm or model that could be integrated as a target detection and recognition component of the AUV. The study demonstrates that YOLOv7 achieved superior accuracy and speed in detecting underwater objects on the Aquarium and enhanced version of the URPC2021 dataset, achieving precision rates of 98% and 92% respectively. However, it struggled with inference time when tested on the Tesla V100 GPU, resulting in slower execution speed compared to other YOLO models. YOLOv8 shows a good balance of accuracy and speed, while YOLOv5 provides the best inference times on GPU. On the Aquarium and UPRC2021 datasets, YOLOv8 and YOLOv5 achieved nearly equal performance in terms of precision, recall, mAP@0.5, and mAP@0,5:0.95, but YOLOv5 was the fastest algorithm that outperformed both YOLOv7 and YOLOv8 across all three datasets.

Overall, the study highlights the potential of vision-based deep learning algorithms in underwater object detection and uses an image enhancement algorithm for improving system performance. The lack of high-quality underwater datasets and images remains a significant challenge in the development of target detection in underwater environments. Future research efforts will focus on optimizing the most effective models by collecting a large and diverse set of underwater datasets and employing image enhancement techniques to improve the overall quality of underwater images, which are crucial for the practicality of the system in real-world applications.

ACKNOWLEDGMENT

This research work was supported by the ADRIATIC project (cooperAtion unDerwater foR effIcient operATions vehICles) co-funded by the MarTERA partners Romanian Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI), the Scientific and Technological Research Council of Turkey (TÜBITAK) and the Research Council of Norway (RCN) and the European Union.

REFERENCES

[1] Z. Liu, M. Ling, T. Zhu, and D. Xu, "Safety Analysis of Shrinkage Monitoring Equipment in Marine Resource Exploration," J. Coast. Res., 2020, doi: 10.2112/JCR-SI105-051.1.

[2] H. Ghafoor and Y. Noh, "An overview of next-generation underwater target detection and tracking: An integrated underwater architecture," IEEE Access. 2019. doi: 10.1109/ACCESS.2019.2929932.

[3] K. Liu and Y. Liang, "Enhancement of underwater optical images based on background light estimation and improved adaptive transmission fusion," Opt. Express, 2021, doi: 10.1364/oe.428626.

[4] G. S. Kumar, U. V. Painumgal, M. N. V. C. Kumar, and K. H. V. Rajesh, "Autonomous Underwater Vehicle for Vision Based Tracking," 2018. doi: 10.1016/j.procs.2018.07.021.

[5] S. Villon, M. Chaumont, G. Subsol, S. Villéger, T. Claverie, and D. Mouillot, "Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and HOG+SVM methods," 2016. doi: 10.1007/978-3-319-48680-2_15.

[6] X. Wang, J. Ouyang, D. Li, and G. Zhang, "Underwater Object Recognition Based on Deep Encoding-Decoding Network," J. Ocean Univ. China, vol. 18, no. 2, pp. 376–382, Apr. 2019, doi: 10.1007/s11802-019-3858-x.

[7] X. Sun et al., "Transferring deep knowledge for object recognition in Low-quality underwater videos," Neurocomputing, vol. 275, pp. 897–908, Jan. 2018, doi: 10.1016/j.neucom.2017.09.044.

[8] M. Fulton, J. Hong, M. J. Islam, and J. Sattar, "Robotic detection of marine litter using deep visual detection models," 2019. doi: 10.1109/ICRA.2019.8793975.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. doi: 10.1109/CVPR.2016.91.

[10] Z. Liu, Y. Zhuang, P. Jia, C. Wu, H. Xu, and Z. Liu, "A Novel Underwater Image Enhancement Algorithm and an Improved Underwater Biological Detection Pipeline," J. Mar. Sci. Eng., 2022, doi: 10.3390/jmse10091204.

[11] A. Jesus, C. Zito, C. Tortorici, E. Roura, and G. DeMasi, "Underwater Object Classification and Detection: First results and open challenges," 2022. doi: 10.1109/OCEANSChennai45887.2022.9775417.

[12] Roboflow, "Underwater Object Detection Dataset," Kaggle, 2020. https://www.kaggle.com/datasets/slavkoprytula/aquari um-data-cots

[13] J. Y. Chiang and Y. C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," IEEE Trans. Image Process., 2012, doi: 10.1109/TIP.2011.2179666.

[14] C. Li et al., "An Underwater Image Enhancement Benchmark Dataset and beyond," IEEE Trans. Image Process., 2020, doi: 10.1109/TIP.2019.2955241.

[15] M. Afifi, B. Price, S. Cohen, and M. S. Brown, "When color constancy goes wrong: Correcting improperly white-balanced images," 2019. doi: 10.1109/CVPR.2019.00163.

[16] Y. Wang, W. Song, G. Fortino, L. Z. Qi, W. Zhang, and A. Liotta, "An Experimental-Based Review of Image Enhancement and Image Restoration Methods for Underwater Imaging," IEEE Access. 2019. doi: 10.1109/ACCESS.2019.2932130.

[17] Y. T. Peng, K. Cao, and P. C. Cosman, "Generalization of the Dark Channel Prior for Single Image Restoration," IEEE Trans. Image Process., 2018, doi: 10.1109/TIP.2018.2813092.

[18] P. Drews-Jr, E. Do Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," 2013. doi: 10.1109/ICCVW.2013.113.

[19] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," Multimed. Tools Appl., 2023, doi: 10.1007/s11042-022-13644-y.

[20] U. Nepal and H. Eslamiat, "Comparing YOLOv3, YOLOv4 and YOLOv5 for Autonomous Landing Spot Detection in Faulty UAVs," Sensors, 2022, doi: 10.3390/s22020464.

[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.02696

[22] J. Solawetz and Francesco, "What is YOLOv8? The Ultimate Guide.," Roboflow, 2023.

[23] M. Själander, M. Jahre, G. Tufte, and N. Reissmann, "EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure," pp. 1–6, 2019, [Online]. Available:http://arxiv.org/abs/1912.05848

Wild Animal Species Classification from Camera Traps Using Metadata Analysis

Aslak Tøn*, Ali Shariq Imran† and Mohib Ullah‡
Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway
Email: *aslakto@stud.ntnu.no, †ali.imran@ntnu.no, ‡mohib.ullah@ntnu.no

***Abstract*—Camera trap imaging has emerged as a valuable tool for modern wildlife surveillance, enabling researchers to monitor and study wild animals and their behaviours. However, a significant challenge in camera trap data analysis is the labour-intensive task of species classification from the captured images. This study proposes a novel approach to species classification by leveraging metadata associated with camera trap images. By developing predictive models using metadata alone, we demonstrate that accurate species classification can be achieved without accessing the image data. Our approach reduces the computational burden and offers potential benefits in scenarios where image access is restricted or limited. Our findings highlight the valuable role of metadata in complementing the species classification process and present new opportunities for efficient and scalable wildlife monitoring using camera trap technology.**

***Index Terms*—Metadata, Camera trap imaging, Neural networks, Data fusion, Scene recognition.**

I. Introduction

Human-induced influences like climate change [1], [2], deforestation [3], and trafficked roads [4], [5] have resulted in a dramatic wildlife strain, ushering in an era termed "Anthropocene" [6]. Monitoring such habitats [7], [8] is crucial, as shown by the 2019-20 Australian wildfires [9]. Camera traps offer rich insights [10]–[12], but growing data volumes necessitate robust filtering [13], [14]. Databases like LILA BC and the Snapshot Serengeti (SS) dataset [15] exist, and this paper utilizes a smaller dataset from the Norwegian Institute for Nature Research [16]. Past studies mainly employed image analysis for species identification [13], [14], [17], with few incorporating metadata [18]–[20]. Our study emphasizes metadata's significance, defining explicit metadata as data accompanying the image (like temperature, date, and location) and implicit metadata as indirect information about the image itself (like scene descriptors and attributes), extracted using pre-trained models on the places365 dataset [21]. We advance species classification by using metadata alongside image data, enhancing accuracy in camera trap research. The paper proceeds with: Related work in section II, section III discusses the methodology for data acquisition and how the classification was done, Results and discussion is in section IV, and finally we conclude our findings in section V.

II. Related Works

Although there are numerous papers discussing various aspects of metadata usage, limited attention has been given to its direct application for classification purposes. In this section, we explored related works concerning image classification, explicitly focusing on animals. For example, Norouzzadeh et al. [13] suggest image classification is enhanced by object detection, filtering irrelevant background data without requiring additional resources. They used an existing pre-trained model for object detection, achieving an accuracy of 91.71%, precision of 84.47%, and recall of 84.24%. Animals in each scene were counted via bounding boxes, and the kind of animal in non-empty images was identified. Despite an imbalanced dataset, they achieved high accuracy for the majority of classes and an overall accuracy of 91.37%. The paper also explores active learning methods. Norouzzadeh et al. [14] focuses on animal classification, object counting [22], action recognition [23], and detecting children's presence. Their multi-stage fusion network outperforms a full classifier model, tackling four objectives: animal species classification [24], social interaction [25], animal count [26], and attribute addition [27]. They achieved 96.8% accuracy with VGG [28] network for the first task, top-1 accuracy of 94.9%, and top-5 accuracy of 99.1% for the second. Binned animal count achieved 62.8% accuracy and 83.6% when counting within one bin. Action detection yielded 75.6% accuracy, 84.5% precision, and 80.9% recall. Similarly, Schindler et al. [29] proposes a two-stage fusion network using Mask R-CNN for animal classification and action determination. Temporal data from the video were used for action recognition, with variations of ResNet-18 handling $3 \times T \times H \times W$ frame input. The SlowFast network proposed by [30] underperformed. The authors also present their own accuracy metrics for segmentation, with the best segmentation method achieving 63.8% average precision and 94.1% action detection accuracy.

III. Methodology

A. Acquisition

The acquisition of the NINA Viltkamera dataset metadata is a complex task. All images and their corresponding metadata are publicly available on the Norwegian Institute for Nature Research (NINA) website. However, direct downloading is not feasible due to the extensive number of potential unique URLs. Therefore, we resorted to web scraping to acquire the necessary data. Within the website's interactive map, each camera trap pin held specific metadata. By creating a script, we automated the extraction process of these URLs and their corresponding metadata. Each URL was linked to a JSON object under the "VM" entity on the website. This JSON object

979-8-3503-4218-5/23/$31.00 ©2023 IEEE

contained essential metadata like the filename and a foreign key referencing the species id (NOR: ”FK_ArtID”). To link the foreign key with the species name, we utilized the function ”vm.arter()”. Furthermore, the ”vm.lokaliteter()” function was used to map the location ID to its corresponding latitude and longitude. This strategy allowed us to automate the extraction of metadata, which was essential for our study. In total, metadata was collected for 170 thousand camera trap images. These samples were split into 65 original classes. These classes were severely imbalanced, to the point where some classes had one or two samples. To combat this, we employed both class combination and data augmentation. More information on this is discussed in Section III-B. In terms of additional metadata, temperature data was often missing. To fill in these gaps, we used the Norwegian Metrological Institute's Frost API[1]. This API provided temperature data from the nearest weather station to the camera trap. We limited temperature readouts to within a 24-hour window of the image capture time. This still left some missing temperature values (16 thousand samples); these were set to the average temperature of the entire dataset. The date and time were stored as a one-hot encoded vector, dubbed the “datetime” vector. This preserves the cyclical nature present in time data while eliminating any ambiguity that may arise. We first considered a sine curve to represent time, as this would also capture the cyclic nature of time. However, this may have confused, as spring and fall would result in the same values. In the same vein, dawn and dusk would also result in the same values. Latitude and longitude were also included to capture potential geographical variations in animal distribution. It is important to note that the positional data acquired is only approximate, as the locations of the pins are only accurate to within about a kilometre radius. Lastly, implicit metadata was obtained through pre-trained models on the Places365 dataset. This provided us with scene attributes and scene descriptors, which offered extra context for species identification. To prevent computation delays during model training, these attributes were pre-extracted and stored alongside the image metadata.

B. Class Imbalance

As mentioned previously, the 170 thousand data points collected were severely imbalanced. The largest class “Roe Deer” consisted of 53 thousand samples alone, while other classes, like “Lemmings” only had three. The birds were especially prone to low sample size, as each individual species of bird was catalogued. Two methods were used to combat this: Class combination and data augmentation. Class combination combines certain classes, like the different bird species, into one larger super-class. In the case of bird species, we combined them to form the “Bird” superclass. Other classes were similarly combined, “Rodent” became one superclass, as did “Deer”. In total, with these combinations, we ended up with 25 classes. Furthermore, to balance out the class representation when running deep learning, we utilized Borderline Synthetic Minority Oversampling Technique (Borderline SMOTE) [31]. Borderline SMOTE generates more valuable sample points than the regular SMOTE algorithm. Borderline SMOTE generates synthetic samples on the boundary region between classes, which gives the network more hard-to-tell samples, which should provide more benefit during training.

[1]https://frost.met.no/index.html

C. Noisy Labels

One issue with this dataset is the lack of validation on the said dataset. Several samples with one given class were, in fact, a different class (see Fig. 1). Unfortunately, due to the sheer number of samples, combined with the lack of relevant expertise from the authors of the paper, reclassifying the animals is infeasible. Luckily, the vast majority of labels are correct, with only around $0.5\%-1\%$ of labels being wrong.

D. Evaluation metrics

Our study primarily focuses on two significant metrics: Accuracy and the Cohen Kappa Score. **Accuracy** quantifies the fraction of true results (including both true positives TP and true negatives TN) in the total number of samples analyzed. Formally, Accuracy is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP stands for True Positives: the number of samples correctly classified to class y_i.
- TN represents True Negatives: the samples correctly not assigned to class y_i.
- FP is False Positives: the samples wrongly assigned to class y_i but should have been classified to a different class y_j.
- FN denotes False Negatives: the samples that should have been classified to y_i but were classified to y_j.

This metric provides a view of our model's overall performance. Due to the imbalanced nature of our dataset, we opted to use a metric sensitive to prediction accuracy that accounts for class imbalance. Thus, we incorporate the **Cohen Kappa Score**. The Cohen Kappa Score measures the agreement between two raters who classify N items into C mutually exclusive classes. The score calculates the possibility of the agreement occurring by chance (p_e) and the observed agreement (p_o). Initially, the probability of random agreement, p_e, is calculated as:

$$p_e = \frac{1}{N^2} \sum_{k=1}^{C} n_k^{(1)} n_k^{(2)}$$

Here, $n_k^{(i)}$ is the number of times the rater i predicted class k. Next, the observed agreement, p_o, is calculated as:

$$p_o = \frac{\sum_{i=1}^{C} x_{i,i}}{\sum_{i=1}^{C} \sum_{j=1}^{C} x_{i,j}}$$

Fig. 1: Animal misclassifications. All are labelled as “Sheep”

Here, the elements $x_{i,j}$ constitute the observed response matrix M. Finally, the Cohen Kappa Score (κ) is calculated using these probabilities:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

This score provides a more robust measure than accuracy as it considers both the class imbalance and the probability of a correct prediction occurring by chance, offering a more nuanced view of our model’s performance.

E. Classification

To properly evaluate what effects metadata would have on classification, we need to perform an exhaustive search of the classes and features available. This involves classifying n classes using m features, where $n \geq 2$ and $m \geq 1$. To run all these combinations, we would have a total of $1,040,186,586$ individual cases to test. This amount of computation is currently unrealistic. Instead, we opted to look at a subset of the classes. The classes we decided to investigate were: ‘Fox’, ‘Deer’, ‘Mustelidae’, ‘Bird’, ‘Lynx’, ‘Cat’, ‘Sheep’, ‘Rodent’, and ‘Wolf’. We also combined temperature and position into one feature. The reasoning is that the single data point of temperature would likely not be a perfect classifier. This left us with nine classes and four features that could be included or excluded. This gives a more manageable 7529 combination that we exhaustively classify. We focused on the quantitative study of all permutations of animals and metadata information. We used a 4-layer fully connected network, with batch normalization and dropout between each layer to combat overfitting. The hidden layers were static, having 64 and 32 neurons, respectively. The input layer had a dynamic number of neurons equal to the number of input features currently selected. Likewise, the output layer was set to the current number of classes to be classified.

F. Data Visualization

Another efficient way of assessing if metadata can be used to classify different species is the use of data visualization tools. Our data consists of 538 data points, meaning we could map the data in a 538-dimensional space and assess what groupings are present in the data. As no currently known technique exists for viewing visual information above three dimensions, four if you include temporal information, we had to rely on dimensionality reduction techniques instead. Dimensionality reduction, in general, aims to preserve the structure of the data as much as possible while reducing the overall information saved for each data point. Our paper utilizes a new approach to dimensionality reduction proposed by [32]. Uniform Manifold Approximation and Projection, or UMAP for short, utilizes topology, higher dimensional manifolds, and graph theory in order to project high dimensional data down to a lower dimension while minimizing the cross entropy between the original projection and the re-projection. The algorithm has been demonstrated to equal or outperform other popular dimensionality reduction techniques such as t-SNE [33], LargeVis [34], and Laplacian eigenmaps [35]. The theory behind UMAP is quite involved, requiring a good understanding of the topic of topology. However, an excellent summary was given by [36]. They break down the process into two major steps and a couple of minor steps in each major step as so:

1 Learn manifold structure
  1.1 Finding nearest neighbours
  1.2 Constructing neighbours graph
    1.2.1 Varying distance
    1.2.2 Local connectivity
    1.2.3 Fuzzy area
    1.2.4 Merging of edges
2 Finding low-dimensional representation
  2.1 Minimum distance
  2.2 Minimizing the cost function

Utilizing UMAP, we can investigate if any patterns emerge on animal clusters. If we find local clusters in the dimensionality-reduced space, we can expect those same patterns to hold in the original 538-dimensional space we cannot investigate.

G. Implementation Details

To create and run the models, we used Python programming language, with PyTorch [37] framework for creating, importing, and training models. The models primarily used categorical cross-entropy [38] as the loss function and the Adam optimizer [39]. The networks were mainly created and trained on a Linux computer using an intel-i9 12900KF, 128 Gigabytes of RAM and an RTX3080-Ti. All weights were randomly initialized, with the optimizer set with an initial

Classes	Features used	Acc	κ
4, 6	Scene attributes	0.948	0.894
6, 12	Position and temperature, Scene attributes	0.982	0.945
4, 6	Places, Position and temperature, Scene attributes	0.967	0.932
6, 12	Datetime, Places, Position and temperature, Scene attributes	0.989	0.964
3, 4, 6	Scene attributes	0.87	0.779
3, 4, 6	Position and temperature, Scene attributes	0.869	0.782
3, 4, 6	Datetime, Places, Scene attributes	0.866	0.775
3, 4, 6	Datetime, Places, Position and temperature, Scene attributes	0.878	0.796
2, 3, 4, 6	Scene attributes	0.696	0.552
3, 4, 6, 12	Position and temperature, Scene attributes	0.731	0.603
3, 4, 6, 12	Datetime, Position and temperature, Scene attributes	0.729	0.614
3, 4, 6, 12	Datetime, Places, Position and temperature, Scene attributes	0.746	0.63

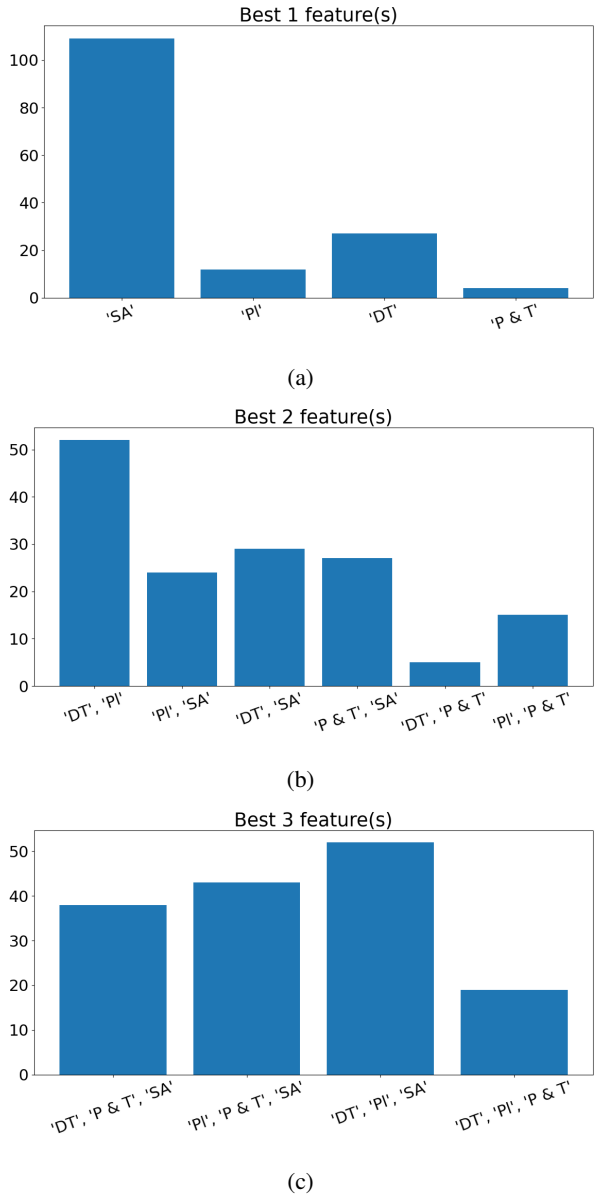
TABLE I: Metadata Predictors Scores

learning rate of $1e-3$. The learning rate was then reduced by an order of magnitude every seven epochs, and a total of 25 epochs ran for each model. The samples were split into mini-batches of 64. For each epoch, the model was validated using 10% of the test samples; if the model performed worse than previous runs, it was reset back to its best-performing iteration. Finally, the model was evaluated using 10% of the data that was left aside before training started.

To ensure balanced representation in the training data. Borderline SMOTE [31] was utilized. By having the same number of samples from each class, the network cannot "cheat" by only predicting the majority class to achieve an acceptable result. The validation sets and testing sets were left unaltered.

IV. Results and Discussion

We can see the results for two or three separate classes using one, two, three or all four features. Looking at Table I, we see a reasonably high accuracy for classifying some animal species, despite not having any image data. We've decided to use an ID for each species instead of the said species' name. The corresponding ID to species is 0: 'Fox', 1: 'Deer', 2: 'Weasel', 3: 'Bird', 4: 'Lynx', 5: 'Cat', 6: 'Sheep', 7: 'Squirrel', 8: 'Rabbit', 9: 'Rodent', 10: 'Cattle', 11: 'Boar', 12: 'Wolf', and 13: 'Bear'. We see that the "Scene attributes" information yields the best single feature to include in the prediction. We also see as we increase the number of features included increases, the best performer is still "Scene attributes". However, including extra attributes does yield diminishing returns. The average performance of the different features is less clear-cut. We can quantify this relation better by looking at the "winner" when pitting n predictors against each other to predict between m classes. By finding and counting the best predictor(s) for all combinations of animals, we get Fig. 2. To save space, we used abbreviated versions of the feature names, 'SA' equates to scene attributes, 'Pl' is short for "Places" which are the Scene descriptors, 'DT' is the datetime vector, and 'P & T' is the position and temperature information. We see that "Scene attributes" is the clear best single predictor. However, it is not among the pair of best predictors, being beaten out by the combination of "Datetime" and "Places". Its worth noting that this method of counting the winner does not take into account



(a)

(b)

(c)

Fig. 2: The best n features to use to distinguish a set of m animals

how much better one predictor performed than another. We do not know whether "Scene Features" dominated the competition as the singular feature or if other features were close seconds to the best performance of "Scene Features". However, we can conclude that accuracy, in general, improves when more features are included. Meaning all the metadata contributes something valuable to the prediction of the animal feature. Remember that these predictions of animal classes are purely based on the metadata information, no image of the animal is given to the model, yet it can quite confidently predict between two classes.

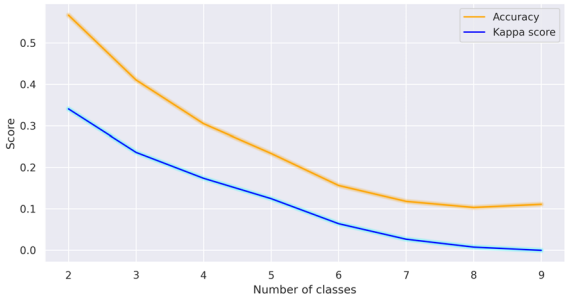
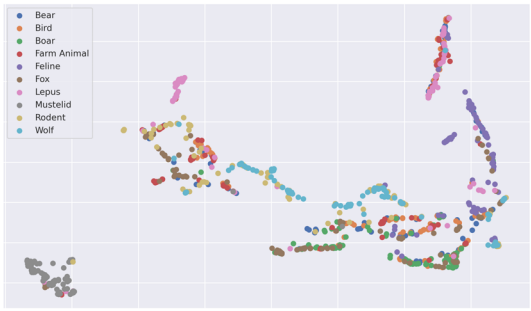
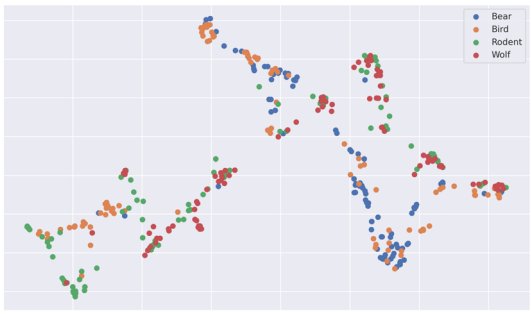


Fig. 3: Prediction score versus the number of classes to distinguish

The prediction score does steadily decrease as more classes are included. Fig. 3 demonstrates this clearly. We postulate this is due to the increased homogeneous actions of the animals. Some animals may be active during the daytime, others during nighttime; some are preferentially spotted in some locations, while others avoid those same locations. When we only have two animals, we can use these facts to separate them. However, once multiple animals act similarly, we can no longer separate them purely using this metadata, and image data are required. This issue of reduced performance when more classes make intuitive sense. It is harder to guess between 5 categories than it is to guess between only two. However, the kappa score should account for this increased performance of randomly guessing the correct class, but it is also declining. Some of the explanations for this can be seen by using UMAP. Fig. 4a shows Mustelidae cleanly separating into its own cluster. This indicates that some higher dimensional line can be drawn that can confidently classify Mustelidae away from other animals. However, once we remove many of the classes, we find that UMAP no longer cleanly separates these classes. This problem can be seen in Fig. 4b. We can summarize that metadata has the ability to help differentiate species from each other without the need for image data to be included. These findings are more valuable when we include image data once again. By designing networks that can incorporate metadata to image feature extraction for networks, we believe we can enhance the classification results over standard network architectures. Metadata should prove even more helpful in cases where there are few classes to choose from or where the existing classes have distinct behavioural patterns that separate them from each other at a metadata level, such as different biomes, locations, or sleep schedules that result in image capture during different hours.



(a) UMAP separating Mustelidae cleanly from other classes



(b) UMAP struggling to separate the remaining classes

Fig. 4: UMAP embedding of metadata features and classes

V. Conclusion

In our study, we have showcased the effectiveness of utilizing explicit and implicit metadata associated with camera trap images for animal prediction. The results obtained highlight the potential of metadata-driven augmentation for deep-learning approaches in the field of animal classification. Building upon these findings, we recommend employing a two-step classification process: First, identifying the appropriate subgroups into which animals can be separated using the available metadata and then utilizing more specific prediction models to assign the final species label to each animal. This coarse-to-fine classification methodology aligns well with the outcomes and implications presented in the paper. Our work holds promise for improving the overall accuracy and efficiency of animal classification in camera trap research.

Acknowledgment

We would like to acknowledge the provision of images by the SCANDCAM project, which is coordinated by the Norwegian Institute for Nature Research and has received funding from the Norwegian Environment Agency and various Norwegian county councils.

REFERENCES

[1] V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, Eds., *Human Influence on the Climate System*, pp. 423–552, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[2] Muhammad Munsif, Hina Afridi, Mohib Ullah, Sultan Daud Khan, Faouzi Alaya Cheikh, and Muhammad Sajjad, "A lightweight convolution neural network for automatic disasters recognition," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[3] Kusum Lata, Arvind Kumar Misra, and Jang Bahadur Shukla, "Modeling the effect of deforestation caused by human population pressure on wildlife species," *Nonlinear Analysis: Modelling and Control*, vol. 23, no. 3, pp. 303–320, 2018.

[4] W Richard J Dean, Colleen L Seymour, Grant S Joseph, and Stefan H Foord, "A review of the impacts of roads on wildlife in semi-arid regions," *Diversity*, vol. 11, no. 5, pp. 81, 2019.

[5] Maryam Hassan, Farhan Hussain, Sultan Daud Khan, Mohib Ullah, Mudassar Yamin, and Habib Ullah, "Crowd counting using deep learning based head detection," *Electronic Imaging*, vol. 35, pp. 293–1, 2023.

[6] Simon L Lewis and Mark A Maslin, "Defining the anthropocene," *Nature*, vol. 519, no. 7542, pp. 171–180, 2015.

[7] Joel Berger, Steven L Cain, and Kim Murray Berger, "Connecting the dots: an invariant migration corridor links the holocene to the present," *Biology Letters*, vol. 2, no. 4, pp. 528–531, 2006.

[8] Toby A Patterson, Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos, "State–space models of individual animal movement," *Trends in ecology & evolution*, vol. 23, no. 2, pp. 87–94, 2008.

[9] Isabel T Hyman, Shane T Ahyong, Frank Köhler, Shane F McEvey, GA Milledge, Chris AM Reid, and Jodi JL Rowley, "Impacts of the 2019–2020 bushfires on new south wales biodiversity: a rapid assessment of distribution data for selected invertebrate taxa," *Technical reports of the Australian Museum online*, vol. 32, pp. 1–17, 2020.

[10] Franck Trolliet, Cédric Vermeulen, Marie-Claude Huynen, and Alain Hambuckers, "Use of camera traps for wildlife studies: a review," *Biotechnologie, Agronomie, Société et Environnement*, vol. 18, no. 3, 2014.

[11] Allan F O'Connell, James D Nichols, and K Ullas Karanth, *Camera traps in animal ecology: methods and analyses*, vol. 271, Springer, 2011.

[12] Francesco Rovero, Fridolin Zimmermann, Duccio Berzi, and Paul Meek, ""which camera trap type and how many do i need?" a review of camera features and study designs for a range of wildlife research applications.," *Hystrix*, 2013.

[13] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune, "A deep active learning system for species identification and counting in camera trap images," *Methods in ecology and evolution*, vol. 12, no. 1, pp. 150–161, 2021.

[14] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[15] AB Swanson, M Kosmala, CJ Lintott, RJ Simpson, A Smith, and C Packer, "Data from: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna," 2015.

[16] John Odden and Jon E. Swenson, "Scandcam project," https://viltkamera.nina.no/, 2023, Images provided by the SCANDCAM project coordinated by the Norwegian Institute for Nature Research with funding from the Norwegian Environment Agency and multiple Norwegian county councils.

[17] Mohib Ullah, Zolbayar Shagdar, Habib Ullah, and Faouzi Alaya Cheikh, "Semi-supervised principal neighbourhood aggregation model for sar image classification," in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2022, pp. 211–217.

[18] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017, Submitted on 7 Feb 2017.

[19] Andre GC Pacheco and Renato A Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE journal of biomedical and health informatics*, vol. 25, no. 9, pp. 3554–3563, 2021.

[20] Weipeng Li, Jiaxin Zhuang, Ruixuan Wang, Jianguo Zhang, and Wei-Shi Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE, 2020, pp. 1996–2000.

[21] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[22] Sultan Daud Khan, Ahmed B Altamimi, Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh, "Tcm: Temporal consistency model for head detection in complex videos," *Journal of Sensors*, vol. 2020, pp. 1–13, 2020.

[23] Mohib Ullah, Muhammad Mudassar Yamin, Ahmed Mohammed, Sultan Daud Khan, Habib Ullah, and Faouzi Alaya Cheikh, "Attention-based lstm network for action recognition in sports," *Electronic Imaging*, vol. 33, pp. 1–6, 2021.

[24] Tinao Petso, Rodrigo S Jamisola, and Dimane Mpoeleng, "Review on methods used for wildlife species and individual identification," *European Journal of Wildlife Research*, vol. 68, pp. 1–18, 2022.

[25] Habib Ullah, Sultan Daud Khan, Mohib Ullah, and Faouzi Alaya Cheikh, "Social modeling meets virtual reality: An immersive implication," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*. Springer, 2021, pp. 131–140.

[26] Colin J Torney, David J Lloyd-Jones, Mark Chevallier, David C Moyer, Honori T Maliti, Machoke Mwita, Edward M Kohi, and Grant C Hopcraft, "A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images," *Methods in Ecology and Evolution*, vol. 10, no. 6, pp. 779–787, 2019.

[27] Milan Kresovic, Thong Nguyen, Mohib Ullah, Hina Afridi, and Faouzi Alaya Cheikh, "Pigpose: A realtime framework for farm animal pose estimation and tracking," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2022, pp. 204–215.

[28] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] Frank Schindler and Volker Steinhage, "Identification of animals and recognition of their actions in wildlife videos using deep learning techniques," *Ecological Informatics*, vol. 61, pp. 101215, 2021.

[30] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[31] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[32] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[33] Geoffrey E Hinton and Sam Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[34] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei, "Visualizing large-scale and high-dimensional data," in *Proceedings of the 25th International Conference on World Wide Web*. apr 2016, International World Wide Web Conferences Steering Committee.

[35] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[36] Hina Afridi, Mohib Ullah, Øyvind Nordbø, Faouzi Alaya Cheikh, and Anne Guro Larsgard, "Optimized deep-learning-based method for cattle udder traits classification," *Mathematics*, vol. 10, no. 17, pp. 3097, 2022.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019.

[38] Zhilu Zhang and Mert R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," 2018.

[39] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

A HITCHHIKER'S GUIDE TO WHITE-BOX NEURAL NETWORK WATERMARKING ROBUSTNESS

Carl De Sousa Trias [1], *Mihai Mitrea* [1], *Enzo Tartaglione* [2], *Attilio Fiandrotti* [2,3], Marco Cagnazzo [4], Sumanta Chaudhuri [2]

[1] Télécom SudParis, Institut Polytechnique de Paris, France
[2] Télécom Paris, Institut Polytechnique de Paris, France
[3] Universitá di Torino, Italy
[4] Universitá di Padova, Italy
Email: carl.de-sousa-trias@telecom-sudparis.eu

ABSTRACT

The present study deals with white-box Neural Network (NN) watermarking and focuses on the robustness property. The first contribution consists of formalizing neuron permutation as a geometric attack, thus demonstrating the very existence of this class of attacks for NN watermarking. The second contribution consists in devising and demonstrating the effectiveness of the corresponding counter-attack. As a side result, the possibility of extending NN white-box watermarking scope beyond image classification is brought to light. The experimental study considers three state-of-the-art methods, four NN models, three tasks (image classification, segmentation, and video coding), and five types of attacks. We underline that none of the existing methods is robust against the geometric attack, and using the counter-attack advanced in this paper effectively ensures the robustness.

Index Terms— watermarking, neural network, white-box, robustness, geometric attacks, counter-attack.

I. INTRODUCTION

Neural Networks (NN) are currently serving as enablers for practically all multimedia-related tasks, such as image classification, segmentation [1] or compression [2]. Design, data collection, and training of NN require huge investment, and protecting the underlying intellectual property rights is not only an ethical issue but an economic one, as well. Moreover, such applications can also be deployed in critical contexts (*e.g.* autonomous driving), where it is key to verify that the NN functioning has not been corrupted.

Watermarking represents a promising solution to the above, and potentially other related problems [3], [4]. Watermarking [5] originally refers to *imperceptibly* and *persistently* embedding into multimedia contents some additional information (referred to as *watermark* or *mark*) according to a *secret key*. Inserted by an *authorized user*, the watermark detection is expected to track down an *unauthorized user* that would illicitly benefit from or modify that content.

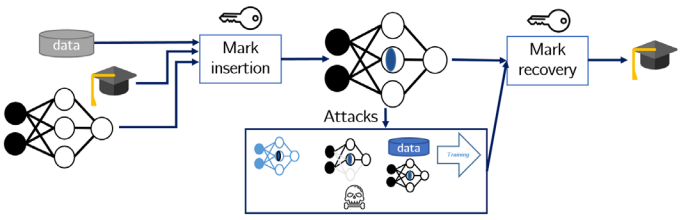


Fig. 1: Neural network watermarking synopsis.

This generic framework inherited from multimedia realm is to be reconsidered and extended to match the NN peculiarities, as detailed here after and illustrated in Fig. 1. First, the watermark is inserted into the NN model (defined as the set of parameters of a neural network, including the input-output functions). The watermark can be either retrieved from the parameters of the model (the so-called *white-box* methods [3], [6], [7]) or from the inference output by the watermarked model (the so-called *black box* ones [4], [8]). The *data payload* represents the size of the watermark, *i.e.* the quantity of information to be inserted and detected.

Second, the *imperceptibility* refers to the impact (if any) of the mark insertion in the task achieved by the NN. For instance, watermarking a NN for image classification is imperceptible when the class score distribution is not modified.

Third, *robustness* is the property of recovering the mark from the protected content even when it is subjected to malicious or mundane operations (commonly referred to as *attacks*).

Finally, the *secret key* refers to the information that should be kept secret and implicitly ensures the method's security (in the Kerckhoff's sense).

In practice, each watermarking method finds a trade-off among these four properties, according to the actual application constraints [5]. For instance, the authorized user can trade the data payload for reaching prescribed imperceptibility and robustness. On the other side, the unauthorized

user is expected to devise attacks that would abide to the imperceptibility constraint, while decreasing the robustness property.

The state-of-the-art analysis carried out in Section II highlights two methodological limitations of the NN watermarking landscape. First, the robustness is solely analyzed against mundane modifications related to NN life cycle (like pruning or fine-tuning, for instance) and implicitly assumes that the unauthorized user would not make any malicious attempt against the watermark. Secondly, although the NN application field is so broad, classification seems to be the only application benefiting from white-box watermarking.

The present study presents two contributions to the state-of-the-art in white-box watermarking [3], [6], [7]. First, the NN permutation attack [9] is formalized, thus demonstrating the very existence of *geometric* attacks for NN watermarking. Secondly, an effective counter-attack is devised and investigated on tasks beyond classification. The experiemntal study is based on three methods ([3], [6], [7]), four architectures (VGG16, ResNet34, DeepLabV3, and DVC), three tasks (image classification, segmentation, and video coding), and five types of attacks (Gaussian noise, fine-tuning, pruning, quantization, and permutation). Beyond analyzing the threats and opportunities related to NN geometric attacks and counter-attacks, this study serves as practical guidelines when designing effective NN watermarking methods.

II. BACKGROUND AND PROBLEM STATEMENT

This section first introduces the attack taxonomy as inherited from multimedia watermarking, then sketches the panorama of NN white-box watermarking solutions before identifying the issues raised by NN permutation attack.

II-A. Watermarking robustness and attack taxonomy

Robustness is the property of detecting the watermark, even when the watermarked model is subjected to modifications commonly referred to as *attacks*. The robustness is evaluated by assessing the ability to detect the watermark. For example, the BER (bit error rate) between the inserted and the recovered watermark can be computed [3], [6]. Alternatively, the correlation coefficient between the inserted and detected watermarks might be computed [7]. Conceptually, when evaluating the robustness, no distinction is made against mundane attacks (*i.e.* operations coming across with the usual NN life-cycle, like fine-tuning for better performances or pruning for lower footprint) and malicious attacks (*i.e.* operations specifically designed by unauthorized users to decrease the robustness).

In the multimedia realm, watermark attacks are classified as *removal attacks*, *geometric attacks*, *cryptographic attacks*, and *protocol attacks*. Removal attacks simply attempt to make the watermark unreadable. Geometric attacks do not try to remove the mark, but rather destroy the detector synchronization. Cryptography attacks aim at detecting and removing the watermark without any knowledge of the key, exploiting the fact that the embedded watermark is public and/or by assuming a detector (working with the proper key) is available. Finally, protocol attacks are meant to create ambiguity and confusion about watermark usage, even if properly detected. Removal and geometric attacks intimately relate to the insertion and detection methods. Cryptography attacks relate to the system security and secret key management and can be, for instance, based on *known text* attacks or on *oracle attacks*, as inherited from cryptography [10]. Protocol attacks deal with the practical watermark usage, as legal proof of copyright and/or integrity. **The present study will focus on removal and geometric attacks, while the last two classes can be conceptually considered complementary with respect to the paper scope.**

II-B. White-box neural network watermarking

The earliest NN watermarking methods [3] considers image classification, namely a wide residual network trained on CIFAR10 dataset or Caltech-101. A binary watermark of M bits is inserted in the so-called *flattened version of the layer* l, where M is lower than the number of input channels N_{l-1}. The key is represented by a random matrix $X \in \mathbb{R}^{N_{l-1} \times M}$. The mark is embedded during training via a regularization term minimizing the distance between the watermark and the projection of the flattened watermarked weights on the key. Watermark detection is achieved by projecting the watermarked (and possibly attacked) layer on the secret key, rounding the product results towards 0 or 1; the BER with respect to M is subsequently computed. The robustness is checked against fine-tuning (additional epoch of training without the embedding term up to 50% of the total training) and magnitude pruning (remove the fraction $\boldsymbol{T} \in [0.1; 0.99]$ of the smallest weights in terms of $L1$-norm).

While [6] inherits its key concept from [3], the mark is now embedded in the activation function of the selected layer. Four architectures are investigated: an MLP trained on MNIST, a test CNN and a WideResNet trained on CIFAR10, and ResNet50 trained on ImageNet. A binary watermark of M bits is inserted, according to a secret key represented by a random matrix $A \in \mathbb{R}^{N_l \times M}$. To embed the watermark, the output of the watermarked layer is estimated by a Gaussian mixture and two regularization terms are designed: the first one selects the Gaussian laws to be watermarked, while the second one, only activated for a subset of the training, minimizes the distance between the projection of those laws on the key and the watermark. Detection is performed by adapting the concepts in [3]. Robustness is checked against fine-tuning (up to 15% of the total training), magnitude pruning ($\boldsymbol{T} \in [0.1; 0.99]$, and watermark overwriting (embedding, with the same method, a new watermark).

The study in [7] randomly selects a set of parameters to be watermarked from multiple layers. Three classification models (ALL-CNN-C and ResNet32 trained on CIFAR10,

and LeNet5-caffe trained on MNIST) are considered. The watermark is represented by an image whose size depends on the model size. A subset of the initial weights is replaced by the pixels in the watermark, and their location is stored to serve as a secret key. The watermark is inserted via a regularization term making the inference highly sensitive to the selected parameters (hence, keeping those parameters unchanged during the training). Mark detection is achieved by recovering the selected parameters and by computing the Pearson's correlation between the original and retrieved watermarks. The robustness is checked against fine-tuning (up to 15% of the total training) and quantization (reduce the number of bits $\boldsymbol{B} \in [2; 16]$ representing the parameters).

II-C. Problem statement

The state-of-the-art analysis highlights two types of limitations in the white-box watermarking landscape. First, the robustness investigation preponderantly considers fine-tuning, pruning, and quantization, all belonging to the class of removal attacks. This originates the first question our study deals with: **"Do geometric attacks exist for NN watermarking? If so, how can they be handled?"**. Second, the application scope is generally restricted to image classification [3], [6], [7]. Moreover, [3] and [7] are *a priori* prone to be generalized to other application domains, while [6] is intimately connected to the classification task, and its conceptual generalization is not straightforward. So, the second question our study deals with is: **"Is NN watermarking restricted to classification tasks, or can it be effectively extended to other tasks? If so, is the robustness property modified?"**

III. GEOMETRIC ATTACKS TO NEURAL NETWORK

This work investigates i) whether geometric attacks can be defined for NN white-box watermarking, and ii) how can they be counter-attacked.

III-A. White-box permutation attacks

By definition, geometric attacks try to desynchronize the detector by altering the locations conveying the watermark. NNs are exposed to geometric attacks because they have many symmetrical, equi-loss representations that can be generated by a random *neuron permutation* within a layer, without affecting the neurons' functions. A corresponding permutation should also be applied to the input channel of the next layer (further referred to as *channel permutation*). Therefore, ensuring *a posteriori* resynchronization of neurons within a layer is a challenge in itself [11]. The process of permuting in-layer neurons can be accommodated by the following equations:

$$\boldsymbol{w}_{l,c,-}^{\pi_l} = \left\langle P_{\pi_l}, (\boldsymbol{w}_{l,c,-})^T \right\rangle \ \forall c, \tag{1}$$

$$\boldsymbol{w}_{l+1,-,n}^{\pi_l} = \left\langle P_{\pi_l}, (\boldsymbol{w}_{l+1,-,n})^T \right\rangle \ \forall n, \tag{2}$$

with $\boldsymbol{w}_l \in \mathbb{R}^{N_{l-1} \times N_l}$ being the weights for the l-th layer, $P(\pi_l)$ the applied permutation, $\langle \cdot \rangle$ inner product, and $(\cdot)^T$ the transpose operator. The equations above were derived for a single fully-connected layer without biases; yet, they can be extended to any other layer typology. This process can also be applied to any pair of consecutive layers.

In order to establish whether state-of-the-art white box methods are *a priori* robust against neuron permutation, they should be confronted to Eq. (1) and/or Eq. (2), as follows.

In [3], the detection is done by projecting the weights of the flattened watermarked layer on the secret key. Consequently, the neuron permutation on the l-th layer has no impact on detection. However, if the neuron permutation is applied to the $l-1$-th layer, the resulting channel permutation will completely desynchronize the watermark.

In [6], the detection is done by projecting the output of the watermarked layer on the secret key. Consequently, a complementary behavior with respect to [3] is encountered: the neuron permutation completely destroys the synchronization while the channel permutation preserves the synchronization.

In [7], the detection is done by using the secret key to locate the watermarked weights; hence, both neuron and channel permutations are likely to destroy the detection synchronization.

The above analysis demonstrates that Eq. (1) **and/or Eq.** (2) **stand for effective geometric NN watermarking attacks, as they jointly meet all the unauthorized user expectancies: (1) they succeed in destroying the mark detection, (2) they have no impact in the imperceptibility, as they preserve the watermarked NN output, and (3) they introduce no additional computational cost (in the sense that they just relate to the NN model representation and do not require any inference-related computation).** As a preliminary step towards ensuring robustness against this new type of attack, the possibility of defining counter-attack methods is investigated hereafter.

III-B. White-box permutation counter-attack

A posteriori resynchronization of neurons inside an NN layer subjected to neuron permutation is, in its general form, an exhaustive search problem in the space of factorial (over the number of neurons in the permuted layer) dimension. Regardless of the potential solution, the problem of recovering the original order for permuted neurons becomes even more complex for NN watermarking, when permuted neurons can also be modified by other types of attacks. The preliminary solution presented in [11] was not designed to be effective when supplementary operations (*e.g.* fine-tuning) are applied on the permuted neurons, while [9] targets the specification of a generic counterattack against the permutation. The advanced counterattack is based on creating a trigger set that differentiates one neuron from another and thus resynchronizes the model before retrieving the watermark. During the experiments, the permutation attack is applied to the

first or the second hidden layer of ResNet18 and ResNet50, with 160 elements in the trigger dataset. The performance of the counterattack is assessed by evaluating the BER (bit error rate) between the inserted and the retrieved watermark. The authors consider the counterattack successful for an experimental configuration when the BER is lower than 0.4 making the capacity of the original method reduced to 1 bit, indicating whether the watermark is inserted or not. The results show that the advanced counterattack has highly sensitive chances of success, depending on the experimental conditions. From a security point of view when using the same configuration as in [9], the information to be protected is the third layer of a ResNet-18, $\boldsymbol{w}_l \in \mathbb{R}^{64\times128\times3\times3}$ that has $73,728$ elements. According to [9], 160 Trigger inputs of size 32×32 (that is, $163,840$ elements) are created and should be kept secret. Hence, [9] requires at least twice more information to be kept secret than the information that is protected.

Our proposed counterattack consists of computing the cosine similarity between the un-attacked model, which is already public, and the attacked model. Indeed, despite the redundancy known to exist in NN models, we can expect the cosine similarity $S_C(\boldsymbol{w}_{l,i}, \boldsymbol{w}_{l,i}) = 1$, and hence, to have the following equation:

$$S_C(\boldsymbol{w}_{l,i}, \boldsymbol{w}_{l,i}) > S_C(\boldsymbol{w}_{l,i}, \boldsymbol{w}_{l,j}) \quad \forall j \neq i. \tag{3}$$

The original positions can be recovered by building the permutation matrix P_{π_l}:

$$(P_{\pi_l})_{i,j} = \begin{cases} 1 & j = \text{argmax}_k \left[S_C(\boldsymbol{w}_l, \boldsymbol{w}_l^{\pi_l})\right] \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

with $\boldsymbol{w}_l^{\pi_l}$ being a permuted version of the original weights. **To conclude with, Eq.** (3) **and Eq.** (4) **ensure effective reversion of the permutations described by Eq.** (1) **and Eq.** (2)**, and they can serve as a theoretical counter-attack in NN watermarking. Yet, there is no *a priori* ground about their behavior when several types of attacks are combined (*e.g.* permutation and fine-tuning), and an in-depth, complementary experimental study is required.**

IV. EXPERIMENTAL STUDY

This section presents a global yet detailed investigation of the robustness property. Section IV-A presents the experimental testbed, Section IV-B the results related to the robustness property in absence of any counter-attack, while Section IV-C illustrates the relevance of the geometric counter-attack.

IV-A. Experimental testbed

Watermarking methods and tasks. Three state-of-the-art methods are considered [3], [6], [7]. As explained in Section II, [3], [7] can be extended from classification towards image segmentation and video compression tasks, and they will be studied accordingly. In each case, the data payload and the imperceptibility are kept from their references. For each task, the imperceptibility criterion is provided by validation metrics considered during their training (*cf.* paragraph here-after). For [3] and [6], the watermark is inserted in one of the biggest convolutional layers and the penultimate layer, respectively; for [7] the watermarked weights are randomly selected through the whole model, respectively.

Watermarked architectures and training datasets. According to the three tasks, the watermarking methods are applied to four NN architectures trained on three datasets, namely: (1) VGG-16 and ResNet34 trained on CIFAR-10 for image classification, (2) DeepLabV3 [1] trained CityScapes [12] for image segmentation, and (3) DVC [2] trained on Vimeo-90k [13] and tested on UVG-dataset [14] for video compression. For the three tasks, the corresponding validation metrics are: (1) top-1 classification error, (2) the complementary mean Intersection over Unions (mIoU), and (3) the mean rate distortion *vs.* image quality, expressed in bit per pixel for a prescribed Multi-Scale Structural Similarity (bpp/MS-SSIM).

Attack parameters. First, four removal attacks are considered: Gaussian noise addition ($\mathcal{N}(0, \sigma_l \cdot \Omega)$, with $\Omega \in [0.01; 0.6]$, where σ_l is the standard deviation of the l-th layer), pruning (remove the $\boldsymbol{T} \in [0.1 : 0.99]$ fraction of the smallest weights in terms of $L1$-norm), fine-tuning (resume the training for up to 5% of the original number of iterations), and quantization (reduce the number of bits $\boldsymbol{B} \in [2; 16]$ used to represents the parameters). These attacks have been applied to the watermarked layers for [3] and [6]; this corresponds to the worst possible case for the authorized user, in the sense that, for a given imperceptibility value, they would provide the most harmful effects. In the case of [7], the attacks are applied over all the layers (as the mark is spread over an arbitrary, unknown, number of layers). In this case, in order to keep a fair comparison with [3] and [6], we target to keep constant the total amount of attacks induced in the watermarked NN, by adjusting the attack parameters accordingly, as detailed in Section IV-B. Second, the geometric attack and its counter-attack are applied to each and every layer in the NN.

IV-B. White-box robustness against attacks

The experimental results consider all the working configurations mentioned above and are illustrated in Table I.

In Table I, rows are first grouped according to the type of watermarked architecture (VGG16, ResNet34, DeepLabV3, DVC). Next, for each architecture, the rows are labeled according to the watermarking method. Columns are of three types, and provide information about the NN model in absence of any watermarking operation, on the watermarked NN in absence of any attack, and on the attacked watermarked NN. The first column is of the first type and presents the baseline performance of the NN model. The

Table I: Robustness evaluation for the different methods and architectures. For each combination, the parameter gives the value for an attack, imperceptibility is the performance on the validation set, and robustness corresponds to the watermarking metrics (C-BER and Pearson correlation coefficient) multiplied by 100. Blue box enlights a successful attack.

					Watermarked and attacked													
		Baseline	Watermarked		Gaussian			Pruning			Fine tuning			Quantization			Permutation	
		Perf.	Perf.	Rob.	Param.	Perf.	Rob.	Param.	Perf.	Rob.	Param.	Perf.	Rob.	Param.	Perf.	Rob.	Perf.	Rob.
VGG16 (classification)	[3]	11.01	11.86	100	0.6	31.4	100	0.99	25.15	100	5	11.79	100	2	90	100	11.86	**49**
	[6]		11.19	100	0.6	25.9	100	0.99	18.54	100	5	11.9	100	2	89.99	50	11.19	**62.5**
	[7]		12.49	99.99	0.06	24.6	73.97	0.9	77.22	99.99	5	8.81	99.99	2	80.77	99.99	12.49	**22.48**
ResNet34 (classification)	[3]	9.76	10.14	100	0.6	14.24	100	0.99	10.14	100	5	8.27	100	2	89.68	100	10.14	**55**
	[6]		8.72	100	0.6	19.35	100	0.99	23.38	100	5	8.21	100	2	90	37.5	8.72	**68.75**
	[7]		11.14	99.99	0.06	30.44	80.23	0.9	90	99.99	5	11.02	99.99	2	91.99	98.99	11.14	**11.04**
DeepLabV3 (segmentation)	[3]	33.1	36.23	100	0.6	42.51	100	0.99	94.72	100	5	36.46	100	2	97.99	62.5	36.23	**68.75**
	[7]		30.14	99.99	0.06	42.54	99.35	0.9	99.99	99.99	5	29.99	99.91	2	99.86	98.99	30.14	**46.39**
DVC (compression)	[3]	0.23/0.97	0.24/0.97	100	0.6	0.25/0.97	100	0.99	0.24/0.97	100	5	0.23/0.97	100	2	13.62/0.11	87.5	0.24/0.97	**37.5**
	[7]		0.23/0.97	99.99	0.06	5.81/0.21	62.56	0.9	0.50/0.63	99.99	5	0.23/0.97	93.62	2	13.62/0.31	99.99	0.23/0.97	**49.57**

next two columns are of the second type and provide the performance of the NN (according to the corresponding validation metric, IV-A) and the Robustness. The differences between the inserted and the recovered watermarks are expressed as complementary BER, denoted by C-BER and computed as ($C-BER = (1-BER)\times 100$) for [3], [6] and as Pearson coefficient (multiplied by 100) for [7]. The other columns are of the third type and are sub-grouped according to each investigated attack. In addition to performance and robustness, the "attacks parameter" is provided, except for the permutation attack where it is irrelevant. For each combination, Table I provides the parameter value for which the watermark can no longer be retrieved or, if the watermark fully withstands the set of values presented in Section IV-A, the value corresponding to the strongest attack. Note that information about the imperceptibility can be obtained by comparing the values of performance between the watermarked and baseline columns; similarly, information about the impact of the attacks in imperceptibility can be obtained by comparing the values of performance between the attack and watermarked columns.

Several conclusions can be drawn from Table I.

First, by comparing the Watermarked and Baseline columns, it is shown that, at least in absence of attacks, the application field of NN watermarking can be extended from classification to segmentation and compression. This conclusion is based on the fact that the three tasks result in quite an equal impact on performance. For classification, the relative differences in performance can be computed from the values presented in Table I; they range between -0.1 and 0.14, with an average of 0.05. Such values become -0.1, 0.1 and 0 for segmentation. In the case of video compression, while the MS-SSIM is constant, the relative variations in bpp become 0, 0.04, and 0.02. Note that actually the regularisation term included in [6], [7] for watermarking purposes also has a beneficial impact on the NN performance that can be increased with respect to the baseline. In each and every case, the watermark can be recovered (C-BER $= 100\%$ and Person's coefficient $= 0.99$). This opens the door to studies devoted to specific NN watermarking methods for segmentation and coding tasks.

Secondly, for each investigated NN and watermarking method, the robustness against the removal attacks is met, as either the watermark can be retrieved or the performance is lowered beyond the application purpose. In this respect, for any of the three tasks, the Gaussian, pruning, and fine-tuning attacks do not have any impact on the watermark detection, as demonstrated by values C-BER $= 100\%$ and Person's coefficient > 0.6. When considering the quantization attack, the watermark can be lost (C-BER $\leq 90\%$) but the performance decreased beyond the application requirements; just for illustration, in the case, [6] and VGG16 architecture, C-BER $= 50$ but the top-1 error becomes $= 89.99$. Similar behavior is encountered for [6] on ResNet34 and [3] on DVC.

In contrast to removal attacks, the geometric attack is always successful: for the same performance as the watermarked model, the mark cannot be anymore detected (C-BER $\geq 70\%$ and Person's coefficient ≤ 0.5). Hence, the effectiveness of the geometric attack defined by Eq. (1) and Eq. (2) is demonstrated, and the need for evaluating the counter-attack defined by Eq. (3) and Eq. (4) is proved.

IV-C. Geometric counter-attack performance

The counter-attack to geometric modifications is applied to each of the working configurations investigated in the previous sub-section. The results are synoptically displayed in Fig. 2 for [3], [6] and in Fig. 3 for [7].

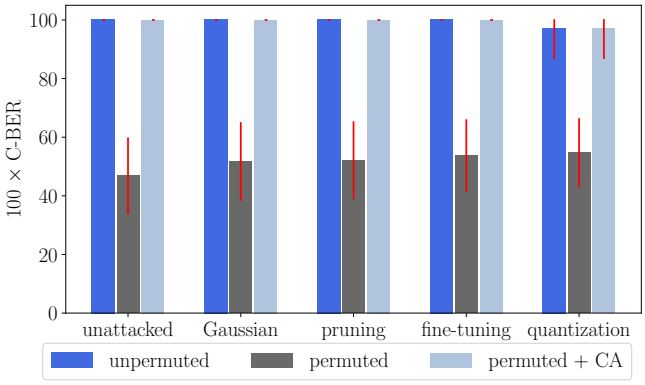


Fig. 2: Robustness evaluation of geometric counter-attack against the different removal attacks for methods [3], [6].

Each of these two figures is structured in five areas. The first area corresponds to the case when no removal attack is applied on the watermarked NN. The other 4 areas correspond to the cases when the four removal attacks are individually applied (from left to right: Gaussian noise addition, pruning, fine-tuning, and quantization, respectively). In its turn, each of these 5 areas shows three bars corresponding to the cases of: no additional geometric attack - labeled by (unpermuted), an additional geometric attack - labeled by (permuted), and an additional geometric attack followed by its counter-attack - labeled by (permuted+CA).

While the abscissas are identical for these two figures, their ordinates are different. Figure 2 provides average C-BER values (multiplied by 100) and their related $\pm$ standard deviation intervals (bounded at the maximum theoretical value of 100). The averages are computed over all the NN architectures, all the investigated attack parameters, and the methods in [3], [6]; the standard deviation is computed as an unbiased estimator over the same data. In Fig. 3, the coordinate corresponds to the Person's coefficient (multiplied by 100) and also presents average and $\pm$ standard deviation intervals (bounded at the maximum value of 100); this time, the average is computed only for [7], over all the NN architectures and all the investigated attack parameters; the standard deviation is also computed as an unbiased estimator Fig. 2 and Fig. 3 demonstrate that Eq. (3) and Eq. (4) are effective geometric counter-attacks: they can synchronize back the mark detection even when the geometric attack is applied in conjunction with any of the 4 investigated removal attacks.

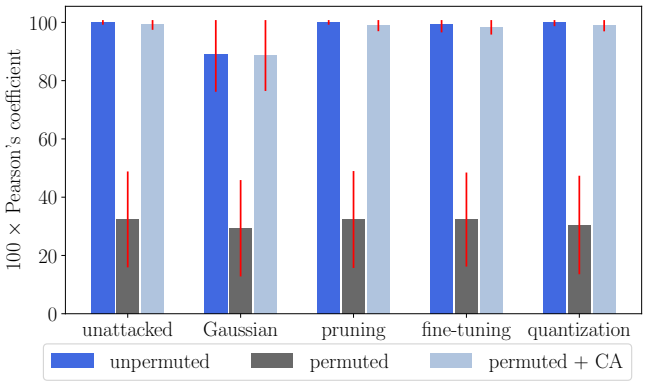


Fig. 3: Robustness evaluation of geometric counter-attack against the different removal attacks for method [7].

V. CONCLUSION

The present paper presents an in-depth investigation of NN watermarking robustness. First, it shows that the neuron and channel permutation operations can be transposed into an effective, new type of attack (the first in the geometric attacks family), and provides the matched counter-attack. Secondly, it demonstrates that the counter-attack is effective in ensuring robustness when the geometric attack is applied by itself or in conjunction with any of the four state-of-the-art removal attacks (Gaussian noise addition, pruning, fine-tuning, and quantization). As a side result, the study establishes that the NN watermarking scope can be extended from classification tasks to segmentation and compression, and identifies the performance gap to be bridged by future methods. Finally, the level of detail of the quantitative results presented in the study can provide guiding information for an experimenter who would like to get to a practical NN watermarking solution. Future work will be devoted to investigating the coupling of several types of removal attacks as well as to identifying the potential synergies and anatomies when coupling removal, geometric and cryptography attacks. Extending the principle from this study to devise a generic regularisation term that can be dynamically used as a counter-attack is also part of future work.

VI. REFERENCES

[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[2] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of CVPR*, 2019.

[3] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding watermarks into deep neural networks," in *Proceedings of ICMR*. jun 2017, ACM.

[4] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *USENIX Security*, 2018.

[5] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*, Morgan kaufmann, 2007.

[6] B. Darvish Rouhani, H. Chen, and F. Koushanfar, "Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks," in *in proceedings of ASPLOS*, 2019.

[7] E. Tartaglione, M. Grangetto, D. Cavagnino, and M. Botta, "Delving in the loss landscape to embed robust watermarks into neural networks," in *2020 25th ICPR*. IEEE, 2021.

[8] E. Le Merrer, P. Perez, and G. Trédan, "Adversarial frontier stitching for remote neural network watermarking," *Neural Computing and Applications*, vol. 32, no. 13, 2020.

[9] Fang-Qi Li, Shi-Lin Wang, and Yun Zhu, "Fostering the robustness of white-box deep neural network watermarks by neuron alignment," in *ICASSP 2022-2022 IEEE*. IEEE, 2022.

[10] A. G. Konheim, *Cryptography, a Primer*, John Wiley & Sons, Inc., USA, 1st edition, 1981.

[11] K. Ganju, Q. Wang, W. Yang, C. A Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *Proceedings of ACM SIGSAC*, 2018.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on CVPR*, 2016.

[13] T. Xue, B. Chen, J. Wu, D. Wei, and W. T Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, 2019.

[14] A. Mercat, M. Viitanen, and J. Vanne, "Uvg dataset: 50/120fps 4k sequences for video codec analysis and development," in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020.

Classical approaches and new deep learning trends to assist in accurately and efficiently diagnosing ear disease from otoscopic images

Dhruv Chetan Jobanputra
Computer Science and Software Engineering
The University of Western Australia
Perth, Australia
dhruvjobanputra8@gmail.com

Mohammed Bennamoun
Computer Science and Software Engineering
The University of Western Australia
Perth, Australia
mohammed.bennamoun@uwa.edu.au

Farid Boussaid
Electrical, Electronic and Computer Engineering
The University of Western Australia
Perth, Australia
farid.boussaid@uwa.edu.au

Lian Xu
Computer Science and Software Engineering
The University of Western Australia
Perth, Australia
farid.boussaid@uwa.edu.au

Jafri Kuthubutheen
Division of Surgery, Medical School
The University of Western Australia
Perth, Australia
jafri.kuthubutheen@health.wa.gov.au

***Abstract*—Rural communities in Australia have limited access to Ear, Nose and Throat (ENT) specialists, resulting in a lack of expertise to provide a diagnosis of complex and chronic ear diseases. This literature review examines previous attempts at creating a computer-aided tool to accurately diagnose ear disease and gaps in the literature. A systematic search was conducted to identify relevant papers and the latest best trends in technology. Four papers showed significant results in ear disease detection with deep learning models providing the best performance. Some studies using larger datasets consisting of endoscopic images obtained accuracies of over 90%. No adequate model was found that used otoscopic images with a sensitivity of over 90%. Endoscopic images provide better quality images, making it unclear how the models would perform on otoscopic images. Advanced techniques such as Transformers have not yet been tested in ear disease detection and could help improve model accuracy.**

***Index Terms*—Convolutional Neural Network, Deep Learning, Image Classification, Otology, Transformers**

I. Introduction

A. Background

A lack of Ear, Nose, and Throat (ENT) specialists, particularly in rural and remote regions of Australia, results in a shortage of specialist-led diagnoses of ear diseases, and in particular, otitis media. Ear diseases are most prevalent amongst children, with approximately five out of six children having an ear infection by the age of three [1]. With such a high prevalence rate, especially amongst Aboriginal and Torres Strait Island populations (ATSI), the ability to accurately diagnose ear conditions is crucial to ensure prompt and appropriate treatment is provided.

Whilst obtaining an accurate diagnosis and initiating treatment according to well-established protocols and guidelines seem straightforward, diseases can be misdiagnosed up to 50% of the time using an otoscope [2]. A study conducted in 2001 [3] evaluated the accuracy of pediatricians and Otolaryngologists (ENTs) in diagnosing ear diseases when using a 30-second video taken from an otoscope. Participants were asked to distinguish between one of four possible conditions: Acute Otitis Media (AOM), Otitis Media with Effusion (OME), retracted but otherwise normal tympanic membranes, and normal ears. A total of 514 pediatricians and 188 ENTs were tested on nine different cases, with the results shown in *table I*. The participants' experience in their field was assessed. While there was on average over 90% accuracy for both types of practitioners for detecting abnormalities, the accuracy for correctly diagnosing the individual disease was much lower. Pediatricians on average made a correct diagnosis only 50% of the time, while the ENTs on average made a correct diagnosis 73% of the time. The videos shown to the participants were taken after the ear canals of the patients' had been completely cleared of any wax, and coupled with the long length of viewing time (30 seconds) suggests that in real-life clinical scenarios, accuracies would be even lower than seen in this experiment [3]. Furthermore, this study used videos rather than still otoscopic images, which should have helped as getting a single clear image recorded of the eardrum can be difficult. Some conditions are known to be particularly difficult to identify, such as otitis media with effusion [4], and including these often challenging conditions can reduce the overall diagnostic accuracy reported by a study. This study highlights that diagnosing ear diseases accurately can be a challenging task, even for an experienced practitioner.

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

Video Examination No.	Correct Diagnosis	Pediatricians (n = 524)	ENTs (n = 188)
1	OME	48	88
2	OME	45	69
3	Retracted TM, otherwise normal	56	76
4	AOM	73	76
5	OME	50	79
6	OME	25	48
7	Retracted TM, otherwise normal	46	83
8	OME	48	84
9	Retracted TM, otherwise normal	59	65
	Overall	50	73

TABLE I
THE RESULTS OF THE STUDY ASSESSING PEDIATRICIANS AND ENTS. [3]

B. Problem statement and specifications

The primary challenge encountered in the diagnosis of ear disease is that the interpretation of the tympanic membrane appearance, especially by novice or less experienced health-care practitioners, is often fraught with difficulty and may be inaccurate. Additionally, there is a shortage of more advanced tools such as otoendoscopes, which can provide a wider field image with increased brightness and clarity, as they are not commonly available in primary health care and often require specialised training to operate. Whilst otoscopy is commonly performed by most health care practitioners, there is a potential gap in accurately interpreting images. The common metrics used to measure the performance of a particular diagnosis are its sensitivity and specificity. Sensitivity refers to the proportion of patients with the disease that are correctly identified as having that disease, while specificity corresponds to the proportion of patients without any disease that are correctly identified as not having a disease. As the diagnosis of ear diseases requires the correct classification of the image into one of several multiple possible causes, sensitivity refers to the proportion of patients with a particular disease that were correctly identified as having that particular disease, and specificity is essentially equivalent to the sensitivity of detecting a patient without any disease, *i.e.*, the proportion of patients with a normal eardrum.

An ideal diagnostic 'test' should have both high sensitivity and high specificity and a test is considered to have good diagnostic accuracy when both are around 90%. The 'test', or in this case, a diagnostic classification model must also be capable of classifying more challenging cases such as OME to a similarly high level. As a result, it is important to ensure that the sensitivity for each disease is equally high, rather than the average for all diseases. The overall accuracy, *i.e.* proportion of correctly classified instances, can also be a useful measure. However, in situations where there is a large proportion of normal ear drums, there is a potential for bias which can lead to a high overall accuracy overall but with a low sensitivity for some diseases. A predictive model with low sensitivity to abnormal eardrums can lead to incorrect treatment, delayed treatment and increased treatment costs, which can have long-term sequelae [5].

II. METHODOLOGY

A. Search and selection criteria

A literature search was conducted using Google Scholar with keywords including 'ear disease detection', 'deep learning', 'machine learning' and 'otoscopic images'. The aim was to identify papers that utlised computer-aided technology to diagnose ear diseases from otoscopic images of the eardrum, captured through otoscopic tools. The scope of computer-aided technology included traditional computer vision techniques, deep learning and machine learning.

Twenty papers closely related to this topic were initially obtained through the search. The papers' ranged from 2016 to 2022, indicating the research in this field is relatively new and still has room for exploration. Each paper was evaluated based on its publication date, the dataset used, methodology and overall results. Four papers were chosen for further analysis. A 2016 paper by Myburgh *et al.* [6] was selected for its status as the earliest study to use computer-aided technology for ear disease detection as well as its use of traditional computer vision techniques in conjunction with machine learning. The remaining papers employed deep learning techniques using CNN (Convolutional Neural Network) models. The paper by Cha *et al.* [7] used a large dataset and obtained a high level of accuracy. The paper by Zeng *et al.* [8] was an improvement over previous studies in terms of dataset size and the final accuracy. The final paper, by Habib *et al.* [9] was the most recent study (2022) and was also an Australian-based study. The dataset used was obtained from remote areas of Australia and was a more representative dataset than the previous studies.

III. RESULTS

A. Analysis of papers

The study conducted by Myburgh *et al.* [6] in 2016 was among the very first to use machine learning for ear disease diagnosis. They aimed to develop computer-aided technologies that could diagnose ear diseases in developing countries, where access to medical doctors and health personnel is limited. They collected 391 images for training representing 5 different classes of diseases: *cerumen, normal Tympanic Membrane (TM), AOM, OME and Chronic Suppurative Otitis Media (CSOM)*, *see Figure 1*. The images had a resolution of at least 500×500. With this limited training set, they achieved an average accuracy of 80.6% using a traditional machine learning technique, *i.e.,* decision trees. To extract the feature list, they applied numerous computer vision operations such as edge detection and colour detection to identify specific features in the image. They achieved a sensitivity ranging from 79% to 82% for each class. However, this study has two limitations: (**i**) it used a small dataset, which may not adequately represent real-life data; (**ii**) the chosen model was a decision tree, which is fixed and would need to be redesigned if more data becomes available. Despite these limitations, the study demonstrated that even with a small training dataset,

relatively good accuracy can be achieved through traditional machine learning and computer vision approaches.

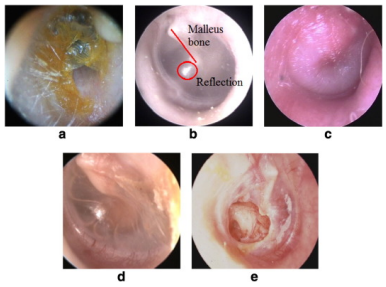


Fig. 1. The five classifications of ear diseases respectively - Cerumen, Normal, AOM, OME, CSOM. Myburgh *et al.* [6]

In 2019, Cha *et al.* [7] conducted a study to address a similar problem, but using a large dataset. They collected 10,544 images, of which 80% (8435) were used for training and the remaining for testing. The resolution of the images was 640×480 pixels. This dataset was collected over four years using otoendoscopic images and it represented a significant improvement in terms of the data size compared to previous studies, such as the one by Myburgh *et al.* [6]. As a result, they were able to achieve a high average accuracy of 93.73% using an ensemble deep learning model - a combined model of two publicly available pre-trained models, *i.e.,* InceptionV3 and ResNet101. To implement the method, they followed a series of steps, including data processing, training individual models and finally combining the best two models to create their final ensemble model. The data processing consisted of labelling the images into relevant categories. From the 10,544 images, they separated the images into six different classes: *normal, tumour, OME, myringitis, TM perforation and attic retraction*. The classes did not necessarily represent one specific disease, rather they combined similar diseases under one umbrella class since some ear diseases only had a few samples. The grouping of similar ear diseases was based on the similarity in diagnosis and treatment. Sample images of the classes are shown in *Figure 2*.

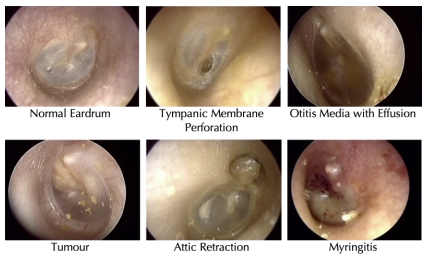


Fig. 2. The six classifications of ear diseases. Cha *et al.* [7]

The dataset was then augmented by applying random rotations, translations, scaling and flipping. To obtain the final models, they first selected nine pre-trained CNN models and evaluated each model individually on the validation set to obtain the accuracy and calculation time, as shown in *Table II*. Subsequently, the two best-performing models were then combined to create an ensemble model. Ensemble models can often combine weaker models to create a more powerful and accurate model. The models this paper proposed to combine were InceptionV3 and ResNet-101. The final accuracy on the entire training set of 8435 images, was 93.73%. This is a relatively high accuracy and can partly be attributed to the large size of the training dataset and the large bias towards normal images, which represented approximately 41% of the dataset. The study also tried different data sizes to test the effects of changing the size of the training data. They conducted three tests using different numbers of randomly sampled training images, *i.e.,* 2k, 5k images, and the full training set of 8435 images, respectively. These tests were performed using each of the nine individual models. The models using 2k training images obtained accuracies ranging from 68.2% to 84.1%, while the models using 5k images resulted in accuracies ranging from 82.8% to 89.5%. Using the full training set resulted in accuracies ranging from 85.6% to 92.1%. Other performance metrics, except training time, were not considered. The results are displayed in *Table II*. This demonstrates that having a larger dataset can increase the accuracy by a significant margin. It also shows that the final ensemble model had greater accuracy than any individual model alone. The positive outcomes of their research were the high accuracy achieved, and also the large dataset that was collected. However, the limitation of this work is that as the model was trained on otoendoscopic images, it may not perform well on standard otoscopic images. Although this dataset is large, it is not an accurate representation of the real-world scenario for rural regions. Another limitation was the sensitivity across all classes was not necessarily high. The sensitivity metric was computed for the final two models and the ensemble classifier. While the ensemble classifier still reported a higher sensitivity than either base model for most classes, the lowest sensitivity recorded for the ensemble classifier was 77.9% for the *myringitis* class. This is much lower than the target value of 90%. Most other classes also reported sensitivity values lower than 90%.

Transferred models	Accuracy Full	Full- H25	Quarter	Half	GPU time (seconds)	Parameters (millions)	Number of layers
SqueezeNet	85.55	85.5	73.5	82.8	4137	1.24	68
Alexnet	87.2	83.6	73.7	82.6	3805	61	25
ResNet18	90.65	90.2	83.4	86	4256	11.7	72
MobileNet-v2	90.75	89.8	79.9	84.9	7032	3.5	155
GoogLeNet	90.9	88.7	68.2	85.5	5104	7	144
Resnet50	91.2	91.4	81.3	86.3	7302	25.6	177
Resnet101	91.55	91.7	83.6	86.1	12,215	44.6	347
Inception-v3	92	92.1	84.1	89.5	11,938	23.9	316
InceptionResnet-	92.1	91.9	82.2	86.9	33,283	55.9	825

TABLE II
THE RESULTS AFTER TRAINING NINE CNN MODELS ON DIFFERENT SIZED DATASETS. CHA *et al.* [7]

The next relatively large study was conducted in 2021 by Zeng *et al.* [8], building upon the previous work in [7] by using even more data and testing out some different pre-trained CNN models. They collected 20,542 otoendoscopic images over three years and classified the ear diseases into 8 different classes including *normal, cholesteatoma, CSOM, External Auditory Canal (EAC) bleeding, cerumen, Otitis Externa (OE),*

OME and Tympanosclerosis. The resolution of these images was 448×448 pixels. Samples of the images are shown in *Figure 3*.

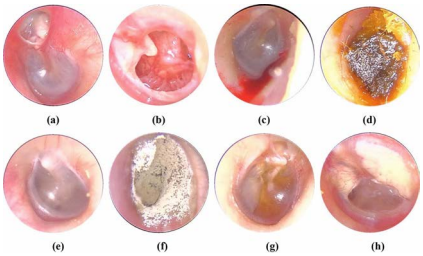


Fig. 3. The 8 classifications of ear diseases respectively - Cholesteatoma, CSOM, EAC Bleeding, Cerumen, Normal, OE, OME, Tympanosclerosis. Zeng *et al.* [8]

They used an 80:20 training/testing split on the overall dataset and achieved an average accuracy of 95.59% using an ensemble model consisting of two models, *i.e.,* DenseNet-BC161 and DenseNet-BC1615. This method is similar to the one used in the previous study [7]. Initially, they tested the dataset against nine individual deep learning models and ranked them based on their accuracy and efficiency, which was measured by their training time, *as shown in Table III*. From this, they selected the top two models and combined them to create an ensemble model to make a more powerful model.
In addition to developing the model, they also built a real-time deep learning system to be used in the clinical workplace. Although they were not able to test their system in a clinical environment with real patients at the time of publication, they claim that the system works in almost real-time on the testing data. The positive aspects of this study were the high accuracy and sensitivity that they were able to achieve and the system they built. Unlike the previous study, the sensitivity values for this study ranged from 90.82% to 100%. However, as with the past study, these results are highly attributed to the dataset and quality of images used. The dataset they were able to obtain was almost double that of the aforementioned study and is far greater than any other study conducted to date. The images obtained were taken via otoendoscopes and so while this study had impressive results, it is not conclusive how this model would perform in a rural environment. However, it does demonstrate that with a large enough dataset, high accuracy and sensitivity are achievable.

The study by Habib *et al.* [9] conducted in 2022 is the most recent study on the topic of the detection of ear diseases. This study was conducted in Australia using a dataset collected from "Aboriginal and Torres Strait Islander children who underwent tele-otology ear health screening in the Northern Territory, Australia, between 2010 and 2018" [9]. Although the data was collected only for children, this dataset is representative of the real-world scenario as it was both taken by an otoscope and from a rural region. A total of 6527 otoscopic images were used to train and test the model. More specifically, 5297 images were used for training and 600 were used for testing; approximately a 91:9 split between training

Transferred models	Accuracy	GPU time (s)	Parameters	Processing time (s)[??]
MoblieNet-V2	93.455	27,240	2,235,200	0.0374
MoblieNet-V3	93.884	24,758	2,946,622	0.0357
Inception-V4	93	98,270	42,681,353	0.1309
ResNet50	93.581	51,098	25,557,032	0.0668
ResNet101	93.632	78,844	42,516,552	0.1099
Inception-ResNet-V2	94.617	111,849	54,318,760	0.1604
DensNet-BC121	94.188	54,192	6,962,056	0.0859
DensNet-BC161	94.564	78,453	26,489,672	0.1707
DensNet-BC169	94.541	56,477	12,497,800	0.109
DensenetBC1215	94.364	56,079	7,548,920	0.0809
DensenetBC1615	95.099	80,895	27,893,456	0.4512
DensenetBC1695	94.339	58,209	13,122,040	0.1318
Ensemble				0.5708

TABLE III
THE RESULTS OF TRAINING THE INDIVIDUAL MODELS ON THE TRAINING SET. ZENG *et al.* [8]

and testing. The resolution of these images was not reported. The dataset was classified into five classes that included *Normal, OME, AOM, CSOM and cerumen*. Sample images are shown in *Figure 4*.

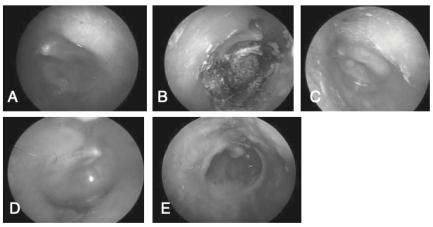


Fig. 4. The five classifications of ear diseases respectively - Normal, Cerumen, AOM, OME, CSOM. Habib *et al.* [9]

Their model was created by using the Custom Vision API provided by Microsoft Azure Custom Vision, and used CNN models with a ResNet backbone to create a deep learning model using transfer learning - which used pre-trained weights from ImageNet. When using all images and classes for their classification, they achieved a test accuracy of 74.5%. Several other tests were also conducted by removing one or more classes of ear disease from the dataset and using the rest of the data for training and testing. When not using images of the "OME" class, the accuracy was 92.8%. This brings an increase of about 17%, indicating that the model had trouble classifying the OME class. OME is also an example of a disease that ENTs have more trouble with. These results are displayed in *Table IV*.

Model	Categories	Classes	Total Images (n)	Training Images (n)	Test Images (n)	Test Accuracy (%)	AUC
1	Normal, OME, AOM, COM, Wax/obstructed EAC	5	6527	5927	600	74.5	0.963 (95%CI: 0.941-0.986)
2	Normal, AOM, COM, Wax/obstructed EAC	4	5125	4675	450	92.8	0.997 (95%CI: 0.994-1.000)
3	Normal, OME, AOM, COM	4	6195	5625	570	74.4	0.972 (95%CI: 0.965-0.989)

TABLE IV
THE RESULTS OF PERFORMING MULTICLASS CLASSIFICATION. HABIB *et al.* [9]

The difficulty of detecting OME accurately was also demonstrated by a series of binary classification (*i.e.*, normal *v.s.* abnormal) tests. Each class of ear disease underwent binary testing against normal images to evaluate the accuracy and sensitivity of differentiating between normal and one specific

disease. As a result, AOM, CSOM and cerumen had a test accuracy ranging from 96.3% to 99.3% and test sensitivity ranging from 90.7% to 100%. In contrast, OME only had an accuracy of 77.8% and a sensitivity of 59.3%. This indicates that even in the case of binary classification, their model did not perform well in detecting OME. This is despite having a large amount of data for the OME class, *i.e.,* 21.4% (1402) of the total images in the original dataset, which is four times greater than the number of EAC images and seven times greater than AOM classes. These results are displayed in *Table V*.

Category	Total Images (n)	Training Images (n)	Test Images (n)	Test Accuracy (%)	Test Sensitivity (%)	Test Specificity (%)	Test PPV (%)	Test NPV (%)
Wax/Obstructed EAC versus normal	3434	W/O-302 N-2852	W/O-30 N-250	98.2	100	98	85.7	100
AOM versus normal	3294	AOM-172 N-2852	AOM-20 N-250	99.3	100	99.2	90.9	100
COM versus normal	4601	COM-1349 N-2852	COM-150 N-250	96.3	90.7	99.6	99.3	94.7
OME versus normal	4504	OME-1252 N-2852	OME-150 N-250	77.8	59.3	88.8	76.1	78.5

TABLE V
THE RESULTS OF PERFORMING BINARY CLASSIFICATION. HABIB *et al.* [9]

This study demonstrates that, while utilising a dataset collected from a rural area - which aligns with the realistic scenario for rural settings - the model struggled to detect OME (a condition that occurs to a high degree in rural and remote patients) with a high level of sensitivity. Additionally, the study was limited by the small number of classes of images included in the dataset. The dataset included only five classes, while other diseases such as OE or tympanosclerosis were not taken into account.

IV. DISCUSSION

A. Current state-of-the-art technologies for image classification

The latest research on image classification has shown that deep learning is the future of this field. Although previous ear disease detection studies [7]–[9] have used deep learning models, new and more powerful models have emerged since then. Transfer learning is a common and efficient way to use these models in medical applications. Due to privacy and ethical concerns, there is a lack of readily available data, which means pre-trained models are needed to ensure accurate results. Creating a model from scratch will not yield the desired level of accuracy when used in real-world settings, as the model weights would not have been tuned to a sufficient degree. Transfer learning uses pre-trained models that have already had their weights and biases turned for accuracy, and alters the last few layers to fit the output requirements.

The traditional practice for deep learning in image classification is to use CNN models, however, a paper in 2017 [10] introduced the idea of using a self-attention architecture, known as transformers. CNNs extract features from an image, but do not take into account the correlation between the features. The self-attention architecture (transformer), is designed to improve upon the traditional CNN model by considering the relationships between features within an image, rather than solely extracting features individually. This allows for a more comprehensive understanding of the image as a whole. These architectures have been shown promise in the area of computer vision and image classification and have been shown to perform better than CNN models. The self-attention architecture also opens up a pathway to image captioning which includes providing a textual description of an image rather than just producing a simple classification. The state-of-the-art models as listed on *Papers With Code* [11], are typically transformers, or a combination of transformer and CNN models, and have a top-1 accuracy of over 90% when trained on large image datasets such as ImageNet. For comparison, InceptionV3, a model used by the 2019 study by Cha *et al.* [7] has a top-1 accuracy of 78.95% while a transformer type architecture such as ViT-G/14 has shown an accuracy of 90.45% on ImageNet. The significant increase in accuracy compared to traditional CNN models highlights the power of these architectures.

B. Deep learning in medicine

Deep learning has overtaken traditional computer vision techniques in the field of medicine rapidly in recent years. A study in 2020 [12] reviewed the use deep learning in various aspects of medicine including target detection, segmentation, classification and registration. The review found CNN based deep learning techniques to be a successful in not only finding lesions, but also in discriminating and classifying specific lesions, and at times segmenting the lesion area. The application of deep learning was also seen across various medical fields to be successful. The major shortcoming of this technique found was that training a model requires a large dataset which makes the dataset acquisition more demanding.

A study in 2022 [13] reviewed the use transformers in medical imaging and found that transformers had pervaded almost all areas of medical imaging, with segmentation and classification impacted most significantly. Though transformers have only recently been applied in medical imaging, they have already been shown to produce comparable results to state-of-the-art CNN models in areas of segmentation and classification. The rapid growth of transformers have prompted further research in the area and "despite their impressive performance, it is anticipated that there is still much exploration to be done with transformers in medical imaging" [13].

C. Publicly available datasets for ear diseases

Despite the lack of publicly accessible datasets due to privacy and ethical concerns, there are still three open-source datasets that can be utilised for training a machine learning model.

In 2019, Başaran *et al.* [14] created their own dataset for the purpose of diagnosing ear diseases through the use of a grey-level co-occurrence matrix technique and artificial neural networks. They created two datasets in total, with the first one consisting of 282 otoscopic images with seven classes; *Normal, AOM, CSOM, Cerumen, Tympanosclerosis, OE and Tube.* The second dataset created later [14] had 956 images and included two additional classes; *Foreign object and Pseudomembrane*. All the images had a resolution of 500×500 pixels, and samples of these images are displayed in *Figure 5*.

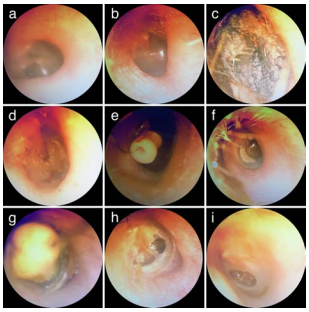


Fig. 5. The nine classes respectively; *Normal, AOM, Cerumen, OE, Tube, Tympanosclerosis, Foreign object, Pseudo-membrane and CSOM*. Başaran *et al.* [14]

In 2020, Viscaino *et al.* [15] conducted a study to test the diagnosis of ear disease using computer-aided technologies and made their dataset publicly available. The dataset consists of 880 otoscopic images taken from patients aged between 7 to 65 and includes four classes; *normal, cerumen, tympanosclerosis and CSOM*. The dataset is well balanced with 220 images for each class and each image has a resolution of 420×380 pixels. Samples of these images are displayed in *Figure 6*.

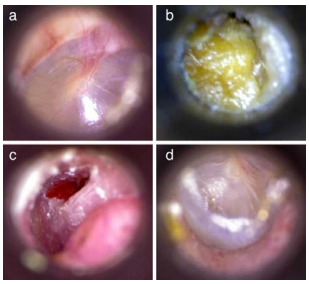


Fig. 6. The four classifications of ear diseases respectively; Normal, Cerumen, CSOM and Tympanosclerosis. Viscaino *et al.* [15]

The study by Camalan *et al.* [16] in 2020 aimed to use deep learning for ear disease detection, and they made their dataset publicly available for others to use. The dataset consists of 454 images from three classes; *normal, OME and tube*. These images were taken using an otoscope and have an approximate resolution of 900×900 pixels. However, the images have been cropped to different sizes, resulting in some having slightly lower resolution, while others have a higher resolution. Samples of these images are displayed in *Figure 7*.

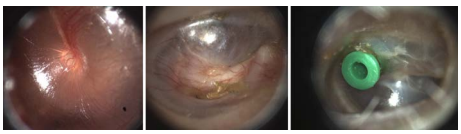


Fig. 7. The three classifications of ear diseases respectively; *Normal, OME and Tube*. Camalan *et al.* [16]

There have been some other studies that, while not making their datasets publicly available, have stated that they are able to share their datasets upon suitable request to the authors and the corresponding university. These datasets are listed in *table VI*.

Source	Dataset Size	Type	Resolution	No. of Classes	Data Sharing
Başaran *et al.* [14]	1238	Otoscopic	640×480	9	Public
Viscaino *et al.* [15]	880	Otoscopic	420×380	4	Public
Camalan *et al.* [16]	454	Otoscopic	∼900×900	3	Public
Cha *et al.* [7]	10544	Otoendoscopic	500×500	6	Requires Permission
Mothershaw *et al.* [17]	8486	Otoscopic	299×299	10	Requires Permission

TABLE VI
PUBLICLY ACCESSIBLE DATASETS

V. Conclusion

A. Research gap

While previous research has been conducted in the area of ear disease detection, there is still a need for a complete solution to accurately diagnose otoscopic images with high sensitivity for a range of different ear diseases. It is uncertain how well the model developed in the study by Zeng *et al.* [8] would perform when applied to non-endoscopic images, despite achieving a high sensitivity (>90%) for all classes when using otoendoscopic images. Otoendoscopic images do not provide a complete representation of the image quality that can be obtained by non-specialists. Studies using otoscopic images, such as Habib *et al.* [9] have reported lower sensitivity values, suggesting a significant difference in the results obtained from otoendoscopic and otoscopic images. Habib *et al.* [9] used a more representative dataset obtained from rural communities using otoscopes, but were unable to achieve sufficient sensitivity levels in diagnosing certain diseases such as OME which only had a sensitivity of 59.3% in a binary classification between OME and normal images.

B. Contribution and future work

The benefits of this technology extend beyond patients and hearing health professionals and in particular rural communities which have limited access to ENT specialists. By utilising telehealth in conjunction with this technology, patients can be protected from receiving unnecessary or incorrect treatments, thereby saving money, time and expediting the resolution of their disease. It can serve as an educational tool for both novice and experienced doctors to interpret images and improve diagnostic accuracy. Computer-aided tools can be provide a prediction and also highlight important parts of the image, providing invaluable information to clinicians. For future work it is recommended to experiment with state-of-the-art models such as transformers or newer CNN models and use large datasets by possibly combining multiple different datasets to produce a more generalised model. The next steps for this research include using videos instead of still images and providing image captioning to describe the image or video to provide a more detailed analysis.

References

[1] "Ear infections in children," 2017. [Online]. Available: https://www.nidcd.nih.gov/health/ear-infections-children

[2] "What a middle ear infection looks like using the tomi scope [otosight middle ear scope]," Mar 2021. [Online]. Available: https://photoni.care/news/what-a-middle-ear-infection-looks-like-using-the-tomi-scope/

[3] M. E. Pichichero and M. D. Poole, "Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media," *Archives of pediatrics & adolescent medicine*, vol. 155, no. 10, pp. 1137–1142, 2001. [Online]. Available: https://jamanetwork.com/journals/jamapediatrics/fullarticle/191139

[4] M. E. Pichichero, "Acute otitis media: Part 1. Improving diagnostic accuracy," *American family physician*, vol. 61, no. 7, p. 2051, 2000. [Online]. Available: https://www.aafp.org/pubs/afp/issues/2000/0401/p2051.html

[5] C. Hayes, "Why is diagnosing ear infections so hard?" Jan 2015. [Online]. Available: https://www.kevinmd.com/2015/01/diagnosing-ear-infections-hard.html

[6] H. C. Myburgh, W. H. Van Zijl, D. Swanepoel, S. Hellström, and C. Laurent, "Otitis media diagnosis for developing countries using tympanic membrane image-analysis," *EBioMedicine*, vol. 5, pp. 156–160, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352396416300500

[7] D. Cha, C. Pae, S.-B. Seong, J. Y. Choi, and H.-J. Park, "Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database," *EBioMedicine*, vol. 45, pp. 606–614, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352396419304311

[8] X. Zeng, Z. Jiang, W. Luo, H. Li, H. Li, G. Li, J. Shi, K. Wu, T. Liu, X. Lin *et al.*, "Efficient and accurate identification of ear diseases using an ensemble deep learning model," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-90345-w

[9] A.-R. Habib, G. Crossland, H. Patel, E. Wong, K. Kong, H. Gunasekera, B. Richards, L. Caffery, C. Perry, R. Sacks *et al.*, "An Artificial Intelligence Computer-vision Algorithm to Triage Otoscopic Images From Australian Aboriginal and Torres Strait Islander Children," *Otology & Neurotology*, vol. 43, no. 4, pp. 481–488, 2022. [Online]. Available: https://www.ingentaconnect.com/content/wk/mao/2022/00000043/00000004/art00021

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] "Papers with code - imagenet benchmark (image classification)." [Online]. Available: https://paperswithcode.com/sota/image-classification-on-imagenet

[12] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine*, vol. 8, no. 11, 2020.

[13] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in Medical Imaging: A Survey," *arXiv preprint arXiv:2201.09873*, 2022. [Online]. Available: https://arxiv.org/pdf/2201.09873.pdf

[14] E. Başaran, A. Şengür, Z. Cömert, Ü. Budak, Y. Çelık, and S. Velappan, "Normal and acute tympanic membrane diagnosis based on gray level co-occurrence matrix and artificial neural networks," in *2019 international artificial intelligence and data processing symposium (IDAP)*. Ieee, 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8875973

[15] M. Viscaino, J. C. Maass, P. H. Delano, M. Torrente, C. Stott, and F. Auat Cheein, "Computer-aided diagnosis of external and middle ear conditions: A machine learning approach," *Plos one*, vol. 15, no. 3, p. e0229226, 2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229226

[16] S. Camalan, M. K. K. Niazi, A. C. Moberly, T. Teknos, G. Essig, C. Elmaraghy, N. Taj-Schaal, and M. N. Gurcan, "OtoMatch: Content-based eardrum image retrieval using deep learning," *Plos one*, vol. 15, no. 5, p. e0232776, 2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232776

[17] A. Mothershaw, A. C. Smith, C. F. Perry, C. Brown, and L. J. Caffery, "Does artificial intelligence have a role in telehealth screening of ear disease in Indigenous children in Australia?" *Australian Journal of Otolaryngology*, vol. 4, 2021. [Online]. Available: https://www.theajo.com/article/view/4451/html

DCAN:DenseNet with Channel Attention Network for Super-resolution of Wireless Capsule Endoscopy

Hiren Vaghela[1], Anjali Sarvaiya[1], Pranav Premlani[1], Abhishek Agarwal[1], Kishor Upla[1], Kiran Raja[2] and Marius Pedersen[2]
[1]*Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India.*
[2]*Norwegian University of Science and Technology (NTNU), Gjøvik, Norway.*
(hvaghela429, anjali.sarvaiya.as, pranavpremlani2002, abhishekag2702, kishorupla)@gmail.com
(kiran.raja, marius.pedersen)@ntnu.no

***Abstract*—Wireless Capsule Endoscopy (WCE) captures images of the gastrointestinal (GI) tract and transmits the images in a wireless manner. Due to the hardware limitations of the capsule and the varying imaging conditions within the GI tract, the recorded images can have a low spatial resolution with a high frame rate or a high spatial resolution with a low frame rate. Although it is generally common to have low spatial resolution to capture details of the GI tract, low spatial resolution limits the detection of minor anatomical features and abnormalities in the small intestine and other portions of the GI tract. Super-Resolution (SR) is a class of software-based techniques that are used to enhance the resolution of a Low-Resolution (LR) image. This work proposes a new model referred as *DCAN*-DenseNet with Channel Attention Network for Super-resolution of LR WCE images. The design of *DCAN* consists of multiple strategies adopted from state-of-the-art methods such as Channel Attention Network (CAN) from RCAN and short dense connections from DenseNet to extract details from LR observation. Additionally, to improve the accuracy of the SR images, we create a derivative dataset of 10,000 images from a publicly available WCE dataset. The proposed approach has been validated against multiple state-of-the-art methods by conducting quantitative evaluation using perceptual metrics. The analysis is complemented by statistical validation to demonstrate the consistency of the proposed method over the other models for the SR task.**

***Index Terms*—Wireless Capsule Endoscopy, Super Resolution, Channel Attention Network, DCAN**

I. Introduction

The Wireless Capsule Endoscopy (WCE) is a minimally invasive medical technology that utilizes a small, swallowable capsule equipped with a wireless camera to capture images and videos of GastroIntestinal (GI) tract. Captured video frames are transmitted to a recording device outside the patient's body. It allows for a comprehensive examination of the small intestine similar to conventional endoscopy but with additional convenience. The recorded images and videos provide valuable diagnostic information about GI disorders such as Crohn's disease, tumors, bleeding, or Inflammatory Bowel Disease (IBD) [1]. It generates an average of 50k to 60k images while moving through the GI tract, and a normal colon video test generates about 8 hours of RGB video data. Thus, the vast amount of data generated by WCE presents a challenge for medical professionals, who must verify numerous images or videos. Continued advancements in technology and image analysis algorithms further enhance the capabilities of WCE, leading to improved patient care and outcomes.

Resolution plays a crucial role in all vision-driven applications including medical diagnosis. A low resolution video/image can lead to incorrect diagnostics for both machines and medical practitioners [2]. The image sensor equipped with High-Resolution (HR) can help visualise intricate details within the digestive tract, such as mucosal irregularities, ulcers, polyps, or early-stage tumors [3]. The clear and detailed images obtained from HR camera allowing doctors for targeted interventions or surgical procedures. However, a capsule consisting of an optical dome, illuminator, imaging sensor, battery, and RF transmitter in a capsule-shaped structure with a length of 26 mm and a diameter of 11 mm [1] can work in two modes. The small-sized structure leads to hardware limitations in terms of spatial resolution of sensor which is usually coarser. The minimum resolution obtained by a capsule used is 336×336 pixels with 24 frames per second (fps) [4] and the maximum resolution of 1 megapixel can reduce the frames rate to 5 fps. Having higher fps is advantageous in covering large area and despite of having numerous benefits of WCE technology, the operational fps suffers from inadequate frame resolution and video quality leading to adverse diagnostics [5]. Thus, there is a clear demand for methods capable of enhancing the resolution of capsule endoscopes to facilitate both subjective and objective analysis.

Image Super-Resolution (SR) is a software-driven method used to enhance the LR image to its corresponding HR one. Single Image SR (SISR) and Multi-Image Super-Resolution (MISR) are the two types of SR methods, with SISR being more popular due to its advantages over MISR, where multiple images of the same scene and image registration are required. However, SISR poses a challenging ill-posed problem as a single LR image may correlate to several HR solutions [6]. The recent advancement of deep learning techniques has resulted in a number of techniques that can be used in SISR making it possible to use for other applications.

Inspired by the success of applications in other domains, we present a SR approach for WCE images using a deep learning-based approach which we refer to as *DCAN*-DenseNet with Channel Attention Network (CAN). The proposed architecture

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

incorporates the CAN mechanism for extracting high-level details by feature scaling in an adaptive way. Such a mechanism allows us to leverage the high frequency details in WCE image to identify and retain abnormality present in the WCE images for downstream classification tasks like pathology classification. Additionally, we also introduce short skip connections to extract low-level features that are common in images from the GI tract. The low level features are then combined with the high-level features to generate information-rich SR images. To enhance the reconstruction process and recover image details, we also employ bottleneck, deconvolution, and reconstruction layers. The potential of the proposed model is evaluated on a new derived dataset created from the original Kvasir capsule endoscopy dataset [4]. Our contributions from this work are:

- A new SR approach that leverages Channel Attention Network (CAN) and Dense connections to generate SR images from LR images.
- The CAN in the proposed model adaptively re-scales the features by taking into account the inter-dependencies among different features. Further, the use of short skip connections in convolution layers specializes in excessing the high-level features to obtain low-level features in the input LR image, which is important for better detailing in reconstruction of the SR image.
- Unlike other approaches, we propose training and testing of WCE images in the Y-channel of $YCbCr$ which provides better performance metrics compared to RGB color scheme and corresponds closely to the Human Visual System (HVS). This is validated through empirically through various metrics such as Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Metric (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) which asserts our intuition of using $YCbCr$ over RGB processing.
- Further, due to unavailability of datasets for SR tasks, we create a new derivative dataset from Kvasir Capsule Endoscopy [4] to train the SR network. The new dataset consists of 10,000 samples which are manually pre-processed to improve the accuracy of the proposed network. All our experiments are conducted using state-of-the-art methods to demonstrate the applicability of proposed approach for SR generation and is supported by detailed analysis of various perceptual metrics.

II. Literature Review

SR methods based on deep learning aim to capture the complex relationship between given LR and HR images. Dong et al. [7] introduced Super-resolution Convolutional Neural Network (SRCNN) consisting shallow network of having 3 layers. Later, Kim et al. increased the network depth to 20-layers proposed VDSR [8] and DRCN [9], which achieved notable improvements over the previous SRCNN indicating the importance of network depth. In similar lines, Lim et al. [10] further advanced this concept by creating the EDSR, a very wide network an exceptionally deep network consisting of approximately 165 layers, using simplified residual blocks. However, it is worth noting that merely stacking residual blocks to construct deeper networks does not necessarily lead to significant improvements.

Tong et al. introduced DenseNet [11] leveraging dense connections between convolution layers and growth rate to quantify the amount of new information added by each layer to the final reconstruction. Also due to dense connections all level of high, average and low level of features can be extracted easily. Thus, the DenseNet model utilizes feature maps from each layer that are merged with the previous layer, and the data is replicated multiple times for effective training of very deep networks. Zhang at el. proposed RCAN [12] model with Residual in Residual (RIR) structure where the Residual Group (RG) acts as the basic module and allows for residual learning in a coarse level through the use of Long Skip Connections (LSCs). This model also introduces a Channel Attention (CA) mechanism, which adaptively re-scales each channel-wise feature by modeling the inter-dependencies across feature channels. This mechanism enables the network to focus on more useful channels, thereby enhancing its discriminative learning ability. Several other SR works have been proposed to enhance the perceptual quality of SR results. For instance, Ledig et al. [13] proposed an SRGAN model that improves the perceptual quality of super-resolved images beyond pixel-level improvements. Similarly, Wang et al. [14] proposed an Enhanced Super Resolution using GAN (ESRGAN), which introduces several improvements over SRGAN. These works have been tested on visible (i.e., RGB scene) images and are also extended to medical data. Mahapatra et al. in [15] used Progressive GAN (P-GAN) for accurate detection and proper segmentation of anatomical landmarks on MRI images. Additionally, a few other SR techniques have also been utilized to improve the quality of images acquired by traditional endoscopic cameras. Yasin et al. [16] learned a mapping from low-to-high resolution mapping using conditional adversarial networks with a spatial attention block to improve the resolution by up to factors of $\times 8$, $\times 10$, $\times 12$ respectively. However, the approach is limited to conventional endoscopy images. Thus, the super-resolution of WCE images is not attempted by researchers in the community to the best of our knowledge motivating us to focus on SR task for WCE images.

III. Proposed method: DenseNet with Channel Attention Network (DCAN)

With the aim of recovering rich high-frequency details from capsule endoscopy images, the proposed approach consists design inspired from RCAN [12] and DenseNet [11]. The RCAN model is one of the state-of-the-art methods for SR of visible images which has introduced novel Channel Attention Network (CAN) to improve the learning ability of CNN network. Similarly, dense connections are usually employed in the CNN network to learn effective features from LR images and also to reduce the effect of overfitting or underfitting. Motivated by these, we incorporated the above concepts in the proposed method which we referred as *DCAN*-DenseNet with Channel Attention Network. In WCE images, low-frequency

components display a relatively homogeneous pattern, the high-frequency elements typically correspond to regions characterized by edges, texture, and other intricate details. Thus, the use of CAN in the proposed model enhances channel-wise feature representations, and hence, *DCAN* gains the advantage to extract information more precisely. The fusion of this with dense connections results in a powerful architecture that leverages the strengths of both RCAN and DenseNet, enabling it to effectively handle the task at hand and achieve superior performance in acquiring intricate details within WCE data.

The architecture of the proposed model for the task of SR of WCE images for upscaling factors $\times 4$ is depicted in Fig. 1. It can be observed that the WCE LR image is given to a convolution layer first to learn low-level features. After this, a single CAN layer is added to learn the features channel wise from the LR image. Subsequently, a series of DCAN blocks are employed to learn high-level features. Towards the end, the bottleneck layer is used to decrease the input feature maps and finally, the deconvolution layer is employed to upsample the feature images, and the output of the reconstruction layer generates an SR image.

A. DCAN Block

In the proposed network, we utilize DCAN blocks as fundamental building unit. This design allows to enhance details and promotes feature reuse throughout the network, leading to more comprehensive and expressive representations at higher layers. The network architecture of DCAN block is depicted in Fig. 2. There are n number of DCAN blocks used in our architecture, which we fix to 8 empirically. Each DCAN block consists of Channel Attention Network (CAN), one convolution layer and m number of DCAN layers (i.e., $m = 8$) that enable to extract high-level features in the output image. Moreover, one skip connection is added to avoid vanishing-gradient problem. The block schematic DCAN layer is shown in Fig. 3 (a). Each DCAN layer consists of a convolution layer having kernel size 3×3 and Relu activation function with short skip connection. Thus, the proposed model consists of short skip connections and also global skip connections for effective learning and also to avoid gradient problem.

B. Channel Attention Network (CAN)

The earlier CNN-based SR methods [7]–[11], [17], treat LR channel-wise features equally, which is not optimal for real-world cases. To address this issue and focus the network on more informative features, CAN mechanism is proposed in RCAN [12] that exploits the interdependencies among feature channels. Generating different attention for each channel-wise feature is a crucial step in this mechanism. An LR information contains both low-frequency and high-frequency components that are valuable for SR. However, the low-frequency parts are relatively homogeneous, while the high-frequency components typically correspond to regions with edges, texture, and other details. Second, each filter in the convolution layer operates within a local receptive field, which limits its ability to exploit contextual information beyond the local region. Thus, the use of CA in the proposed method is helpful to learn the features effectively by assigning proper weights to each feature. The architecture for CAN is depicted in Fig. 4. It consists of adaptive average pooling with a convolution layer with kernel size 3×3, attached with a ReLU activation function, which is passed to another convolution layer with kernel 1×1, and a skip connection is used to add the input values with the output of the sigmoid function.

C. BottleNeck Layer

In order to enhance the compactness and computational efficiency of the model, we use a bottleneck layer to decrease the quantity of feature maps prior to their input into the deconvolution layers. The bottleneck layer shown in Fig. 3(b) is used to reduce the output features from DCAN blocks to a lower dimension. It consists of a convolution layer having kernel size 1×1 with ReLU activation function. In our proposed model, we reduce the features to 256 features using the bottleneck layer.

D. Deconvolution and Reconstruction Layers

Deconvolution layers can be seen as the inverse operation of convolution layers, allowing for the learning of diverse upscaling kernels that work together to predict HR images. It provides two advantages: By conducting computations in the LR space, the SR reconstruction process is accelerated. Additionally, the inclusion of a deconvolution layer enables the utilization of contextual information from LR images to infer high-frequency details. The network design of the Deconvolution layer consists of two pixel-shuffle layers as shown in Fig. 3(c), where each layer upsamples the image by a factor of $\times 2$. Pixel-shuffle rearranges the feature maps by reshaping them into a higher resolution. Then it rearranges the pixel values to get the final image. We are using two pixel-shuffle layers which give us total upsampling of factor 4. Finally, a reconstruction layer, consisting of a convolution layer with a 3×3 kernel, is used to generate SR images from the feature maps in the RGB space.

IV. Experimental Analysis

The design of the proposed model is validated by conducting subjective and quantitative evaluations. We empirically verify DCAN's effectiveness with state-of-the art architectures qualitatively by taking a patch from the output images from all state-of-the-art models. Additionally, the same is quantitatively verified using different standard SR metrics such as Peak Signal to Noise Ratio (PSNR) & Structural Similarity Index Metric (SSIM) and using perceptual metric i.e., Learned Perceptual Image Patch Similarity (LPIPS). Finally, the statistical analysis of the proposed model is also carried out to show the consistency of the SR results. Our method is benchmarked against state-of-the-art models such as SRGAN [13], CycleGAN [18], DenseNet [11], and RCAN [12] for comparison purposes.

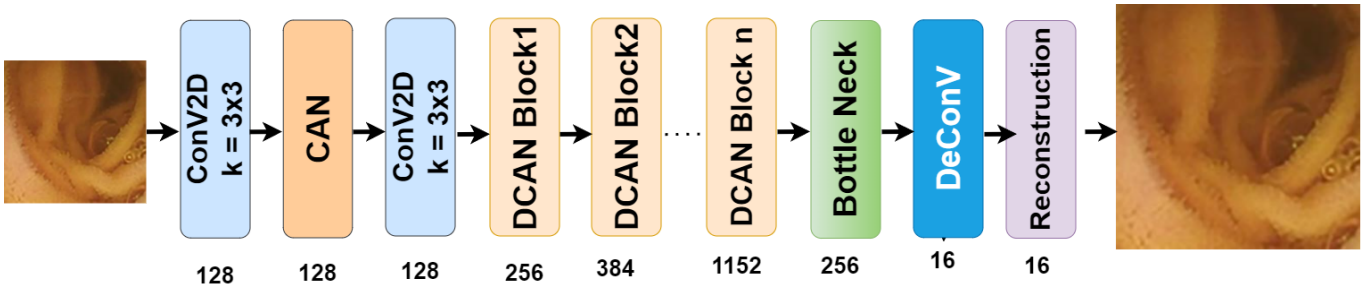


Fig. 1. The network architecture of proposed model *DCAN*, where k denotes kernel size and numerical values mentioned below every layer indicates size of output features.

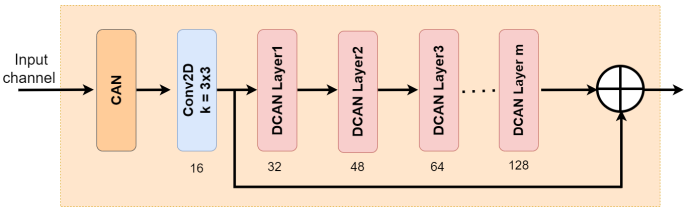


Fig. 2. The network architecture of DCAN block used in proposed model-*DCAN*. The values below every layer indicate the number of output features.

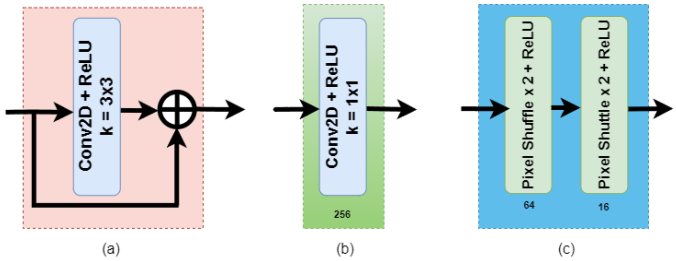


Fig. 3. The design of (a) DCAN Layer, (b) Bottleneck Layer and (c) Deconvolution Layer in the proposed method.

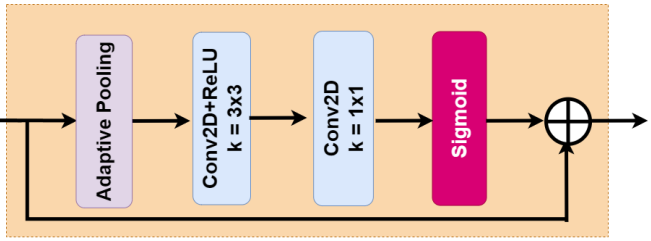


Fig. 4. The architecture design of Channel Attention Network (CAN) used in *DCAN* model.

A. Dataset

One of the novel contributions of the work is the creation of new derivative dataset from the available Kvasir Capsule Endoscopy Dataset [4] consisting of WCE images. In the original Kvasir dataset, each image is in the RGB color space with size of 336×336 pixels. The dataset contains a total of $47,236$ images, which are categorized according to different medical anomalies. As the original dataset contained redundant images with many border areas with black pixels, we have curated the dataset for the SR task by manually selecting images and removing redundant images from the Kavasir dataset. The new SR dataset therefore consists of $10,000$ training images, 550 validation images and 1000 testing images [1]. As mentioned earlier, the WCE images from the original Kvasir dataset containing non-informative part in the border area. Those regions are removed manually through cropping resulting in images of 280×280 pixels for all images. The proposed model along with all other models are experimented on the new datset and SR results are generated.

B. Training details

Firstly, to prepare LR-HR pair of WCE images, we consider the original images as the HR image and applied bicubic down-sampling with factor $\times 4$ and obtained LR image. These LR-HR pairs are fed to the proposed model to train it to generate SR images. Further, each LR image has also been transformed into $YCbCr$ space and only the Y-channel was used for training which represents a gamma-encoded channel that predominantly contains high-level feature information. On the other hand, the Cb and Cr channels are chroma-encoded channels that do not contain as many high-level features. To save computational time during the process and improve the extraction of high-level features in the SR image, the Cb and Cr channels are directly interpolated and added to the output image. The training process aimed to minimize the loss function, which was taken as the Mean Squared Loss (MSE). The training was carried out for a total of 300 epochs with a batch size of 32. Additionally, the Adam optimizer was used with a learning rate of 0.0001. This protocol was used on all the state-of-the-art models and SR results are generated. While testing, we use the YCbCr space of test LR image to generate the SR image.

C. Comparison with state-of-the-art models

Qualitative Analysis: The qualitative comparison of various SR methods on scaling factor of $\times 4$ is depicted in Fig. 5. One can inspect by looking at the zoomed-in patches that the proposed model generates better SR solutions than other models. Also, the SSIM map of each patch is shown below the

[1]The dataset will be made available for researchers provide they have license agreement for original Kvasir Capsule dataset.

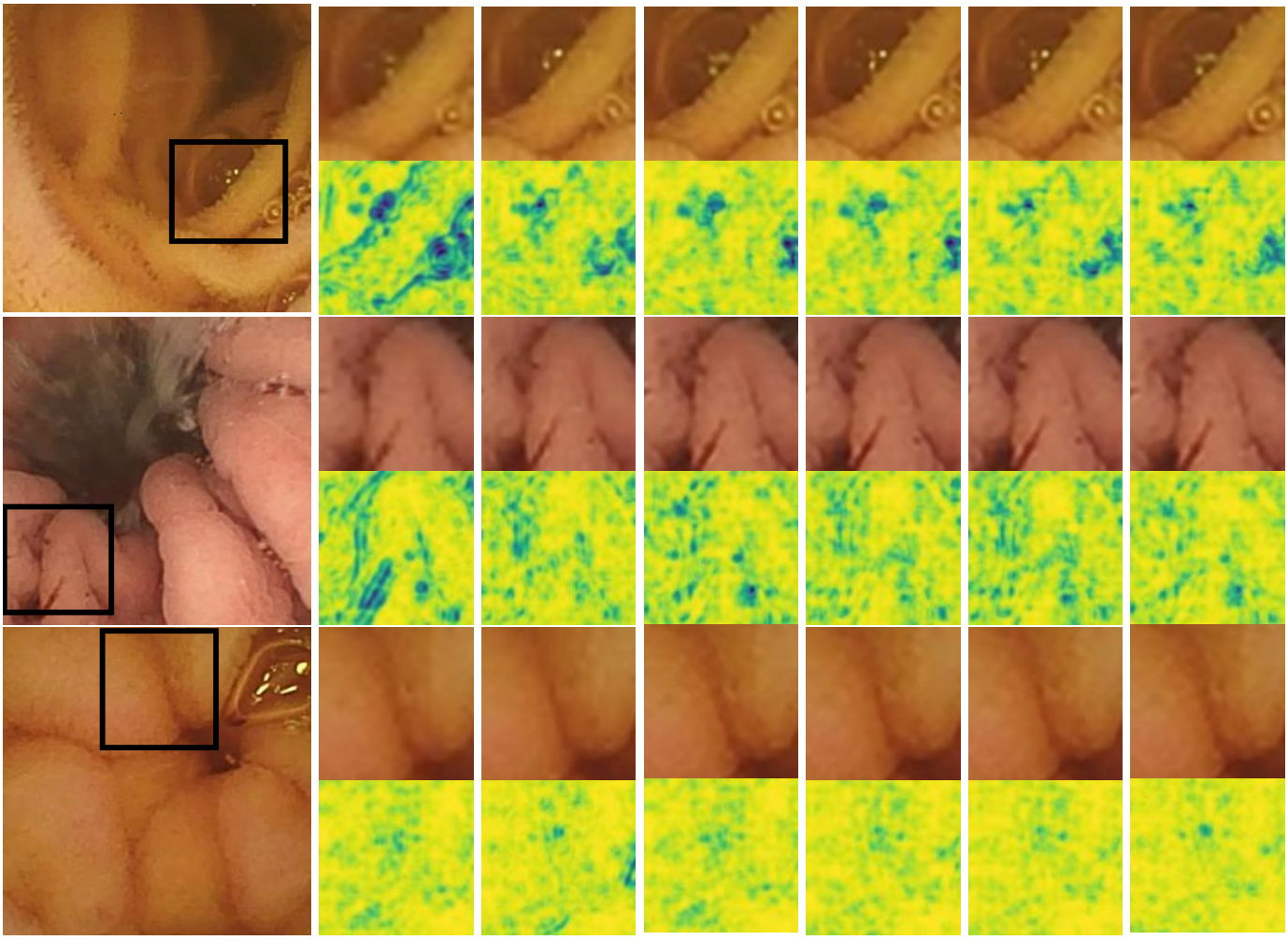


Fig. 5. Qualitative comparision of proposed models with state-of-the-art models using SSIM maps (where yellow region shows similarity and blue region shows disimilarity.)

patch SR image, which shows the similarity between generated SR and HR images. The yellow part in the SSIM maps shows the similarity between SR and HR images. However, the green and blue regions in SSIM maps show the dissimilarity between SR and HR images. We can observe that the proposed model has more similar parts than the other methods. In the first row SSIM maps in Fig. 5, it can be observed that the bicubic output image exhibits the highest dissimilarity. Comparatively, other models such as RCAN and DenseNet perform better than other models as well as the bicubic method. However, the proposed model demonstrates the lowest dissimilarity among the bicubic method and all other state-of-the-art models. Additionally, in the second row, it is apparent that SRGAN and DenseNet exhibit better performance in comparison. However, the proposed model demonstrates the highest structural similarity, indicating superior performance in terms of preserving the structural characteristics of the image.

Quantitative Analysis: To validate the SR results quantitatively, the average SSIM and PSNR values for the testing images of each model are provided in Table I. We calculated the average PSNR and SSIM on Y channel as well as the RGB channels. From the table, it can be observed that the proposed DCAN model has the highest PSNR in both the Y channel and the RGB channel. Additionally, the DCAN model also demonstrates the highest SSIM values, implying a better structural similarity. When considering LPIPS for perceptual image comparison, lower LPIPS values emphasize the model's ability to capture perceptual similarity effectively. Remarkably, the DCAN model exhibits the lowest LPIPS values among all different models, suggesting superior perceptual similarity.

Statistical Analysis: Finally, the statistical analysis is also conducted on the SR results of the proposed model along with the others to ensure the model consistency compared to state-of-the-art models. Standard deviation demonstrates the deviation in the values from the mean and hence it should be low for an algorithm. The values of standard deviation of each model are presented in Table II. As we can observe the values of our proposed model DCAN is lowest in comparision to all state-of-the-art models. From the given values, it can be noticed that while comparing PSNR consistency Y channel has

TABLE I
QUANTITATIVE COMPARISON OF THE PROPOSED MODEL OVER OTHER MODELS USING DIFFERENT METRICS SUCH AS PSNR, SSIM AND LPIPS ON RGB AND Y-CHANNELS.

Model	PSNR ↑		SSIM ↑		LPIPS↓
	Y-channel	RGB	Y-channel	RGB	RGB
Bicubic	38.1069	37.2111	0.9296	0.9057	0.2310
SRGAN [13]	38.0377	37.0021	0.9291	0.9049	0.1972
CycleGAN [18]	38.0121	36.9441	0.9123	0.9012	0.1984
DenseNet [11]	39.6842	38.8596	0.9401	0.9369	0.1353
RCAN [12]	40.1438	39.4613	0.9427	0.9371	0.1359
Proposed	**40.2261**	**39.5389**	**0.9486**	**0.9378**	**0.1346**

TABLE II
THE STATISTICAL COMPARISON OF THE PROPOSED MODEL WITH OTHER DIFFERENT METHODS USING PARAMETER OF STANDARD DEVIATION OF PSNR AND SSIM VALUES OVER MEAN VALUES IN RGB AND Y-CHANNELS.

Model	STD. dev. of PSNR ↓		STD. dev. of SSIM ↓	
	Y-channel	RGB	Y-channel	RGB
Bicubic	3.8943	2.2102	0.0353	0.0347
SRGAN [13]	3.3490	2.6924	0.0348	0.0324
CycleGAN [18]	3.6802	2.1937	0.0356	0.0319
DenseNet [11]	4.2025	3.1996	0.0378	0.0372
RCAN [12]	3.5612	2.8753	0.0367	0.0350
Proposed	**3.0006**	**2.0186**	**0.0270**	**0.0306**

lower consistency than the RGB channels and while comparing SSIM, its vice-a-versa. Thus, one can conclude from this that there is high peaks in Y-channel i.e., Y-channel works great on majority images and provides better PSNR values, but as it focuses only on Y-channel, the features in Cb and Cr channels are not percieved properly, so when the high-level features are in the Cb and Cr channels, it loses important information which is although a rare case as all important information is in Y-channel majorly.

V. CONCLUSION

Due to the hardware limitations of the WCE sensors, the captured data results in coarser resolution which affects the diagnosis accuracy of the diseases. We present a new SR approach *DCAN* using dense connections and channel attention modules to convert LR images to SR images. As the proposed network integrates the advantages of Channel Attention Network (CAN) from RCAN and utilizes short dense connections inspired by DenseNet, the proposed approach is able to effectively extract details from LR observations. Experiments show that the proposed network can perform better than other existing state-of-the-art SR models both quantitatively and qualitatively. The results are supported with a detailed analysis of quality assessment metrics such as PSNR, SSIM, and LPIPS and statistical analysis of obatained results. A future direction in this work is to focus on improving the perceptual quality and assess it with medical practitioners.

ACKNOWLEDGMENT

Authors are thankful to Research Council of Norway (RCN) for International Network for Capsule Imaging in Endoscopy (CapsNetwork) project managed by Department of computer science, Norwegian University of Science and technology (NTNU), Norway for providing support for this research work.

REFERENCES

[1] P. Muruganantham and S. Balakrishnan, "A survey on deep learning models for wireless capsule endoscopy image analysis," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 83–92, 06 2021.

[2] S. B. and A. P., "Recent developments in wireless capsule endoscopy imaging: Compression and summarization techniques," *Computers in Biology and Medicine*, vol. 149, p. 106087, 2022.

[3] O. Gilja, J. Hatlebakk, S. Ødegaard, A. Berstad, I. Viola, C. Giertsen, T. Hausken, and H. Gregersen, "Advanced imaging and visualization in gastrointestinal disorders," *World journal of gastroenterology : WJG*, vol. 13, pp. 1408–21, 04 2007.

[4] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021.

[5] C. F. Sabottke and B. M. Spieler, "The effect of image resolution on deep learning in radiography," *Radiology. Artificial intelligence*, vol. 2, p. e190015, January 2020.

[6] H. Chen, X. He, C. Ren, L. Qing, and Q. Teng, "Cisrdcnn: Super-resolution of compressed images using deep convolutional neural networks," *Neurocomputing*, vol. 285, 09 2017.

[7] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 391–407, Springer, 2016.

[8] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, 2016.

[9] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645, 2016.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

[11] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE international conference on computer vision*, pp. 4799–4807, 2017.

[12] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018.

[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

[14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[15] D. Mahapatra, B. Bozorgtabar, and R. Garnavi, "Image super-resolution using progressive generative adversarial networks for medical image analysis," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 30–39, 2019.

[16] Y. Almalioglu, K. Bengisu Ozyoruk, A. Gokce, K. Incetan, G. Irem Gokceler, M. Ali Simsek, K. Ararat, R. J. Chen, N. J. Durr, F. Mahmood, and M. Turan, "Endol2h: Deep super-resolution for capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4297–4309, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

Enhancement of Colour Reproduction for Capsule Endoscopy Images

Léo Watine
Department of Biomedical Engineering
Université de Strasbourg
Strasbourg, France
leo.watine@gmail.com

Pål Anders Floor*, Marius Pedersen**, Peter Nussbaum, Bilal Ahmad
Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
*paal.anders.floor@ntnu.no, **marius.pedersen@ntnu.no

Øistein Hovde
Department of Gastroenterology
Innlandet Hospital Trust
Gjøvik, Norway

***Abstract*—During a clinical examination, a capsule endoscope, like the PillCam, allows the entire digestive system of the patient to be photographed. The images can be analysed to detect abnormalities. However, the rendering is of low quality compared to colonoscopy. Distorted colour reproduction of tissues and blood vessels can be one reason for overseeing abnormalities, and improvement of colour reproduction could improve the examination. For this purpose, a colourimetric calibration must be performed in order to deduce the correction to be applied. The ColorChecker is a chart containing a number of standard colour patches sampled over the gamut of perceivable colours, providing ground truth reference. Thus, it is possible to measure the rendering of these colours and then compare and adjust its values. However, the colours that appear inside the gastrointestinal system are a subset of perceivable colours. It also appears like typical PillCam's are constructed accordingly. As shown in this paper, colour correction based on the ColorChecker becomes saturated. Therefore, we develop a novel colour checker based on colours sampled from colonoscopy images and perform corrections based on that. An evaluation of both inter- and intra-camera variation indicates that the same calibration method should perform quite well for all PillCam Colon2 capsules. A subjective evaluation of colour corrected videos was performed by two experienced gastroenterologist, indicating that the proposed method obtains a more faithful colour reconstruction.**

***Index Terms*—Colour correction, capsule endoscopy video**

I. Introduction

Serious diseases of the digestive system, such as Crohn's disease, inflammatory bowel disease, and cancer, affect many people. In Europe, colorectal cancer is the second most common cause of cancer deaths[1].The survival rate is closely related to early detection of cancer.

Lately, the primary goal of preventing death from diseases has moved towards prevention of development of disease through screening [1]. However, fear of pain and the taboo associated with the colonoscopy, is a major drawback in making people volunteer for such screening [2]. The Wireless Capsule Endoscope (WCE) [3], which is a pill-sized capsule that the patient swallows, is a good alternative for screening, as it avoids the problems mentioned above. The WCE carries cameras that record video of the gastrointestinal (GI) tract. However, WCE video streams are long and time-consuming to evaluate, and current WCE has lower resolution, lower frame rate, and lower quality rendering than colonoscopy, making it more difficult to detect pathologies of the GI wall.

Generally, it is of great importance that colours are reconstructed properly, as the colour of tissues and blood vessels can help to detect abnormalities. What *properly* means depends on context. For automatic pathology detection *colour accuracy* is important, whilst for gastroenterologist it is *colour consistency* that matters [4]. Colour accuracy refers to the ability of a system to produce exact colour matches from input to output. Colour consistency refers to the ability to produce image data with a similar response in a human interpreter.

To deal with the problem of colour distortion, in this paper we aim at improving colour reproduction in PillCam Colon2[2] videos through post-processing. For this purpose, a colorimetric calibration must be performed in order to deduce the relevant correction to be applied. The *ColorChecker* [5], is a chart containing a number of standard colour patches sampled over the gamut of perceivable colours, providing ground truth reference. Thus, it is possible to measure the rendering of these colours and then compare and adjust the values accordingly. However, the colours appearing inside the GI system are a subset (or sub-gamut) of perceptible colours. Also, it appears like many WCE's, like PillCam Colon2, are constructed accordingly, emphasising colours typically present in the human GI-system [6]. Therefore, colour correction based on the ColorChecker leads to saturated colour reconstruction. Therefore, we develop a novel colour checker for the task, named *ColonColorChecker* (CCC), based on colours

This work was supported by the Research Council of Norway (RCN), under the project CAPSULE no.300031

[1]http://www.euro.who.int/en/health-topics/noncommunicable-diseases/cancer/news/news/2012/2/early-detection-of-common-cancers/colorectal-cancer (1/5-23)

[2]https://www.medtronic.com/covidien/en-us/products/capsule-endoscopy/pillcam-colon-2-system.html (1/5-23)

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

sampled from colonoscopy images, and perform corrections based on this. We also analyse variability in colour processing both within one Pillcam, as well as inter-variability across several different PillCam's. Two experienced gastroenterologists assessed the results of colour correction subjectively.

In Section II preliminaries are provided. In Section III our colour correction methods are described. In Section IV results are presented. Section V, summarises and concludes the paper.

II. Preliminaries

A. The GretagMacbeth ColorChecker

For several experiments we will use the *GretagMacbeth ColorChecker* [5], a colour calibration device consisting of a cardboard arrangement of 24 coloured sample squares. These patches have spectral reflectances intended to mimic those of additive and subtractive primaries as well as natural objects, such as human skin and foliage, and a grey scale of six different values ranging from white to black. The patches have consistent colour appearance under a variety of lighting conditions and are stable over time.

B. Data sets

We have ten videos from different WCE's of type PillCam Colon2 available, obtained from ten different patients. These videos were collected during clinical trials by PhD candidates Anuja Vats and Bilal Ahmad under consultation of Prof. and Gastroenterologist Øistein Hovde, at Innlandet Hospital Trust, Gjøvik in 2021. For all PillCams, captures of objects that could later be used for geometric, colourimetric, and radiometric calibration, among them the ColorChecker, was taken prior to the patients ingesting them. The ColorChecker was placed 4cm away from the tip of the PillCam inside a black box with diffuse absorbing material, reflecting no light. With these videos it is possible to analyse each camera separately as well as inter-camera variation.

C. $L^*a^*b^*$ space, colour difference and correction

The *CIELAB colour space* is a *perceptually uniform* colour space in three dimensions with orthonormal basis L^*, a^*, b^*, established in 1976 by the International Commission on Illumination (CIE) [7, pp.94-95]. The L^*-axis corresponds to brightness or luminosity according to a psychometric scale ranging from 0 to 100, where 100 represents white, or total reflection, and 0 represents black, or total absorption. Two orthonormal axes, a^* and b^*, are chromaticity coordinates determining the hue and saturation of a given colour, where a^*-axis represents colours from red to green ($-a^*$) and the b^*-axis represents colours from yellow to blue ($-b^*$).

The colour difference, ΔE, measures the difference between two colours (points) in CIELAB colour space [7, pp.95]

$$\Delta E = \sqrt{(L_1 - L_2)^2 + (a_2 - a_1)^2 + (b_2 - b_1)^2}, \quad (1)$$

with L_1, a_1, b_1 the coordinates of the first color and with L_2, a_2, b_2 the coordinates of the second. ΔE is on a scale from 0 to 100, where 0 means imperceptible difference and 100 means total distortion. For example: $\Delta E \leq 1.0$ is not perceptible by the human eye, $\Delta E < 2$ is perceptible by a professional, $\Delta E \in [11, 49]$ refers to similar colours, and $\Delta E = 100$ means that the colours are opposite.

To correct for colour distortion one applies a *Colour Correction Matrix* (CCM) which transforms and input RGB-, XYZ- or $L^*a^*b^*$ vector, typically a known colour target captured by a camera, to an output (corrected) RGB-, XYZ- or $L^*a^*b^*$ vector. The CCM is a 3×3 (or 4×3) matrix whose entries are determined to satisfy the relationship between a known target, like the ColourChecker, and the output of particular camera, like the PillCam, so that the average brightness remains constant. If captures are taken with the relevant camera of a known target, like the ColourChecker, one can estimate the CCM that minimizes ΔE, then use the resulting CCM to correct colours in subsequent images. Estimating the CCM given a reference input is a standard procedure implemented in many computation tools, like MatLab[3].

Typically one can map from some measured RGB values, $[R, G, B]^T$, to the CIE *tristimulus values*, $[X, Y, Z]^T$, via a 3×3 matrix [7, p.64-67], then to $L^*a^*b^*$ as [7, p.94]

$$\begin{aligned} L^* &= 116 f(Y/Y_w) - 16, \ a^* = 500\big(f(X/X_w) - f(Y/Y_w)\big) \\ b^* &= 200\big(f(Y/Y_w) - f(Z/Z_w)\big), \end{aligned} \quad (2)$$

where X_w, Y_w, Z_w are the XYZ values for the reference white being used, and where

$$f(u) = \begin{cases} u^{1/3}, & u > (24/116)^3 \\ (841/108)u + 16/116, & u \leq (24/116)^3. \end{cases} \quad (3)$$

One can measure the XYZ values from some object directly using a spectroradiometer (See Section III): Take a coloured object that reflects light by some *spectral reflection function*, $R(\lambda)$, from some light source with spectrum $I(\lambda)$, where λ denotes wavelength. The corresponding XYZ values are then

$$\begin{aligned} X &= \frac{K}{N}\int_\lambda R(\lambda)I(\lambda)\bar{x}(\lambda)\mathrm{d}\lambda, \ Y = \frac{K}{N}\int_\lambda R(\lambda)I(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda, \\ Z &= \frac{K}{N}\int_\lambda R(\lambda)I(\lambda)\bar{z}(\lambda)\mathrm{d}\lambda, \ \text{with } N = \frac{K}{N}\int_\lambda I(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda, \end{aligned} \quad (4)$$

where $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ are the *CIE 1931 colour matching functions* [7, p. 68], and K is some scaling factor.

III. Colour Correction Methods

A. Correction with ColorChecker

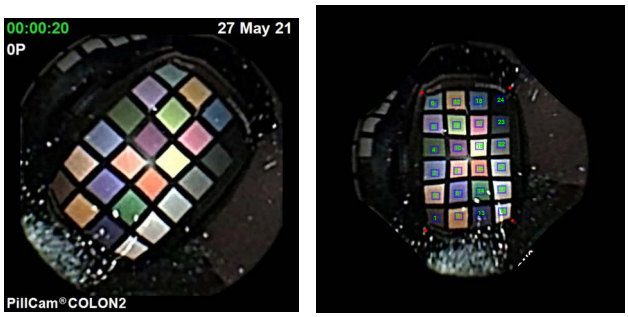
The ColorChecker is present in some captures at the beginning of the available PillCam videos. Each image includes additional information surrounding the actual captures as seen in Fig. 1(a). The frames are cropped before analysis.

We will use standard algorithms for recognition and analysis of the ColorChecker in MatLab[4]. However, due to the recording conditions and the size of the ColorChecker, parts

[3]https://uk.mathworks.com/help/images/correct-colors-using-color-correction-matrix.html (1/5-23)

[4]https://www.mathworks.com/help/images/ref/colorchecker.html (1/5-2023)

of it can't be distinguished in the image. Additionally, there is a strong lens distortion at the edges of the camera. This can lead to erroneous detection. Also, it is necessary that the ColorChecker is oriented vertically (or horizontally) to use the dedicated algorithm. Therefore, we rotate the frames using the *Hough transform* [8] prior to analysis.



(a) (b)

Fig. 1. (a) Example of ColorChecker capture as it appears in one of the trials. (b) Correct detection of ColorChecker.

One needs to check if the detection of the grid is correct before the image can be applied for analysis. An example of a correct detection is shown in Fig. 1(b). The reference points, or *landmarks*, marked in red are well placed, correctly identifying the corners of the ColorChecker, leading to correct identification of the grid. The order of the patches and the size of the *regions of interest* (ROIs) are shown as blue numbers and rectangles, respectively. Each ROI consists of 15×20 pixels, which we average to determine the measured colour value. The whole patch is not covered as it is essential not to include pixels outside the patches as the RGB values would then be biased by the presence of many black pixels. Regions with specularities and other distortions should be avoided.

We then have 24 numbered ROIs, each containing 300 pixels, where the average RGB intensity of each is extracted. Then we know the $L^*a^*b^*$ reference values, and its possible to calculate ΔE introduced in Section II-C, characterising the difference in tone between ground truth (ColorChecker) and reproduction of the Pillcam. Then one can derive a CCM that applies throughout the relevant video.

B. Colon Gamut

As the results reveal in Section IV, correction based on the ColorChecker leads to saturation. One reason may be, as identified in [6], that PillCams do not process all colours equally. That is, there is a sub-gamut of colours being emphasised. For this reason, we determine the gamut constituting the set of $L^*a^*b^*$ triplets present in the colon, named *colon gamut*.

Obtaining ground truth colours of a human colon directly is not possible. As colonoscopes possess high quality rendering, they are considered the *gold standard*. Therefore, we created a data set from colonoscopy images available on the *Gastrolab Image Gallery* web page[5] and applied them to determine the colon gamut. We chose a data set consisting of 50 images from healthy patients, and applied the *colorcloud* function in MatLab[6] to obtain the colon gamut, depicted in Fig. 2.

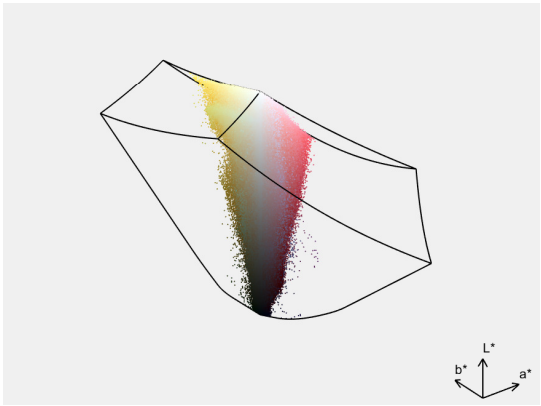
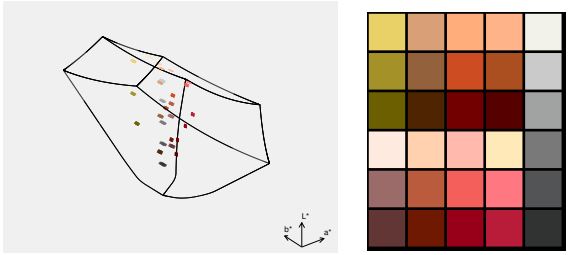


Fig. 2. The $L^*a^*b^*$ values obtained from colonoscopy images, named colon gamut. The black borders refer to the gamut of sRGB.

C. ColonColorChecker (CCC)

We created our own ColorChecker based on the colon gamut, which we named *ColonColorChecker* (CCC). To obtain this, we had to address which colours to choose, what printing material to apply when printing the CCC, and how to measure the $L^*a^*b^*$ values of the printed patches.

It was decided to select 24 colours as shown in Fig. 3(a), to which we added 6 patches for the gray-scale. The patches were first distributed uniformly over the whole colon gamut. Then the CCC was test-printed, and patches with similar colours were further adjusted by moving them further apart in the gamut. The resulting CCC is shown in Fig. 3(b).



(a) (b)

Fig. 3. (a) Distribution of the 30 selected colours allowing to visualise the realised grid (b) ColonColorChecker of 30 custom patches.

Then there is the question of the choice of paper and ink to print the CCC, as this affects the appearance of the colours. Also, during clinical use, the PillCam passes through the patient's digestive system in fluids. In order to reproduce the conditions of a realistic PillCam capture, it was decided to capture the CCC in water (this cannot be done with the ColorChecker, as it is not water resistant). Therefore, ink and

[5]https://www.gastrolab.net/(1/5-23)

[6]https://www.mathworks.com/help/images/ref/colorcloud.html(1/5-23)

paper that do not change significantly when exposed to water were chosen. To ensure the high print quality and accurate colour reproduction, the printing was carried out on an Océ ColorWave 600pp plotter with a coated Tyvek paper[7]. Once the CCC was printed, initial water resistance tests were performed.

1) $L^*a^*b^*$ measurements of printed patches: As we make our own ColorChecker, deviations due to the printing process are of no big concern as long as one stays within the colon gamut, and the patches do not become too similar. The relevant properties can be measured after printing, and it is these values we use as ground truth (or reference). The input arguments for the computation procedure of the correction matrix are the measured $L^*a^*b^*$ values from the PillCam and the ground truth $L^*a^*b^*$ values measured from the printed CCC.

To measure the $L^*a^*b^*$ values of the CCC, we first used the Eye-one Pro spectrophotometer[8], which measures the reflectance of the patches for wavelengths from 380nm to 730nm in increments of 10nm. Five measurements were made after printing, with variation of medium and brightness in the room, to obtain an average and reduce the uncertainties linked to the device and the experimental conditions. The CCC was then placed in water for 30 minutes. Once the CCC was dry, five new measurements were taken to detect any colour variation due to water exposure. Fig. 4 shows the results. The grey bars, 'ref', corresponds to the color values originally chosen before printing. The difference between the blue/orange bars and the grey bars is mainly due to the printing process. What is important is the proximity of the measured values between the dry and wet conditions (blue vs. orange bars). From this, it can be concluded that the paper and the colour patches do not change significantly when exposed to water.

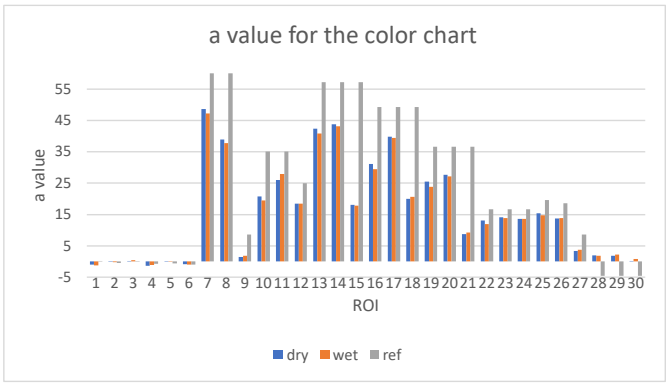
To approach a more realistic situation, we measure the values of $L^*a^*b^*$ with the CCC in water. With no waterproof measuring device available, the measurements had to be done at some distance. We used a TSR CS-2000 spectroradiometer[9], then measured the patches with the CCC placed in a water container. The TSR CS-2000 measures radiance for the given object over wavelengths from 380nm to 780nm with 1nm increment. The TSR CS-2000 was positioned at about 50cm from the target at an angle of $45°$ to the water surface in a room with known and fixed illumination.

The first step was to measure the radiance of a reference plate, an optimal diffuser, to determine the maximum radiance values for the given lighting conditions, and thus to find $L^*a^*b^*$ values independent of the room's lighting. Then the 30 CCC patches were measured. Further, it is necessary to determine the reference values of the reference tile and the power spectral density (PSD) of the PillCam lighting.

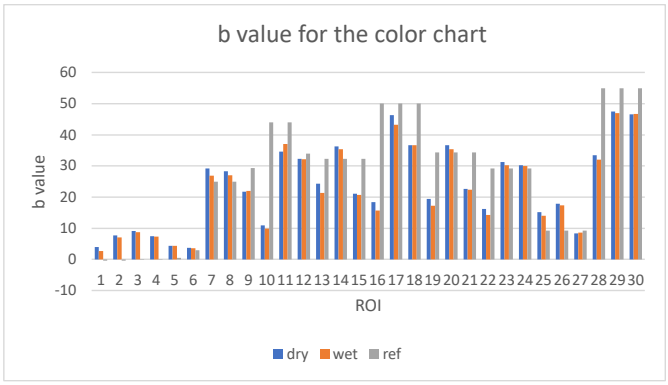
Given the measured data, the following calculations are needed to arrive at the $L^*a^*b^*$ values: First, we have

$$T_c = T_m/T_r, \quad R(\lambda) = P_m/T_c, \tag{5}$$

[7]https://www.dupont.com/tyvekdesign/design-with-tyvek/why-tyvek.html
[8]www.xrite.com/categories/calibration-profiling/i1-solutions (1/5-23)
[9]https://sensing.konicaminolta.us/us/products/cs-2000-spectroradiometer/



(a)



(b)

Fig. 4. Bar graph of a^*- and b^* values for CCC measured with the Eye-one Pro spectrophotometer. In blue, the measurements after printing. In orange, the measurements after immersing the CCC in water for 30 minutes and drying. The grey bar represents the values chosen for the CCC (before printing).

where T_r are values between 0 and 1 correcting for imperfections of the reference tile, T_m are radiance values measured with spectroradiometer with the reference tile as target, and P_m are the radiance values measured with the spectroradiometer against a colour patch immersed in water. Then $(X, Y, Z) = h(R(\lambda), I_{\text{PillCam}}(\lambda))$ and $(L, a, b) = g(X, Y, Z)$, where $I_{\text{PillCam}}(\lambda)$ is the PSD for the PillCam lighting, h is as in Eq (4) using the CIE 1931 1nm colour matching function, and g, the function mapping from the CIE 1931 XYZ colour space to the CIE 1976 $L^*a^*b^*$ space, in Eq. (2) (see Section II-C).

Four different measurements with two different intensity levels in the room's lighting was made to make sure that room lighting had no influence on the obtained values.

The analysis of the 30 patches is done more or less in the same way as for ColorChecker.

IV. Results

A. Deviations within and among videos in data set

With videos from 10 different PillCams it is possible to check if the cameras has a similar reproduction of colours.

We compare ΔE of each patch, for all frames of a video and over all videos, to quantify the deviation of colours. In Fig. 5 we have given one example of boxplots representing

the distribution of $L^*a^*b^*$ values for colour patch 4 (*foliage* or *grass green*), for 5 different PillCams The dashed line

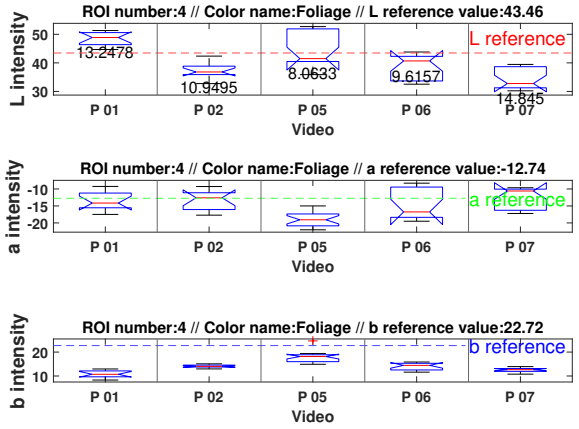


Fig. 5. Boxplots representing $L^*a^*b^*$ values for 5 videos for patch no. 4, foliage green, computed over several captures of the ColorChecker.

represents the true (reference) value (foliage). Note that there is some deviation both within as well as between different videos. However, the deviations are not very large, particularly for the a^* and b^* values. This is also the case for the other 23 patches. Furthermore, the average ΔE measured over all 24 patches for all available videos and cameras is 23.6, with a standard deviation of 12.7. These are quite high values considering that beyond a $\Delta E > 5$, the human visual system can perceive differences.

The statistical study on the 10 videos shows that the colour reproduction is poor (average $\Delta E = 23.6$). In the following, we provide a statistical study on the origin of this large average ΔE value. For this purpose we change the reference value for the ΔE calculation: The first reference are the values from the ColorChecker, the second reference are the values obtained in the 1st video, the third reference are the values obtained from the 2nd video, and so on. In this way, we obtain a comparison of inter-camera variation. Selected results are shown in Fig. 6. These two figures are made for two different patches (ROIs): 1 dark skin and 9 moderate red. Both show that the cameras react in similar ways. Indeed, the comparison of the ΔE deviations for the ColorChecker (dark blue bars), with the other references, shows that each camera captures the colours in a similar way. This capture will be erroneous in view of the ΔE values in relation to the ColorChecker. The analysis shows a high ΔE, but one that is similar for all PillCams. However, as the inter-camera deviation is small, one can determine a global model, and the correction derived can be applied on any PillCam Colon2 without large deviations.

B. Colour Correction

1) Correction using ColorChecker: We select relevant frames and process them as described in Section III-A to derive a colour correction matrix. Knowing the deviation from the ground truth for each patch, the objective is to correct the image to reduce ΔE. To do this, a correction matrix is calculated using a linear least squares fit. The two inputs are the reference $L^*a^*b^*$ values for each ColorChecker patch and the captured $L^*a^*b^*$ values from PillCam.

At first sight, one might think that each frame generates the same correction matrix. However, between two frames under the same experimental conditions (fixed camera and identical lighting conditions), there is a perceptible difference. This leads to different $L^*a^*b^*$ values for each patch and therefore a different correction matrix for each frame. Our approach is to use the average of the correction matrix over all available frames for the same camera. In order to be as general as possible, one could do this for each individual PillCam. However, with the results of Section IV-A in mind, one can assume that different PillCams have similar characteristics. Assuming that the cameras of PillCam COLON 2 have similar characteristics under the same experimental conditions, one can apply the colour correction derived here without having to record a video with the ColorChecker in advance.

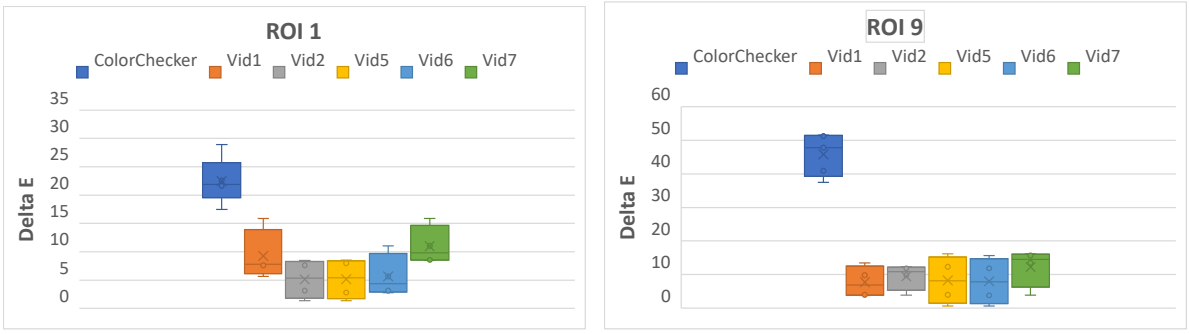
An example of colour correction of a PillCam image is shown in Fig. 7(b) and the original image is shown in Fig. 7(a). One can observe that the colours are quite saturated, and its for this reason that we created the CCC.

2) Correction using ColonColorChecker (CCC): In order to get as close as possible to realistic conditions, the PillCam and CCC were immersed in water. The CCC was placed at 3cm distance from the PillCam in a dark room. That is, the only light present is from the PillCam itself. We had only one PillCam available, so we made two videos with that one. The video did not change significantly between the two takes, in line with the results in Section IV-A. Fig. 7(c) shows an example image correction with the matrix derived on the basis of CCC. The colours reconstructed based on CCC in water are less saturated than those obtained with the ColorChecker, but they are also more realistic than the original image, which is confirmed by the subjective test.

C. Subjective evaluation by Gastroenterologists

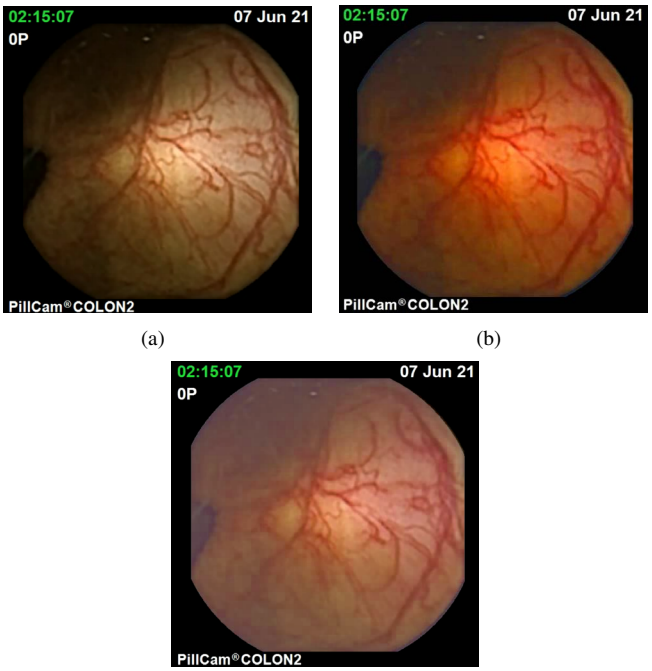
We involved two gastroenterologists with long experience to evaluate 5 videos of 10 seconds length, where the original PillCam video was displayed side by side with videos corrected based on ColorChecker and CCC. The experiment was conducted in a room at Innlandet Hospital Trust, Gjøvik, with the same type of lighting conditions found in rooms typically used for assessment of colonoscopy images. The monitor was calibrated accordingly. We posed the question: *Which video provides the most similar colour reconstruction to that of colonoscopy?* The correction based on CCC and ColorChecker was chosen in 70% and 10% of the cases, respectively, while the original was chosen in 20% of the cases.

A discussion with the gastroenterologists revealed that they preferred the colour reproduction of the CCC corrected videos, but that the contrast is reduced, making it harder to discern detail. This indicates that the colour correction does its thing, but contrast enhancement should be considered in conjunction to obtain better diagnostic value. Also, it would be easier for



(a) (b)

Fig. 6. Boxplots of ΔE values using varying reference for computation for patches: (a) 1 dark skin. (b) 9 moderate red.



(a) (b)

(c)

Fig. 7. Color correction result: (a) Original image. (b) Correction using ColorChecker. (d) Correction using CCC in water.

the gastroenterologists to evaluate diagnostic value if specific pathologies are present in the test videos.

V. Summary and Conclusion

Improvement of colour reproduction could allow better detection of anomalies during clinical examination if done properly. We considered colour correction of videos from PillCam Colon 2 in this paper. For this purpose, we created our own ColorChecker, named ColonColorChecker, which allowed us to select the colours that typically appear in the human colon. The correction matrices we derived should be applicable for any camera of type PillCam Colon 2. On the face of it, the corrections made provide a clear improvement of colour in clinical videos. This was also concluded by two gastroenterologist who evaluated the colour correction in several videos. However, currently the contrast is somewhat lower in the corrected videos than in the original, reducing the diagnostic value of the approach.

In future work we will aim at improving the diagnostic value where contrast enhancement will be applied in conjunction with the colour correction. We will conduct more thorough subjective tests with several more candidates choosing video clips containing specific pathologies. This will make it possible to conclude more firmly about the diagnostic value of the approach. Another investigation is to check if artificial intelligence algorithms can better detect anomalies if they are fed images with corrected colours.

VI. Acknowledgement

We would like to give our appreciation to gastroenterologist Snorri Olafsson for participating in the subjective evaluation.

References

[1] "The guide to clinical preventive services 2014," Online: https://www.ncbi.nlm.nih.gov/books/NBK235846, recommendations of the U.S. Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US). 2014 May.

[2] M. Bugajski, P. Wieszczy, G. Hoff, M. Rupinski, J. Regula, and M. F. Kaminski, "Modifiable factors associated with patient-reported pain during and after screening colonoscopy," *Gut*, vol. 67, no. 11, pp. 1958–1964, 2018.

[3] A. Wang, S. Banerjee, B. Barth, Y. Bhat, S. Chauhan, K. Gottlieb, V. Konda, J. Maple, F. Murad, P. Pfau, D. Pleskow, U. Siddiqui, J. Tokar, and S. Rodriguez, "Wireless capsule endoscopy," *Gastrointestinal Endoscopy*, vol. 78, pp. 805–815, Dec. 2013.

[4] A. Badano, C. Revie, A. Casertano, and e. al, "Consistency and standardization of color in medical imaging: a consensus report," *Journal of Digital Imaging*, vol. 28, pp. 41–52, Feb. 2015.

[5] C. S. McCamy, H. Marcus, and J. G. Davidson, "A color-rendition chart," *Journal of Applied Photographic Engineering*, vol. 2, no. 3, pp. 95–99, 1976.

[6] H. Vu, T. Echigo, K. Yagi, H. Okazaki, Y. Fujiwara, Y. Yagi, and T. Arakawa, "Image-enhanced capsule endoscopy preserving the original color tones," in *Abdominal Imaging. Computational and Clinical Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–43.

[7] M. S. Tooms, *Colour Reproduction in Electronic Imaging Systems: Photography, Television, Cinematography*. West Sussex, UK: John Wiley & Sons, 2015.

[8] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972.

A Quality-Oriented Database for Video Capsule Endoscopy

Tan-Sy Nguyen[1,2], Marie Luong[2], John Chaussard[1], Azeddine Beghdadi[2], Hatem Zaag[1], Thuong Le-Tien[3]
[1] *Université Sorbonne Paris Nord, LAGA, CNRS, UMR 7539, Laboratoire d'excellence Inflamex*,
F-93430, Villetaneuse, France
[2] *Université Sorbonne Paris Nord, L2TI, UR 3043, F-93430*, Villetaneuse, France
[3] *Ho Chi Minh City University of Technology*, Vietnam
{tansy.nguyen, chaussard, zaag}@math.univ-paris13.fr
{marie.luong, azeddine.beghdadi}@univ-paris13.fr, thuongle@hcmut.edu.vn

***Abstract*—In this paper, we propose a novel dataset called the Quality-Oriented Database for Video Capsule Endoscopy (QVCED) which serves as the primary crucial resource for evaluating the quality of Wireless Capsule Endoscopy (WCE) images and videos. Serving as a benchmark, the QVCED encourages the design of learning-based enhancement methods to address image quality assessment and enhancement challenges in WCE. This comprehensive dataset consists of a large number of WCE videos encompassing common distortions encountered in clinical practice, including noise, defocus blur, motion blur, and uneven illumination. Moreover, video quality has been intentionally degraded at varying distortion severity levels to faithfully replicate real-world conditions. The extensive analysis demonstrates the diversity and practical relevance of this dataset in the WCE domain that motivates the advancement of a more precise diagnosis regarding gastrointestinal disorders. The complete dataset is publicly available through the following link: https://cloud.math.univ-paris13.fr/index.php/s/b74TQk7mMpHDXKT.**

***Index Terms*—Wireless Capsule Endoscopy, Video Quality Assessment Dataset, Subjective Evaluation.**

I. Introduction

Wireless Capsule Endoscopy (WCE) has revolutionized medical practices for gastrointestinal (GI) disease screening and diagnosis [1]. However, a major challenge in WCE is obtaining optimal image quality, which directly affects diagnostic accuracy. Indeed, WCE image quality suffers from distortions due to the limitations of the sensor technology and the constrained acquisition environment. For example, narrow apertures and small sensors with limited dynamic range and sensitivity generate noise within captured frames [2]. Especially, additive white Gaussian noise in WCE images is the accepted standard model [3]. Unstable environments result in excessive blurriness [4] due to the uncontrolled and random motion of the capsule, while the capsule's limited lighting coverage cause uneven illumination [5]. These distortions can decrease the performance of tasks like lesion detection, recognition, and tracking in the gastrointestinal tract.

To address image quality limitation issues, due to the aforementioned distortions, numerous learning-based algorithms [2], [6], [7] have been proposed. Specifically, recent advancements in image restoration and enhancement techniques rely on learning-based methods that require pairs of corrupted and clean images for training. However, in the case of WCE, the absence of a dedicated quality assessment dataset poses a significant challenge. Therefore, a dataset specifically designed for assessing the quality of WCE images, with varying levels of distortions, is crucial for developing accurate and reliable image enhancement algorithms. To the best of our knowledge, there is currently no specialized dataset available specifically for assessing video quality in WCE. Existing datasets commonly used for quality assessment, such as LIVE Mobile VQA [8], KoNViD-1k [9], TID2013 [10], CSIQ [11], and CID:IQ [12], have primarily focused on natural images for over two decades. In the field of medical imaging, datasets like RIQA [13] and LVQ [14] have been developed specifically for retinal and laparoscopic image/video evaluation, respectively. However, it is important to note that these datasets are not efficient for training learning-based quality enhancement techniques for WCE images due to the inherent dissimilarities in medical imaging types and modalities. Moreover, most medical databases are tailored for segmentation and classification tasks, making this work a valuable contribution to fulfill a real requirement in medical imaging and in particular on evaluating and improving WCE image quality.

Consequently, toward the demand for a comprehensive video quality assessment dataset, we propose the Quality-Oriented Database for Video Capsule Endoscopy (QVCED), derived from the Kvasir-Capsule dataset [15]. QVCED covers a wide range of scenarios with different pathologies and multiple types of distortions, prioritizing realistic conditions. The dataset is produced through a two-stage process. In the first stage, reference videos that meet the required quality criteria are carefully selected from the Kvasir-Capsule dataset [15]. Next, the reference video is thus subjected to a degradation process in which a controlled level of degradation is applied by means of the physical parameters of the used distortion generation model.

The subsequent sections of this paper are structured as follows: Section II describes the creation process of the QVCED dataset. Within this section, Section II-A outlines the initial selection process for reference videos, while Section II-B explains the generation of simulated distortions in the chosen reference videos. Afterward, Section III focuses on the analysis and discussion of the proposed dataset. Specifically,

979-8-3503-4218-5/23/$31.00 ©2023 IEEE

Section III-A presents the implementation and results of a subjective test, including opinion scores from expert and non-expert observers regarding the simulated distortions. Furthermore, Section III-B analyzes the content diversity of the QVCED dataset. Finally, this paper concludes in Section IV, summarizing the key findings and implications of this work in terms of the creation and analysis of the proposed dataset.

II. Proposed Dataset - QVCED

In this section, we describe the dataset creation process. We provide a comprehensive overview of the methodologies encompassing the selection of reference videos (Section II-A) and the simulation of distortions applied to the chosen reference videos (Section II-B).

The first step is to select reference videos from an existing WCE dataset. We have created a dataset comprising 20 original reference videos extracted from the Kvasir-Capsule dataset [15]. These videos have a duration of 10 seconds, a resolution of 336×336 pixels, and a frame rate of 30 frames per second (fps). The following subsection provides a comprehensive description and the selection process of the reference videos.

A. Reference Videos Selection

The selection of the reference videos aimed at optimizing a wide range of pathological scenarios and maximizing continuous temporal information which enables a thorough evaluation and analysis of quality enhancement algorithms, facilitating advancements in the WCE domain.

To ensure scene content diversity, the QVCED dataset includes fourteen distinct categories. These categories encompass various WCE findings such as Pylorus (PY), Ampulla of Vater (AV), Ileocecal Valve (IV), Normal Clean Mucosa (NCM), Reduced Mucosal View (RMV), Blood-Fresh (BF), Blood-Hematin (BH), Foreign Body (FB), Erythema (ERY), Angiectasias (ANG), Erosion (ERO), Ulcers (ULC), Lymphangiectasia (LYM), and Polyp (PYL). This diverse selection allows for a comprehensive evaluation of algorithms and techniques in the field of wireless capsule endoscopy, covering a broad spectrum of medical scenarios commonly encountered in clinical practice.

Some images from reference videos are shown in Fig. 1. These metrics help identify the highest-quality videos for each finding, ensuring that the chosen reference videos meet the required quality standards which enhances the dataset's reliability and usefulness for various research purposes.

1) Noise Assessment: To estimate the noise level, we employ the fast noise variance estimator proposed by Immerkaer [16]. The process begins by converting the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ into grayscale, resulting in the grayscale image $\mathbf{I}_{gray} \in \mathbb{R}^{H \times W}$ where H, W are the height and width of the images. A noise estimation mask, denoted as $\mathbf{M} \in \mathbb{R}^{3 \times 3}$, is then used to estimate the standard deviation of additive white Gaussian noise in the image. This mask is derived from two elements approximating the Laplacian of the image. The estimated standard deviation of the noise $\hat{\sigma}_n$ is computed as follows:

$$\hat{\sigma}_n = \sqrt{\frac{\pi}{2}} \frac{\Sigma_{x,y} |\mathbf{I}_{gray}(x,y) * \mathbf{M}|}{6(W-2)(H-2)}, \tag{1}$$

2) Blur Assessment: To measure the level of blur in an image, the Perceptual Blur Index (PBI) [17] was used as a thresholding metric. The PBI metric takes into account the perceptual differences in how the Human Visual System (HVS) perceives the addition of blur to an already blurred image compared to a sharp one. Mathematically, the PBI is defined as the difference between the total radial energy of the input image, denoted as $ER(w)$, and the total radial energy of its binomial filtered version, denoted as $ER_f(w)$. The formula for calculating the PBI is as:

$$PBI = log\left(\frac{1}{w_{max}} \Sigma_w |ER(w) - ER_f(w)|\right) \tag{2}$$

$$ER(w) = \frac{1}{K} \Sigma_K |F(w, \theta_k)|, \theta_k = \frac{k\pi}{K}, \tag{3}$$

$$ER_f(w) = \frac{1}{K} \Sigma_K |F_f(w, \theta_k)|, \theta_k = \frac{k\pi}{K}, \tag{4}$$

where $F(w, \theta_k)$ and $F_f(w, \theta_k)$ represent the centered Fourier coefficients of the input images and its binomial filtered version, respectively, in the polar coordinates.

3) Uneven Illumination Assessment: To assess the presence of uneven illumination, the Illumination Histogram Equalization Difference (IHED) [18] is employed. IHED measures the impact of histogram equalization (HE) on the spatial distribution of background illuminance (BI). The evaluation process involves converting the image into the HSV color space to extract the brightness channel $\mathbf{V} \in \mathbb{R}^{H \times W}$. Subsequently, the background illuminance $\mathbf{BI}(x,y) \in \mathbb{R}^{H \times W}$ is extracted through the application of a low-pass filtering method of size $h = \frac{H}{4}$. Finally, IHED is calculated using the following formula:

$$IHED = \frac{\sigma_D}{\frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \mathbf{BI}(x,y)}, \tag{5}$$

where σ_D is the standard deviation of the difference signal $\mathbf{D}$, which is computed as:

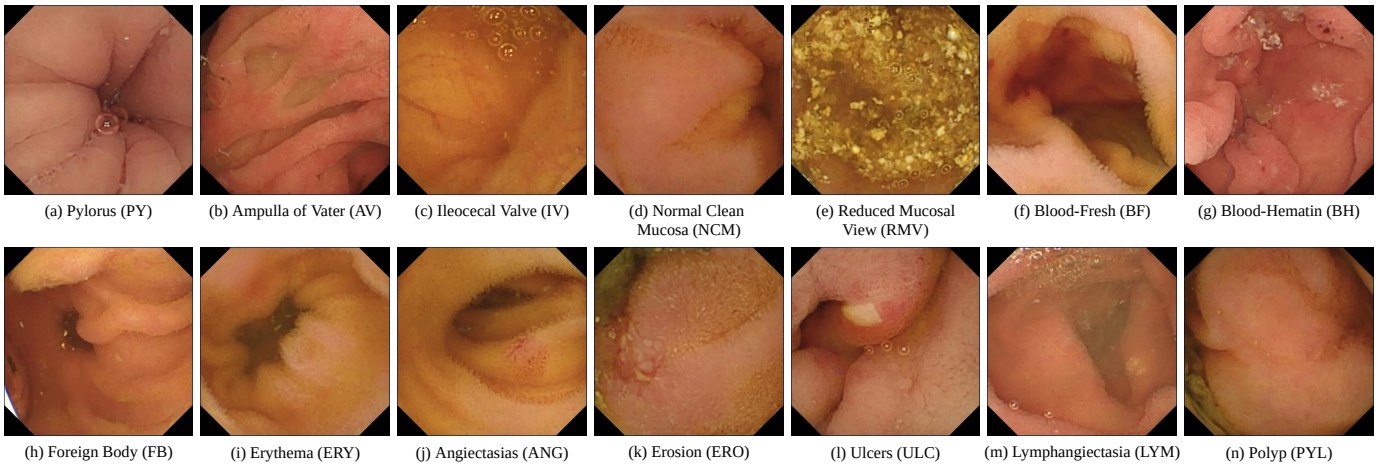
$$\mathbf{D}(x,y) = \mid \mathbf{BI}(x,y) - \mathcal{T}(\mathbf{BI}(x,y)) \mid, \tag{6}$$

where $\mathcal{T}$ denotes the histogram equalization transformation.

A video is considered acceptable for use as a reference only if the levels of all three distortions (i.e., noise, blur, and uneven illumination) are below a predetermined threshold. Once the reference video is chosen, the next subsection will outline how we simulated distortions on these selected reference videos.

B. Distortion Generation

We have integrated four prevalent degradations consisting of noise, uneven illumination, defocus, and motion blur into our extensive dataset. To ensure a faithful reproduction of each distortion, we have applied suitable mathematical models to every individual frame of the reference video. In the current stage of our research, we only added one type of distortion to each video, with the same severity throughout its entirety.



(a) Pylorus (PY) (b) Ampulla of Vater (AV) (c) Ileocecal Valve (IV) (d) Normal Clean Mucosa (NCM) (e) Reduced Mucosal View (RMV) (f) Blood-Fresh (BF) (g) Blood-Hematin (BH)

(h) Foreign Body (FB) (i) Erythema (ERY) (j) Angiectasias (ANG) (k) Erosion (ERO) (l) Ulcers (ULC) (m) Lymphangiectasia (LYM) (n) Polyp (PYL)

Fig. 1. The frames extracted from reference videos in the QVCED dataset represent a diverse range of findings.

1) Noise Model: Noise is a common distortion in video systems, particularly in WCE. It is caused by narrow apertures, small sensors, and limited dynamic range [2] and negatively impacts the effectiveness of the endoscopic examination process. In our study, we included the Additive White Gaussian Noise (AWGN) model in our dataset which assumes that the video noise follows a Gaussian distribution. Mathematically, the distorted image can be represented as:

$$\mathbf{I}_{noisy} = \mathbf{I} + \mathbf{N} \tag{7}$$

where $\mathbf{I}$ represents the original image, and $\mathbf{N} \sim \mathcal{N}(0, \sigma_n^2)$ represents the random noise value following a Gaussian distribution with standard deviation σ_n. To control the severity, AWGN level is configured with $\sigma_n \in \{5, 10, 20, 30\}$.

2) Defocus Blur Model: In WCE, the wireless capsule is equipped with a fixed-focus lens endoscope [19]. This design introduces defocus blur when objects in the scene are not precisely at the camera's focal distance. To simulate defocus blur, a low-pass filtering of the input image using an isotropic Gaussian impulse response as shown in Fig. 2a is commonly used. The isotropic Gaussian kernel is used to simulate the rotational symmetry around the optical axis of the blurring effect. The impulse response associated with this blur, denoted as $h_{db}(x, y)$, is defined as:

$$h_{db}(x, y) = \frac{1}{2\pi\sigma_{db}^2} \exp \frac{-(x^2 + y^2)}{2\sigma_{db}^2}, \ \sigma_{db} \in \{1, 2, 3, 5\} \tag{8}$$

The blurring extent is determined by the standard deviation parameter σ_{db}. Increasing σ_{db} leads to stronger smoothing and more noticeable blurring effects. The size of the convolution mask, denoted as W_{db}, is chosen to preserve the energy of the filtered image signal. To preserve 99% of the total energy of the Gaussian, a size of $6\sigma_{db}$ at least is required. The filter size W_{db} should be an odd number as:

$$W_{db} = 2 \times \lceil 3\sigma_{db} \rceil + 1 \tag{9}$$

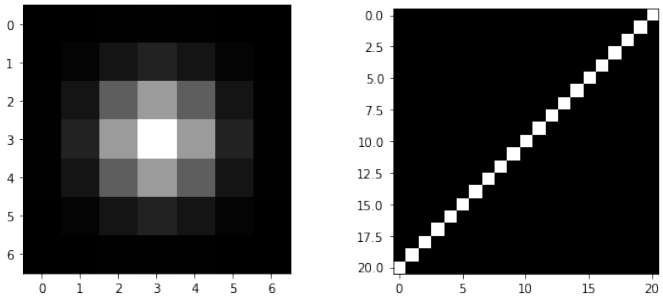
Fig. 2a illustrates an example of a defocus blur kernel of standard deviation $\sigma_{db} = 1$.

3) Motion Blur Model: The rapid and sudden movements of the capsule endoscope can cause blurring, influenced by factors such as fast camera motions at low frame rates, the inability to adjust lens focus, camera mechanism instability, and sensor sensitivity to light variations [19]. When the capsule endoscope moves in a straight line, it results in linear motion blur. The blur kernel, denoted as h_{mb}, can be formulated using two known parameters: the direction of motion blur θ_{mb} and the length of motion blur $L_{mb} \in \{5, 10, 15, 25\}$. The formulation is as follows:

$$h_{mb}(x, y) = \begin{cases} \frac{1}{L_{mb}}, & \text{if } \sqrt{x^2 + y^2} \leq \frac{L_{mb}}{2}, -\tan\theta_{mb} = \frac{x}{y}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

We evaluate the motion in each frame in order to add a motion blur corresponding to the actual motion of the video. In the initial stages, the Lucas-Kanade method [20] is commonly used for estimating the movement direction of a capsule endoscope through optical flow estimation. Two consecutive frames captured by the capsule endoscope are subtracted and Otsu thresholding technique [21] is applied to generate a foreground binary map. The optical flow is then estimated using the Lucas-Kanade method on the center of gravity of the foreground. Fig. 2b illustrates an example of a motion blur kernel, which is configured by two parameters: the direction θ_{mb} and the length L_{mb}.

4) Uneven Illumination Model: The motion of the capsule endoscope, caused by the gastrointestinal tract's peristaltic activity and limited capsule light, can introduce uneven illumination. To simulate this effect, the reference image is first converted from the RGB color space to the HSV color space. Then, we perform a pointwise multiplication of the reference image with a mask. The coefficients of this mask, which are determined in the spatial plane (Fig. 3a), can be represented mathematically as a hybrid distribution that integrates two-dimensional distributions from one the normal variable and one log-normal variation [22]. Fig. 3b,c show the generated masks of the hybrid distribution in two different



(a) Defocus blur kernel (b) Motion blur kernel

Fig. 2. Visual representation of the blurring kernel, characterized by a defocus blur standard deviation of $\sigma_{db} = 1$ and motion blur with $L_{mb} = 20, \theta_{mb} = \frac{\pi}{4}$, respectively.

angles θ, respectively. However, in this preliminary work, only a simulated circular-gradient mask was taken into account. As depicted in Fig. 3d, the mask $\mathbf{M}(x, y) \in \mathbb{R}^{H \times W}$ is defined to conform to the dimensions of the original image.

$$\mathbf{M}(x, y) = 255 - \left[\frac{2\Delta_I}{W} \sqrt{(x - x_c)^2 + (y - y_c)^2} \right], \quad (11)$$

where $\mathbf{M}(x_c, y_c)$ is the circle center at coordinates $(x_c, y_c) \in \{(112, 224), (168, 168), (224, 224)\}$. To achieve varying levels of illumination, the difference in intensity between the brightest pixel of the image and the darkest pixel is set as $\Delta_I \in \{80, 100, 135, 170\}$. In the future, we plan on using the previously defined hybrid distribution as a mask.

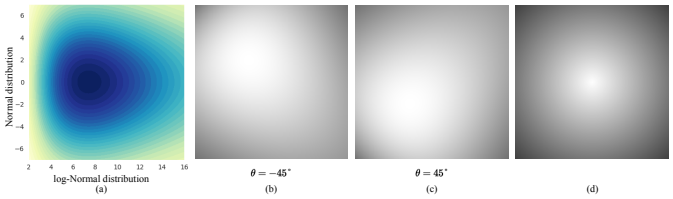


Fig. 3. Gradient masks to simulate the Uneven Illumination.

In summary, the dataset creation process involved the simulation of four distinct types of distortion, each type having four different severity levels. This resulted in a total of 320 degraded videos in the QVCED dataset. A comprehensive overview of the dataset, including specific details such as the types of distortions and severity levels, can be found in Table I.

In the following section, some extensive experiments are conducted to analyze the diversity and practical applicability of the proposed dataset.

III. Dataset Analysis

To evaluate the proposed dataset, some experimental studies were conducted. First, a subjective test (Section III-A) was performed to assess and validate the dataset's quality and its alignment with human perception. This test evaluates how well the dataset is perceived by human observers and ensures its overall quality. In a second round of experiments, statistical features of the dataset were analyzed to verify its content diversity (Section III-B).

TABLE I
Summary of the proposed wireless capsule endoscopy video quality assessment dataset.

Number of Reference Videos	20	**Number of Distorted Videos**	320
Number of Findings	14	**Pathologies**	6
Resolution of Videos	336×336	**Frame Rate**	30
Duration	10s	**Video Type**	.mp4
Number of Distortions	4	**Level of Distortion**	4
Distortion Types	Noise, Defocus Blur, Motion Blur, Uneven Illumination		

A. Subjective Test

Prior to the main WCE subjective test, observers underwent the Ishihara 38 plates CVD verification [23] to detect any red-green color deficiencies. Participants with an accuracy above 70% were selected to complete the WCE subjective test, ensuring normal color perception for accurate evaluation.

To conduct the WCE subjective testing process efficiently, a pairwise-comparison protocol based on the ITU-T standard [24] was implemented, following the described testing environment.

1) Testing Environment: In the WCE quality assessment subjective test, observers were presented with randomized pairs of distorted videos and corresponding reference videos. Randomization was implemented to eliminate presentation order bias. For each video pair, observers were asked to provide an opinion score indicating the perceived severity of distortion. The implemented four-point scale corresponds to four distortion severity levels including: (1) Hardly Visible, (2) Just Noticeable, (3) Annoying, and (4) Very Annoying. The obtained opinion scores allowed us to assess the subjective quality of the distorted videos compared to the corresponding reference videos. The Mean Opinion Score (MOS) for a video is the average score given by observers for that video.

An online platform (shown in Fig. 4) was developed and designed to facilitate the conduction of subjective tests. The platform underwent thorough optimization to ensure usability, including aspects such as background color, button size, and position. These optimizations aimed to enhance the testing experience and make it convenient for observers to effectively complete the task following the ITU-T standard. The source code of the platform is available at: https://github.com/tansyab1/WCETest.

A total of 34 individuals, comprising 12 experts and 22 non-experts, with diverse age groups and backgrounds, took part in subjective the experiments. Fig. 5 displays the distribution of participants' age and the duration of their involvement in the subjective test. Participants across various age groups were included in the subjective test, as shown in Fig. 5a. This diverse age distribution enhances the reliability of the test

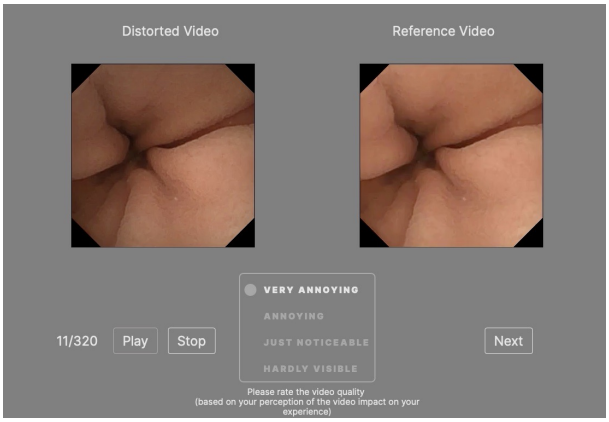
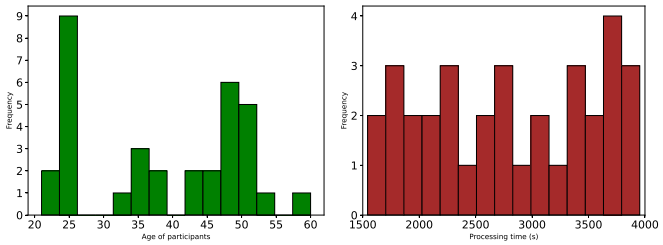


Fig. 4. WCE subjective test window.



(a) Age (b) Processing time (s)

Fig. 5. Age and the processing time distributions of observers participated in the subjective experiment.

outcomes by avoiding biases toward any specific age category. Moreover, it is clearly noticeable from Fig. 5b that each participant dedicated a minimum of approximately 5 seconds to evaluate each video, demonstrating their focused attention and commitment to efficient testing. This statistic affirms the credibility and applicability of the test's outcome.

2) Video Quality Score: As mentioned earlier, the evaluation includes four levels of distortion. Level 1 represents the minimal degradation of distortion, while level 4 represents the most severe condition, where a higher value indicates a lower-quality perception for the video observer. Fig. 6 compares

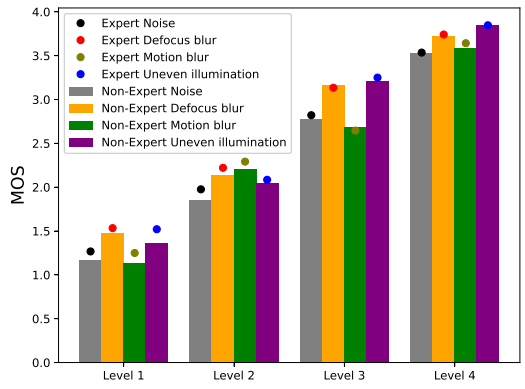


Fig. 6. Comparison of the subjective score regarding experts and non-experts

the mean scores of experts and non-experts for the proposed dataset. The presented data illustrates that experts and non-experts hold a significant correlation between the obtained scores. However, the experts exhibit a heightened level of attention toward specific tasks, which enhances their sensitivity to even the slightest deviations. Therefore, the dissimilarity is more conspicuous when examining videos exhibiting low levels of distortion.

B. Diversity Data Analysis

To analyze the dataset's diversity and broad applicability, experiments were conducted to verify significant variations in video content. These experiments provide valuable insights and benefits for various image-processing tasks. A diverse dataset serves as a valuable resource for benchmarking, validation, and training, facilitating significant advancements in image processing.

To evaluate the content diversity of the datasets, we used deep features of dimension 4096 extracted from a pre-trained VGG-16 [25] on ImageNet [26]. By employing t-SNE (t-distributed Stochastic Neighbor Embedding) [27], we projected these high-dimensional features onto a 2D subspace. The resulting visualization, shown in Fig. 7, succinctly represents the content diversity across the datasets. The broad spectrum displayed in the visualization illustrate the extensive range and variety of visual content present in the dataset.

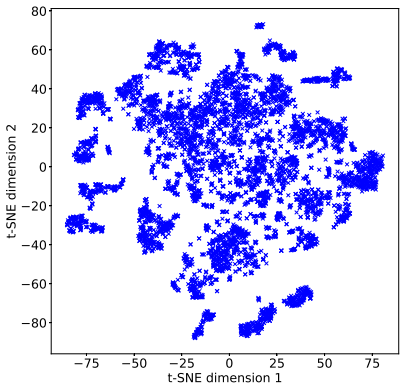
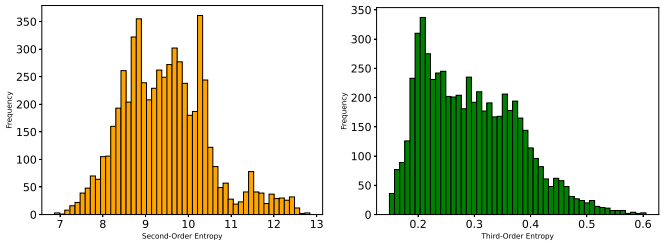


Fig. 7. t-SNE visualization of the embedded feature generated from 20 reference videos by VGG-16 pre-trained network

In addition to the primary analysis, a further examination of the dataset is conducted to analyze the distribution of image entropies, considering both spatial and temporal information. Second-order entropy analysis provides insights into the spatial features within the dataset, where higher entropy indicates a greater content diversity of images. Furthermore, the incorporation of third-order entropy analysis takes into account inter-frame information. By considering the relationships and dependencies between consecutive frames, the third-order entropy provides a deeper understanding of the temporal dynamics and variations within the dataset. This analysis offers a comprehensive perspective on the dataset's complexity and richness. In this work, calculations were performed on

a dataset of 6000 images from 20 reference videos. Fig. 8 illustrates the broad histogram of the entropies, indicating a wide range of visual features and affirming the practical applicability, diversity, and usefulness of the QVCED dataset.



(a) Second-Order entropy (b) Third-Order entropy

Fig. 8. Distribution of Variations and Entropy among 20 reference videos.

IV. Conclusion

In this work, we have introduced a preliminary quality assessment dataset specifically designed for wireless capsule endoscopy. QVCED comprises four distinct distortion types, with each type further subdivided into four levels, resulting in a total of sixteen variations. This dataset serves as a quality assessment resource specifically targeting the WCE domain. Especially, it addresses the previously neglected data challenge and offers valuable insights for evaluating and analyzing the effectiveness of image and video processing algorithms in this particular field. The dataset's strength lies in the extensive diversity of its visual content, enabling researchers to tackle demanding real-world contexts. In this work, the addition of synthetic distortion to a given frame may not have a noticeable impact if the frame is already affected by authentic distortion. In the future, we could remove any existing distortion before applying a synthetic one.

Acknowledgment

We acknowledge support from the Investissements d'Avenir programme ANR-11-IDEX-0005-02 and 10-LABEX-0017, Sorbonne Paris Cité, Laboratoire d'excellence INFLAMEX.

References

[1] M. W. Alam, M. H. A. Sohag, A. H. Khan, T. Sultana, and K. A. Wahid, "Iot-based intelligent capsule endoscopy system: A technical review," *Intelligent Data Analysis for Biomedical Applications*, pp. 1–20, 2019.

[2] S. Zou, M. Long, X. Wang, X. Xie, G. Li, and Z. Wang, "A cnn-based blind denoising method for endoscopic images," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4.

[3] S. Suman, F. A. Hussin, A. S. Malik, N. Walter, K. L. Goh, I. Hilmi, and S. h. Ho, "Image enhancement using geometric mean filter and gamma correction for wce images," in *Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part III 21*. Springer, 2014, pp. 276–283.

[4] H. Liu, W.-S. Lu, and M. Q.-H. Meng, "De-blurring wireless capsule endoscopy images by total variation minimization," in *Proceedings of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2011, pp. 102–106.

[5] Y. Chen and J. Lee, "A review of machine-vision-based analysis of wireless capsule endoscopy video," *Dia. and The. Endoscopy*, 2012.

[6] Y. Wang, C. Cai, and Y. Zou, "Single image super-resolution via adaptive dictionary pair learning for wireless capsule endoscopy image," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 595–599.

[7] V. B. S. Prasath, D. N. Thanh, L. T. Thanh, N. San, and S. Dvoenko, "Human visual system consistent model for wireless capsule endoscopy image enhancement and applications," *Pattern Recognition and Image Analysis*, vol. 30, pp. 280–287, 2020.

[8] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[9] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[10] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Color image database tid2013: Peculiarities and preliminary results," in *European workshop on visual information processing (EUVIP)*. IEEE, 2013, pp. 106–111.

[11] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.

[12] X. Liu, M. Pedersen, and J. Y. Hardeberg, "Cid: Iq–a new image quality database," in *Image and Signal Processing: 6th International Conference, Cherbourg, France*. Springer, 2014, pp. 193–202.

[13] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," in *Medical Image Computing and Computer Assisted Intervention–MICCAI*. Springer, 2019, pp. 48–56.

[14] Z. A. Khan, A. Beghdadi, F. A. Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, Å. A. Fretland, B. Edwin, and O. J. Elle, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, vol. 11316, 2020, pp. 129–136.

[15] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021.

[16] J. Immerkaer, "Fast noise variance estimation," *Computer vision and image understanding*, vol. 64, no. 2, pp. 300–302, 1996.

[17] A. Chetouani, A. Beghdadi, and M. Deriche, "A new reference-free image quality index for blur estimation in the frequency domain," in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2009, pp. 155–159.

[18] T.-S. Nguyen, J. Chaussard, M. Luong, H. Zaag, and A. Beghdadi, "A no-reference measure for uneven illumination assessment on laparoscopic images," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 4103–4107.

[19] J. Park, Y. Hwang, J.-H. Yoon, M.-G. Park, J. Kim, Y. J. Lim, and H. J. Chun, "Recent development of computer vision technology to improve capsule endoscopy," *Clinical endoscopy*, vol. 52, pp. 328–333, 2019.

[20] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[21] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[22] S. J. Fletcher and M. Zupanski, "A hybrid multivariate normal and log-normal distribution for data assimilation," *Atmospheric Science Letters*, vol. 7, no. 2, pp. 43–46, 2006.

[23] J. Clark, "The ishihara test for color blindness." *American Journal of Physiological Optics*, 1924.

[24] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Recommendation P910*, 2008.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[27] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

FEES-IS: Real-time Instance Segmentation of Flexible Endoscopic Evaluation of Swallowing

Weihao Weng, Xin Zhu
Graduate School of Computer Science and Engineering
The University of Aizu
Aizuwakamatsu, Japan
{d8232117, zhuxinl}@u-aizu.ac.jp

Mitsuyoshi Imaizumi, Shigeyuki Murono
Department of Otolaryngology
Fukushima Medical University
Fukushima, Japan
{ima-mitu, murono}@fmu.ac.jp

***Abstract*—Instance segmentation offers advantages over semantic segmentation in medical image analysis by providing more detailed information for accurate identification and tracking of individual objects. However, existing instance segmentation methods for medical image do not always account for limited and variable data, resulting in overfitting. In order to overcome the aforementioned limitations, this paper prposes a novel one-stage end-to-end deep learning framework, named FEES-IS, which is tailored to perform real-time instance segmentation on Flexible endoscopic evaluation of swallowing (FEES) videos. The model incorporates sparse attention to prevent overfitting. Moreover, we propose a loss function that improves instance medical image segmentation accuracy. The study used a dataset of 199 annotated FEES videos to train the model, which was subsequently evaluated on an additional 40 videos from patients who underwent consecutive FEES procedures at Fukushima Medical University Hospital between December 2016 and August 2019. The results show that FEES-IS achieved a mean average precision (mAP) of 61.64 at a frame rate of 41.7 frames per second on a single NVIDIA GeForce RTX 3090 graphics processing unit. This performance is promising and suggests that the proposed FEES-IS model has the potential to aid in the accurate identification and tracking of individual objects in medical images obtained from FEES procedures.**

***Index Terms*—Real-time, Instance segmentation, Deep neural network**

I. Introduction

Dysphagia is a common swallowing disorder that occurs in patients with neurological diseases and strokes [1]. Oropharyngeal dysphagia is usually caused by abnormalities in the oral cavity, pharynx, or esophageal sphincter, and it can lead to severe complications, including aspiration pneumonia, malnutrition, and dehydration. Therefore, the prevention and treatment of aspiration are critical in improving the survival rates of stroke survivors.The gold standard methods for studying oropharyngeal dysphagia are flexible endoscopic evaluation of swallowing (FEES) and videofluoroscopic swallow study (VFSS). While FEES has advantages over VFSS, inexperienced doctors may find FEES difficult to interpret accurately [2], [3]. For example, Imaizumi et al. [4] revealed a notable discrepancy in the texture-modified diet recommendations made by experienced and inexperienced examiners after FEES. Specifically, the experienced examiner advised a diet that was more similar to a normal diet compared to the inexperienced examiner. This difference in dietary recommendations is noteworthy because an improperly prescribed texture-modified diet can increase the risk of malnutrition or aspiration, which can have significant clinical implications. To address this issue, AI-assisted FEES using instance segmentation is a potential solution. Instance segmentation offers a higher level of detail and precision in understanding the swallowing process compared to semantic segmentation. It not only labels different regions of interest but also distinguishes individual objects or instances within those regions. This level of granularity is crucial in accurately identifying and tracking specific structures involved in swallowing, including the epiglottis, vocal folds, or pharyngeal walls. By differentiating these instances, instance segmentation enables a more comprehensive analysis of swallowing, leading to improved diagnosis and treatment planning for dysphagia.

We observe a significant imbalance between the number of studies dedicated to semantic segmentation and instance segmentation in the context of medical videos. This discrepancy can be attributed to the well-established status of semantic segmentation, which has been extensively used in medical image processing, such as tumor and organ segmentation [5]. On the other hand, instance segmentation is rather new, and its use in medical video analysis is still relatively limited. Nevertheless, instance segmentation offers several advantages over semantic segmentation in the video recording of FEES for dysphagia diagnosis [6]. By enabling the separation of objects into individual instances, instance segmentation can accurately delineate object boundaries, thereby facilitating diagnosis and treatment planning.

The FEES-IS system achieves a remarkable frame rate exceeding 30 frames per second (FPS) while accurately segmenting specific regions of interest. This paper presents several notable contributions:

- Introduction of the novel framework, FEES-IS, which is specifically designed for real-time instance segmentation in FEES videos. This framework effectively addresses the inherent limitations commonly observed in existing instance segmentation methods typically employed in medical image analysis.

This study was partially supported by the Competitive Research Fund, The University of Aizu (2023-P-4).

- Integration of sparse attention within the FEES-IS model to alleviate the issue of overfitting. By incorporating this attention mechanism, the model achieves superior generalization capabilities, particularly in scenarios characterized by limited and variable data. As a result, the model's performance in instance segmentation tasks is significantly enhanced.
- Proposal of a novel loss function that enhances the accuracy of instance medical image segmentation. The implementation of this loss function facilitates more precise identification and tracking of individual objects within FEES videos, leading to improved overall outcomes.

II. RELATED WORK

A. Instance Segmentation

Instance segmentation accuracy has been a key focus of research, with MaskRCNN [7] serving as a popular two-stage approach. However, the need for re-pooling features for each region-of-interest renders these methods unsuitable for real-time applications. While one-stage methods are faster, they still entail complex computations that limit their speed. YOLACT [8] is a real-time instance segmentation method that achieves rapid performance by leveraging anchor-free detection and instance-aware feature normalization. YOLACT-like methods integrate additional optimizations or modifications to enhance performance, such as prototype refinement. These real-time instance segmentation methods have significant implications for medical image analysis applications where swift and precise segmentation is essential for diagnosis and treatment planning. For instance, YOLACT++ [9] incorporates several optimizations, including prototype refinement and a more sophisticated ResNet101-FPN backbone network, to improve feature representation. PolarMask [10] uses a polar coordinate representation to enhance detection of small and overlapping instances in an image. On the other hand, YOLACTEdge [11] integrates edge detection to enhance instance segmentation performance in low-contrast or cluttered environments by fusing edge maps with instance segmentation results to improve boundary detection. Nonetheless, these YOLACT-like methods are not customized to medical image analysis, which typically involves limited training data.

B. Attention mechanisms

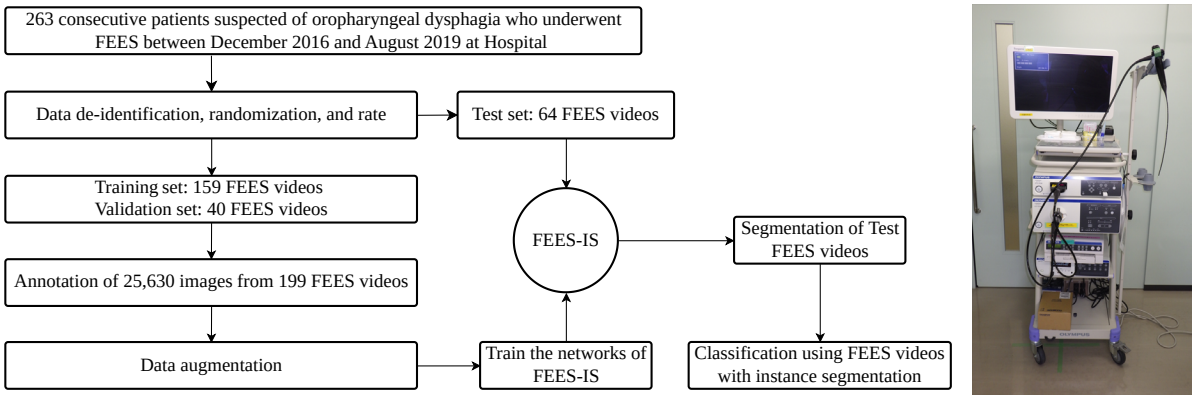
Attention mechanisms have been utilized in instance segmentation to improve object proposals and instance masks. Vaswani et al. [12] first introduced attention mechanisms to selectively weight the contributions of various feature maps or spatial locations during mask prediction. Recent studies have explored the use of attention mechanisms in medical image processing for instance segmentation. For example, Ren et al. [13] developed an RNN architecture with an attention mechanism that improved upon earlier formulations and achieved state-of-the-art results on non-medical instance segmentation datasets. CenterMask, developed by Lee et al. [14], utilizes an attention mechanism and an anchor-free box prediction scheme to achieve state-of-the-art performance in real-time non-medical instance segmentation.

However, attention mechanisms can be computationally expensive, which may limit their ability to achieve real-time speeds in medical image tasks with limited training data. The computation and memory resources required to compute attention weights for each element in a set can be significant. Moreover, most instance segmentation models for medical image processing are based on pre-trained models that were initially trained on non-medical image datasets, such as ImageNet. These pre-trained models are then fine-tuned on the medical image dataset for the specific task. However, this pre-training on non-medical image datasets may not be sufficient for learning effective attention patterns for medical images [15]. Therefore, attention mechanisms may not work effectively in these scenarios, and alternative methods such as transfer learning or data augmentation may need to be employed to improve performance. Obtaining labeled data can be challenging and time-consuming in medical image tasks, further complicating the use of attention mechanisms for real-time applications. Therefore, while attention mechanisms can benefit the performance of medical image processing tasks with limited training data, their computational and data requirements may make them unsuitable for real-time applications.

III. MATERIALS AND METHODOLOGY

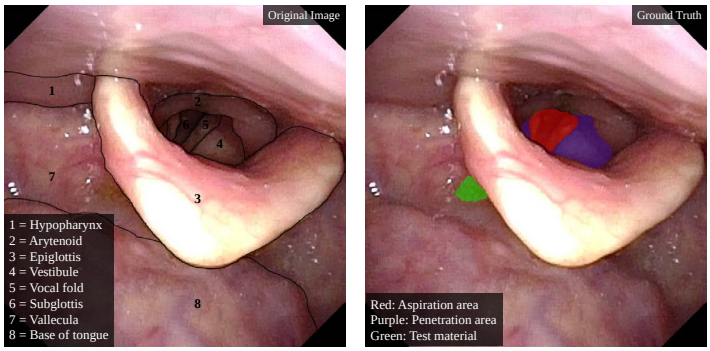
A. Materials

Fig. 1(a) depicts the flowchart of this study. Fig. 1(b) illustrates the use of a 2.6 mm diameter laryngeal flexible endoscope for FEES. This study was conducted at Fukushima Medical University Hospital on a consecutive series of patients who were suspected of having oropharyngeal dysphagia and who underwent Flexible Endoscopic Evaluation of Swallowing (FEES) between December 2016 and August 2019. Prior to the commencement of the examination, written informed consent was obtained from all participants. Aspiration and penetration are common complications associated with oropharyngeal dysphagia, and detecting subglottis regions in swallowing endoscopy is challenging. The act of aspiration is characterized by the downward passage of food beyond the vocal folds into the subglottis, whereas penetration occurs when food enters the laryngeal vestibule without entering the subglottis. In this study, the aspiration area was defined as the region comprising the vocal fold and subglottis. The original video recording obtained during FEES facilitates the evaluation of various anatomical structures. The segmentation of the FEES video highlights the aspiration area (in red), the penetration area (in purple), and the test bolus (in green). The FEES-IS system can recognize "none" and "white-out" images and records the starting point of "white-out" to remind laryngologists. In the following subsection, we present the development process of the FEES-IS system, which was designed to tackle the aforementioned multi-class segmentation task.



(a) (b)

Fig. 1. (*a*) Flowchart of this study. (*b*) FEES was performed using a laryngeal flexible endoscope with a diameter of 2.6 mm (ENF-V3, O, OLYMPUS Corp., Tokyo, Japan).



(a) (b)

Fig. 2. (*a*) shows the various anatomical structures that can be observed in a FEES video, such as the hypopharynx, arytenoid, epiglottis, vestibule, vocal fold, subglottis, vallecula, and base of the tongue. (*b*) displays annotations of video frames indicating the aspiration area in red (created by the vocal fold and subglottis landmarks), the penetration area in purple (created by the laryngeal vestibule), and the annotation of jelly as a test bolus in green.

B. Methodology

Fig. 3 illustrates the architecture of FEES-IS, which is composed of various essential modules such as the feature backbone, feature pyramid, prediction head, fast non-maximum suppression (fast NMS), protoNet, crop, and threshold.

The feature backbone module adopts ResNet-50 [16] with deformable convolutional layers, which have been proven to improve instance segmentation performance by adapting to object boundaries, handling scale variations, and capturing spatial context more effectively than conventional convolutional layers. These deformable convolutional layers have achieved state-of-the-art results on several benchmark datasets and are a promising technique for enhancing instance segmentation models [17].

The feature pyramid is a modified version of YOLACT++'s feature pyramid, which generates feature maps at different scales through lateral and top-down connections and employs deformable convolutional layers. However, instead of integrating spatial attention modules into the feature pyramid, FEES-IS incorporates proposed sparse attention before and after the feature pyramid. The sparse attention method first divides the feature maps of the feature backbone and feature pyramid into 8 groups, and then further splits each group's feature maps into K branches (K equal to the number of channel/the number of groups). These branches are randomly used to build 4 streams, as shown in Fig. 3. One of the streams passes through a channel attention module, which learns channel attention weights using a global average pooling layer and a set of trainable weights and biases. The attention weights are then applied to the original feature maps for obtaining the channel-wise attended feature maps. One of the streams passes through a spatial attention module, which learns spatial attention weights using a group normalization layer and another set of trainable weights and biases. The attention weights are then applied to the normalized feature maps for obtaining the spatially attended feature maps. Two of the streams in each group are passed without any attention operation to address the issue of overfitting due to the complexity introduced by attention mechanisms. Feature maps of these two streams are concatenated with the channel and spatial attention maps, and channel shuffle operators [18] are used for the incorporation of cross-group information flow along the channel dimension.

The prediction head predicts the coefficients of each instance mask, as well as its class probability and location. The

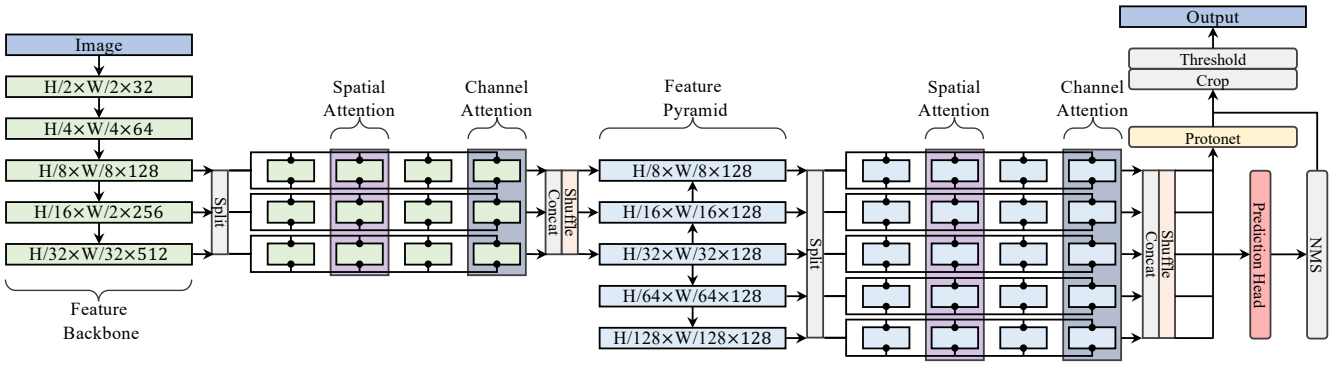


Fig. 3. FEES-IS architecture.

mask encoding process reduces the dimensionality of the mask data, making it computationally efficient.

Fast NMS optimizes the NMS [19] process used in object detection by sorting bounding boxes by their confidence scores and using a maximum weight matching algorithm to find the optimal matching in a bipartite graph. ProtoNet is a Fully Convolutional Network (FCN) that encodes instance masks into a condensed vector representation, generating high-quality masks. The crop module extracts the relevant feature map regions based on the predicted instance masks, and the thresholding step removes low-confidence instance masks, ensuring that only high-quality predictions are retained.

IV. Experiments

The proposed FEES-IS (FEES-IS-50) is compared with several existing methods for instance segmentation, including MaskRCNN [7], Yolact++ [9] with ResNet-50 (Yolact++-50) and ResNet-101 (Yolact++-101), and FEES-IS with Convolutional Block Attention Module [20] (FEES-IS-50-CBAM and FEES-IS-101-CBAM) and with Shuffle Attention [21] FEES-IS-50-SA and FEES-IS-101-SA). All the methods are evaluated with the same settings to ensure a fair comparison.

Datasets: FEES-IS was trained on 199 FEES videos and evaluated on 40 additional FEES videos (see Table I). The training and test datasets were matched with respect to the distribution of disease, age, height, and weight. The videos were anonymized, randomly selected, and evaluated by a panel of experienced laryngologists and dysphagia experts, each with over 15 years of experience in conducting FEES.

Training Methodology: The model loss is a weighted sum of the confidence loss (L_{conf}), segmentation loss (L_{seg}), and location loss (L_{loc}). L_{conf} is the softmax loss over multiple classes confidences (c), i.e.,

$$L_{conf} = \sum_{i \in Pos}^{N} x_{ij}^{p} \log\left(\hat{c}_i^p\right) + \sum_{i \in Neg} \log\left(\hat{c}_i^0\right) \tag{1}$$

where the weight term α is set to 1 by cross validation and

$$\hat{c}_i^p = \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)}. \tag{2}$$

Intersection over Union (IoU) is a commonly used evaluation metric for bounding box regression. It measures the overlap between the predicted and ground-truth bounding boxes by calculating the ratio of their intersection to their union. Specifically, given a ground-truth bounding box $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ and a predicted box $B = (x, y, w, h)$, the IoU is calculated as

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}. \tag{3}$$

The L_{seg} is a loss function used to train a bounding box regression model. Its aim is to minimize the distance between the predicted and ground-truth bounding boxes while maximizing their IoU similarity. The loss is defined by the following equation:

$$L_{seg} = 1 - IoU + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{gt}\right)}{c^2} + \frac{v^2}{(1 - IoU) + v}. \tag{4}$$

Here $\mathbf{b}$ and $\mathbf{b}^{gt}$ are the central points of the predicted (B) and ground-truth (B^{gt}) bounding boxes, respectively. The Euclidean distance between these central points is denoted by $\rho(\cdot)$. c is the diagonal length of the smallest enclosing box that covers both the predicted and ground-truth bounding boxes. v measures the consistency of the ratio between the width and height of the bounding box. It is defined as:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2. \tag{5}$$

Here, w^{gt} and h^{gt} denote the width and height of $\mathbf{b}^{gt}$, and w and h indicate the width and height of $\mathbf{b}$. The $\arctan$ function is used to ensure that the range of v is between 0 and 1. The L_{seg} loss is used during training to adjust the parameters of the bounding box regression model so that it produces accurate predictions that have high IoU values with $\mathbf{b}^{gt}$

L_{loc} is the combination of Polar IoU Loss [10] and Cartesian IOU loss [22], which can be defined as Equ 6.

The model's predicted polygon is represented by N vertices, where x_n and y_n are the x and y coordinates, respectively, of the n^{th} vertex. The center of the polygon is denoted by x_c and y_c, which are the x and y coordinates, respectively.

TABLE I
CHARACTERISTICS OF THE PATIENTS.

	Number	Oral intake	Dysphagia	Age (mean)	Age (median)	Height (mean)	Height (median)	Weight (mean)	Weight (median)
Training data	199	94	178	62.75	71.00	154.31	161.20	49.09	48.40
Test data	40	18	37	64.28	72.50	152.44	160.00	47.13	47.40

$$L_{loc} = \log\left(\frac{(x_1^{\max}y_2^{\max} - y_1^{\max}x_2^{\max}) + (x_2^{\max}y_3^{\max} - y_2^{\max}x_3^{\max}) \cdots + (x_n^{\max}y_1^{\max} - y_n^{\max}x_1^{\max})}{(x_1^{\min}y_2^{\min} - y_1^{\min}x_2^{\min}) + (x_2^{\min}y_3^{\min} - y_2^{\min}x_3^{\min}) \cdots + (x_n^{\min}y_1^{\min} - y_n^{\min}x_1^{\min})}\right) \tag{6}$$

TABLE II
AP FOR AND FPS DIFFERENT IOU THRESHOLDS ON THE FEES DATASET.

Method	AP_{50}	AP_{55}	AP_{60}	AP_{65}	AP_{70}	AP_{75}	AP_{80}	AP_{85}	AP_{90}	AP_{95}	mAP	FPS
MaskRCNN	80.48	77.76	73.71	69.32	68.11	65.32	60.38	51.82	39.20	34.70	62.08	6.28
Insta-YOLO	72.39	69.33	65.51	63.62	59.82	52.38	45.22	38.78	30.51	25.32	52.29	42.65
Yolact++-50	72.91	70.30	67.58	64.96	61.42	54.96	48.15	39.93	31.31	25.38	53.69	35.70
Yolact++-101	74.24	70.98	68.48	64.87	60.68	54.96	48.97	40.84	32.43	25.65	54.21	32.53
FEES-IS-50-CBAM	79.15	77.27	74.62	70.73	67.95	61.42	48.30	43.97	31.36	25.97	58.07	28.48
FEES-IS-101-CBAM	78.63	76.79	73.26	71.03	65.51	63.07	52.40	43.87	27.01	25.03	57.66	23.45
FEES-IS-50-SA	78.82	76.81	71.73	69.05	66.48	60.82	54.51	42.54	38.52	29.12	58.84	30.36
FEES-IS-101-SA	78.89	76.29	72.99	69.38	66.76	60.07	53.55	44.58	36.77	29.74	58.90	27.90
FEES-IS	80.06	76.19	73.27	70.91	67.33	62.18	58.60	52.08	41.74	31.99	61.64	33.18
FEES-IS-101	79.53	76.76	74.21	69.51	66.70	61.36	54.28	46.73	37.85	30.79	59.77	31.23
MaskRCNN-ori	79.91	76.98	73.45	69.67	65.91	61.72	56.25	47.92	35.20	30.35	59.74	6.28
Insta-YOLO-ori	73.59	70.41	67.85	61.63	57.39	52.27	45.75	37.25	29.56	25.36	52.11	42.65
Yolact++-101-ori	74.77	70.14	67.94	64.07	59.02	52.68	48.05	39.90	33.21	28.38	53.82	32.53
FEES-IS-ori	78.34	74.85	71.55	68.72	64.12	60.30	58.48	48.98	38.07	31.10	59.45	33.18

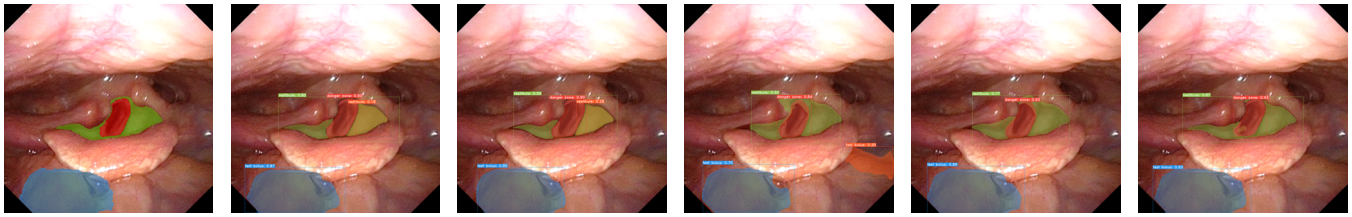
During training, a batch size of 8 was used on a single NVIDIA GeForce RTX 3090 graphics processing unit. The model was initialized with ImageNet pretrained weights and optimized using stochastic gradient descent (SGD) for 400k iterations. The initial learning rate was set to 10^{-3} and was reduced by a factor of 10 at iterations 200k, 250k, 300k, and 350k. A weight decay of 5×10^{-4} and a momentum of 0.9 were applied to prevent overfitting. The training data was augmented using techniques such as random cropping, color jittering, random horizontal flipping, random rotation, random scaling, Gaussian blur, and random noise, which are commonly used in object detection algorithms such as SSD [23]. The evaluationmetric was the average precision (AP) over a range of IoU thresholds from 0.5 to 0.95 [24].

V. RESULT

Table II presents the performance comparison of our proposed FEES-IS method with other instance segmentation approaches on the FEES dataset. We highlight the best result in pink and the second-best result in yellow. Our proposed method achieves competitive accuracy compared to YOLACT while maintaining real-time performance (over 30 FPS). Although FEES-IS has a slightly slower processing speed than Insta-YOLO and Yolact++-50, it significantly outperforms them in terms of mean AP(mAP) by 17.88% and 14.80%, respectively. FEES-IS achieves the second-best mAP, which is slightly lower than MaskRCNN by 0.70%, but at 4.28 times the speed of MaskRCNN. Moreover, FEES-IS achieves the best average precision at the AP_{85} and AP_{90}, respectively, and the second-best AP at the AP_{50}, AP_{65}, and AP_{80}, respectively, outperforming the widely used Yolact++-101 network by a significant margin. The results of our experiments indicate that implementing a deeper backbone (ResNet-101) does not necessarily result in higher performance. In terms of mAP, FEES-IS with ResNet-50 outperforms FEES-IS with ResNet-101 by 3.13%.

Table II also shows the performance comparison of the proposed loss function with the original loss function. The methods labeled as MaskRCNN-ori, Insta-YOLO-ori, and Yolact++-101-ori indicate these methods implement their original loss function, while FEES-IS-ori denotes FEES-IS with the same loss function as Yolact++. It is observed that the proposed loss function is necessary for FEES-IS and MaskRCNN, while it has little influence on Insta-YOLO-ori and Yolact++-101-ori. In terms of mAP, FEES-IS, Yolact++-101, MaskRCNN, Insta-YOLO outperform their original loss function version by 3.68%, 0.72%, 0.35%, and 3.92%, respectively.

Fig. 4 shows An example of the original video frame with (*a*) segmentation ground truth, and with qualitative results of (*b*) MaskRCNN, (*c*) Yolact++-101, (*d*) FEES-IS-50-CBAM, (*e*) FEES-IS-101, (*f*) FEES-IS. We can observe from the results that neural networks equipped with attention mechanisms can effectively identify the Vestibule as a coherent region, thereby improving the interpretability of the segmentation results for dysphagia doctors. However, in Fig. 4(d), we observe that FEES-IS-50-CBAM misclassifies the background as test bolus, which we attribute to overfitting caused by the limited training data. Specifically, the network may struggle to classify pixels



(a) (b) (c) (d) (e) (f)

Fig. 4. An example of the original video frame with (*a*) segmentation ground truth, and with qualitative results of (*b*) MaskRCNN, (*c*) Yolact++-101, (*d*) FEES-IS-50-CBAM, (*e*) FEES-IS-101, (*f*) FEES-IS.

under varying illumination and viewing conditions, which can hinder its performance.

VI. Conclusion

This paper presents an algorithm for real-time instance segmentation in medical images, specifically for Flexible endoscopic evaluation of swallowing (FEES) videos. By incorporating sparse attention and a specialized loss function, the model is able to effectively prevent overfitting and improve accuracy. The results of the study show that FEES-IS achieves competitive accuracy compared to existing instance segmentation methods while maintaining real-time performance. The dataset used for training and testing was extensive and collected from a large number of patients, providing a robust evaluation of the model's performance. Overall, the FEES-IS model has the potential to enhance medical image analysis by providing more detailed and accurate information for identifying and tracking individual objects in medical images.

References

[1] C. Gordon, R. L. Hewer, and D. T. Wade, "Dysphagia in acute stroke." *Br Med J (Clin Res Ed)*, vol. 295, no. 6595, pp. 411–414, 1987.

[2] G. Mann, G. J. Hankey, and D. Cameron, "Swallowing function after stroke: prognosis and prognostic factors at 6 months," *Stroke*, vol. 30, no. 4, pp. 744–748, 1999.

[3] D. Kidd, J. Lawson, R. Nesbitt, and J. MacMahon, "Aspiration in acute stroke: a clinical study with videofluoroscopy," *QJM: An International Journal of Medicine*, vol. 86, no. 12, pp. 825–829, 1993.

[4] M. Imaizumi and S. Murono, "Will levels of experience of examiners affect the diet provided for patients with swallowing impairment?" *Auris Nasus Larynx*, 2023.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[6] W. Weng, M. Imaizumi, S. Murono, and X. Zhu, "Expert-level aspiration and penetration detection during flexible endoscopic evaluation of swallowing with artificial intelligence-assisted diagnosis," *Scientific Reports*, vol. 12, no. 1, p. 21689, 2022.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[8] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[9] J. C. A. Cerón, L. Chang, G. O. Ruiz, and S. Ali, "Assessing yolact++ for real time and robust instance segmentation of medical instruments in endoscopic procedures," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1824–1827.

[10] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 193–12 202.

[11] H. Liu, R. A. R. Soto, F. Xiao, and Y. J. Lee, "Yolactedge: Real-time instance segmentation on the edge," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9579–9585.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6656–6664.

[14] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 906–13 915.

[15] M. R. Hosseinzadeh Taher, F. Haghighi, R. Feng, M. B. Gotway, and J. Liang, "A systematic benchmarking analysis of transfer learning for medical image analysis," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3*. Springer, 2021, pp. 3–13.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.

[18] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[21] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2235–2239.

[22] E. Mohamed, A. Shaker, A. El-Sallab, and M. Hadhoud, "Insta-yolo: Real-time instance segmentation," *arXiv preprint arXiv:2102.06777*, 2021.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[24] L. Chen, M. Strauch, and D. Merhof, "Instance segmentation of biomedical images with an object-aware embedding learned with local constraints," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 451–459.

Identification of Children with ADHD from EEG Signals Based on Entropy Measures and Support Vector Machine

Md. Maniruzzaman
School of Computer Science and Engineering
The University of Aizu
Aizuwakamatsu 965-8580, Japan

Md. Al Mehedi Hasan
Dept. of Computer Science & Engineering
Rajshahi University of Engineering and Technology
Rajshahi-6204, Bangladesh

Taro Suzuki
School of Computer Science and Engineering
The University of Aizu
Aizuwakamatsu 965-8580, Japan

Jungpil Shin
School of Computer Science and Engineering
The University of Aizu
Aizuwakamatsu 965-8580, Japan

***Abstract*—Attention deficit hyperactivity disorder (ADHD) is one of the major psychiatric and neurodevelopment disorders that affects 11% of children worldwide. Moreover, the prevalence of ADHD has rapidly increased over time worldwide. According to DSM-V, three types of symptoms such as inattention, hyperactivity, and combined type (inattention with hyperactivity). So, it is necessary to use a simple, non-invasive, and automatic detection system for identifying children with ADHD at an early stage. The objective of this study was to propose a machine learning (ML)-based ADHD-combine type (ADHD-CT) detection system from electroencephalogram (EEG) signals. EEG signals were recorded from nineteen ADHD-CT children and fourteen healthy children. We extracted five entropy-based features such as approximate-based entropy, Shanon-based entropy, permutation-based entropy, sample-based entropy, and singular value decomposition (SVD)-based entropy from each signal. The subset of the most informative and discriminative features was selected for ADHD-CT using sequential forward floating selection (SFFS). Following that, support vector machine (SVM) was implemented with leave-one-out cross-validation for the identification of ADHD-CT children and assessed its performances based on classification accuracy. Our results illustrated that SVM with polynomial kernel provided 96.87% classification accuracy to discriminate children as ADHD-CT and healthy children. Our findings showed that our proposed system can be used to detect children with ADHD-CT.**

***Index Terms*—ADHD, Identification, EEG Signals, Entropy Measure, Feature Selection, Support Vector Machine**

I. Introduction

Attention deficit hyperactivity disorder (ADHD) is one of the most common neurobehavioral disorders. Globally, 5% of children are affected by ADHD [1]. Approximately 11% of children between the ages of 4 and 17 years are affected by ADHD in the USA [2]. ADHD is mainly diagnosed in children aged 6-12 years and can last until adulthood [3]. Children with ADHD have various challenges, including lack of attention, carelessness, impulsivity, hyperactivity, and combine types (inattention and hyperactivity) [4]. As a result, children have suffered from severe complications, including depression, anxiety, and attempts to commit suicide [5]. The prevalence of males who have ADHD is comparatively higher than females [6]. This figure has been rapidly increasing day by day. So, diagnosing children with ADHD at an early stage is still a very important research problem. This study mainly focuses on detecting ADHD-CT children based on their electroencephalogram (EEG) signals.

Various EEG-based studies concern children with ADHD and they summarized various features like statistical features, frequency-domain, and so on to discriminate ADHD patients from healthy control [7]–[11]. Nowadays, EEG-based neuroimaging is used in order to diagnose numerous various disorders, including ADHD. It is the most popular due to its portability and its ability to image human brains. Although a lot of existing studies has been proposed an ADHD detection system based on EEG signals, there is still scope to improve the efficiency and accuracy of ADHD detection systems. In order to detect ADHD, it is essential to extract features from EEG signals. Various existing studies have attempted to identify potential features or biomarkers of ADHD, which are extracted from EEG signals [7]–[11]. In this study, five entropy-based features such as approximate-based entropy (Approx-Ent), Shanon-based entropy (Shan-Ent), permutation-based entropy (Per-Ent), sample-based entropy (Sam-Ent), and singular value decomposition-based entropy (SVD-Ent) were extracted from EEG signals. Previously, various linear and non-linear classifiers, including support vector machine (SVM), linear discriminant analysis, neural networks, and so on, were widely used to differentiate ADHD subjects from healthy controls [3], [8]–[10], [12], [13]. In this work, we implemented to differentiate ADHD subjects from healthy controls. The contributions of this study are listed as follows:

- Extract five entropy-based features such as Approx-Ent,

Shan-Ent, Per-Ent, Sam-Ent, and SVD-Ent from each channel.
- Determine the subset of the most discriminative and informative features using textcolorbluesequential floating forward selection (SFFS)-based method.
- Implement an SVM-based classifier to distinguish ADHD-CT children from healthy classify and evaluate its performance using classification accuracy.

II. Related Work

Tenev et al. [14] introduced an ML-based classifier for the discrimination of adults with ADHD and healthy controls by analyzing EEG power spectra with four measurement conditions: (i) eyes open; (ii) eyes closed, (iii) visual continuous performance test (VCPT), and (iv) emotional continuous performance test (ECPT). Whereas 19 EEGS signals were recorded from 117 subjects (ADHD: 67 vs. HC: 50) aged 18 to 50 years. They performed spectral analysis using a fast Fourier transform and computed delta, theta, alpha, and beta frequency bands from each signal which were used as input features in the prediction model for classification. They implemented a forward selection scheme with an SVM-based classifier over four different conditions to identify the best combination of the most relevant features. After identifying the combination of the most relevant features for each dataset, they also implemented SVM with a 10-fold cross-validation protocol and obtained a classification accuracy of 82.3%.

Khoshnoud et al. [15] also recorded 19 EEG signals with 256 sampling rate frequency and 16-bit resolution from 12 children with ADHD and 12 healthy controls during eyes-closed resting. They extracted 8 types of features (4 frequency band powers+4 non-linear) from each EEG signal and used principal component analysis (PCA) for dimension reduction. SVM and neural networks (NN) with a four-fold cross-validation protocol were employed for the discrimination of ADHD from healthy controls and obtained 83.03% classification accuracy.

Kaur et al. [16] developed a diagnosis system for EEG signals using the phase space reconstruction method to discriminate adults with ADHD from healthy controls. They collected EEG samples from 47 ADHD and 50 healthy subjects using three conditions: (i) eyes open, eyes closed, and (iii) continuous performance test(CPT). They extracted various statistical features like maximum, minimum, mean, median, and so on were extracted from Euclidean distances using phase space reconstruction of signals. Moreover, they also extracted Katz's and Higuchi's fractal dimensions, power of scale-freeness in VG (PSVG), and graph index complexity (GIC)-based features from EEG signals. Two techniques, such as correlation and particle swarm optimization were implemented to select more efficient features. Then, five classification methods (neural dynamic classifier (NDC), SVM, EPNN, k-nearest neighbor (k-NN), and naive Bayes (NB)) were implemented and achieved the accuracy of 93.3% for eyes-open, 90.0% for eyes-closed, and 100% for CPT conditions by NDC.

Maniruzzaman et al. [8] also developed an ML-based system for identifying ADHD children based on their EEG signals. They used 121 subjects with 61 ADHD children and 60 healthy children. They extracted various features (morphological and time-domain) extracted from each EEG channel and then, the optimal features were identified using independent t-tests and least absolute shrinkage and selection operator (LASSO)-based methods. They trained four ML-based classifiers with LOOCV and achieved a 94.2% classification accuracy and 0.964 AUC by SVM.

Chow et al. [17] developed a novel ADHD detection technique based on Hjorth Mobility (HM) using EEG signals. They recorded 32 EEG signals from 30 ADHD and healthy children and extracted HM and theta beta ratio-based features. An Independent t-test was implemented to determine the most prominent channels using TBR and mobility (p¡0.05) and chose 12 channels, which were fed into a logistic regression-based classification model and obtained classification accuracy of 79.2%, recall of 79.6%, and AUC of 0.885, respectively.

Tor et al. [7] proposed an automated ADHD detection system using EEG signals. They recorded EEG signals using 10-20 international systems from 45 children with ADHD, 16 children with conduct disorder (CD), and 62 children have both ADHD and CD. They decomposed EEG signals using empirical mode decomposition (EMD) and discrete wavelet transform (DWT) methods. Relative energy and autoregressive modeling coefficients were computed from EEG signals. Eight types of non-linear features such as activity, entropies, fractal dimension, Hurst exponent, largest Lyapunov exponent, Lempel-Ziv complexity, Kolmogorov complexity, and recurrence qualitative analysis were extracted from each EEG signal. Then Z-score normalization was performed to standardize the data after feature extraction and implement an adaptive synthetic sampling (ADASYN)-based technique in order to make balance the dataset. The most prominent features were chosen using the SFS algorithm and significant features were determined with p-value¡0.05 which were fed into five ML-based classifiers (DT, k-NN, SVM, AB, and Bagged Tree) with K=3, 10-fold CV.

III. Materials and methods

A. Proposed Methodology

The proposed ADHD-CT detection system methodology is shown in Fig. 1. First, the raw EEG signal dataset was recorded from ADHD-CT and healthy children. Second, we preprocessed the extracted raw dataset, i.e., making it stationary and removing outliers. We implemented a p-th order difference equation to make the EEG dataset stationary and z-scores for removing outliers ($|\text{Z-score}| \leq 2$). Third, five entropy-based features such as Approx-Ent, Shan-Ent, Per-Ent, Sam-Ent, and SVD-Ent were extracted from each preprocessed EEG signal and then combined all features (40=5*8). The next step is to split the experimental dataset into training ((N-1) children) and test sets (One children). We selected the most relevant features using SFFS. Then, we developed a prediction model for ADHD-CT detection based on SVM

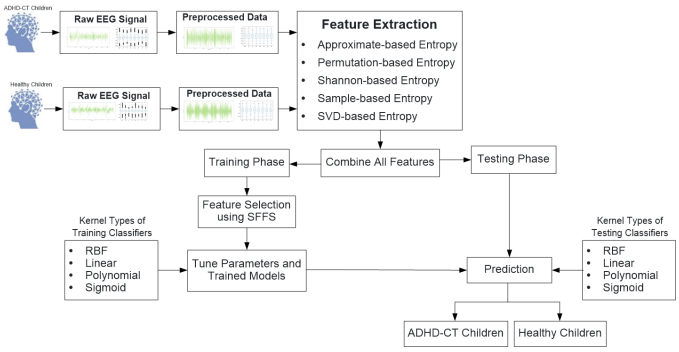


Fig. 1. Proposed Methodology of ADHD-CT detection system.

with four kernels (RBF, linear, polynomial, and sigmoid) and tuned SVM hyperparameters. Using the test set, this trained model was used to predict ADHD-CT and healthy children.

B. Dataset Acquisitions and EEG Recordings

We utilized a dataset, which was publicly available [18], [19]. The dataset had two groups of subjects aged 6-10 years. The first group was ADHD combined type (ADHD-CT), which consisted of only 19 boys with an average age of 8.0± 0.3 years. The second group was the healthy children, which consisted of 14 boys (male: 31 and female: 16) with an average age of 8.10± 0.48 years. A psychiatrist and psychologist were recruited to diagnose ADHD subjects. All subjects needed to meet the criteria of ICT-10 Hyperkinetic disorder [20] or the criteria of DSM-IV [4] of ADHD-CT. EEGs were recorded with a sampling frequency of 256 Hz from 8 channels (Fp1, Fp2, C3, C4, T3, T4, O1, and O2) based on 10-20 standard international systems (See in Fig. 2). The data were collected with resting-state conditions: (i) closed eyes (EC), and (ii) opened eyes (EO). Before recording the EEG session, written consent was obtained from each subject and his parents.

C. Preprocessing

In this work, raw recorded EEG datasets were preprocessed using Python to reduce the computational demand and noise effects on the signal. The raw dataset was filtered using a low bandpass of 0.05Hz, and a high bandpass of 80 Hz. Moreover, we used the p-th order difference equation to make stationary signals. Furthermore, we also the outliers of each channel using Z-score. We included the recordings whose values were lies between (mean -2SD) and (mean + 2SD).

D. Feature Extraction

In this work, we have extracted four types of entropy-based features from each signal. These four entropy-based features were approximate-based entropy (Approx.-Ent), Shanon-based entropy (Shan-Ent), Permutation-based entropy (Per-Ent), Sample-based Entropy (Sam-Ent), and singular value decomposition-based entropy (SVD-Ent). The calculation procedure for these features is described as follows:

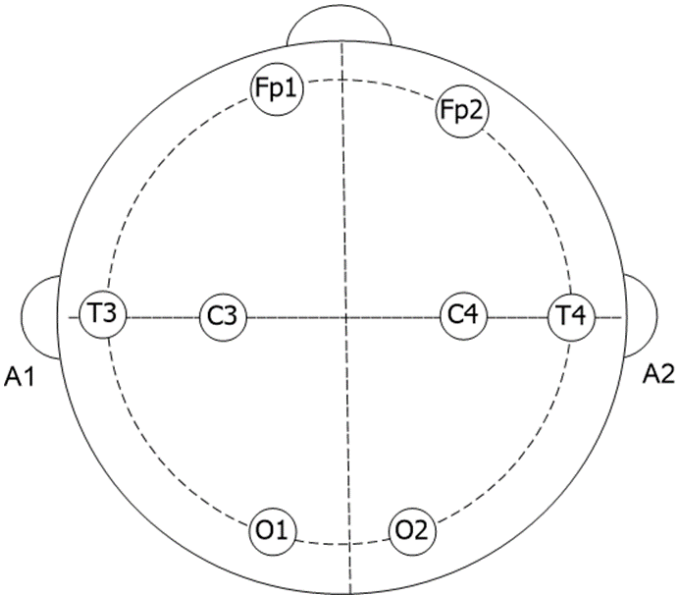


Fig. 2. Position of channels/electrode used in this experiment.

1) Approximate Entropy: Approximate Entropy (Approx-Ent): Approx-Ent is a method that is used to quantify the regularity or predictability of a time series data [21]. The formula for calculating ApEn is as follows:

$$\text{Approx-Ent} = \varphi(m+1,\ r) - \ \varphi\ (m,\ r) \tag{1}$$

Here, $\varphi(m+1,\ r)$is the conditional probability that two similar sequences of length "m" remain close within a tolerance r; m is the pattern length; r is the tolerance or similarity criterion. In this study, we set the value of r as 0.20*SD (x).

2) Shannon Entropy: Shannon Entropy is used to measure the uncertainty in a dataset [22]. It is an information-based entropy and is mathematically defined as:

$$\text{Sha-Ent} = \sum_{i} \mathrm{p}(\mathrm{x_i})\ \log_2 \mathrm{p}(\mathrm{x_i}) \tag{2}$$

3) Permutation Entropy: Permutation Entropy (Per-Ent) is a measure used to quantify the complexity or randomness of time series data. It is based on the concept of ordinal patterns, which represent the orderings of values in a time series. Per-Ent was first introduced in 2002 by Bandt and Pompe [23] as a nonlinear complexity measure of time series data. Per-Ent is mathematically defined as follows

$$\text{Per-Ent (m)} = -\sum \mathrm{p}(\pi)\log_2(\pi) \tag{3}$$

Here, m is the length of the pattern; p is the probability of each unique ordinal pattern, and the summation is taken over all unique patterns. The value of Per-Ent ranged from 0 to $\log_2(m!)$.

4) Sample Entropy: Sam-Ent is a modification of Aprrox-Ent that is used to assess the complexity of time-series data [24] [Richman et al., 2000]. Sam-Ent is mathematically defined as

$$\text{Sam-Ent (x, m, r)}=-\log\frac{C(m+1,\ r)}{C(m,\ r)} \tag{4}$$

Here, m represents the embedding dimension; r is the tolerance point; C (m+1, r) and C (m, r) represent the no. of emended vectors of length (m+1) and m.

5) SVD Entropy: SVD-Ent is also a method to assess the complexity of time-series data based on its singular value spectrum. It is also used as a dimension-reduction method. The SVD-Ent of a signal x is defined as:

$$\text{SVD-Ent}=-\sum_{i=1}^{M}\bar{\sigma_i}\log_2(\bar{\sigma_i}) \tag{5}$$

Here, M represents the no. of singular values of matrix X; and σ_i $(i = 1,\ 2, ..,\ M)$ are the i^{th} normalized singular values of Matrix X.

E. Feature Normalization

Min-Max normalization is a data preprocessing method widely used to transform continuous data into a specific range (0-1). It is computed using the following formula:

$$\text{Z}=\frac{\text{X}-X_{min}}{X_{max}-X_{min}} \tag{6}$$

Where, X_{max} and X_{min} are the maximum and the minimum values of the feature, respectively. The value of Z ranged from 0 to 1.

F. Feature Selection

Feature selection (FS) is a technique widely used in statistics and machine learning to choose a subset of features to build a predictive model. The FS is designed to increase the performance of predictive models, reduce overfitting, and enhance interpretability and understanding [1-3]. By selecting the most discriminative and informative features, the FS-based technique can improve the computational efficiency and accuracy of predictive models. This study adopted the SFS-based method [25] to select the most informative and discriminative features for ADHD-CT, which were used in SVM for discriminating ADHD-CT children from healthy children.

G. Classification using SVM

Support vector machine (SVM) is one of the most popular supervised techniques [26] that is widely used in various fields. In this study, we used to determine an optimal line (called a hyperplane) that can easily differentiate ADHD-CT children from healthy children by solving the constraints:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,\ x_j) \tag{7}$$

Subject to

$$\sum_{i=1}^{n}{y_i}^T\alpha_i=1, 0\le\alpha_i\le C, i=1, ..., n\ \&\ \forall\ i=1, 2, 3, ..., n \tag{8}$$

The final discriminate function takes the following form:

$$\text{f(x)}=\sum_{\text{i=1}}^{\text{n}}\alpha_{\text{i}}\text{K}(\text{x}_{\text{i}},\ \text{x}_{\text{j}})+\text{b} \tag{9}$$

In this study, we implemented SVM with four kernels such as radial basis function (rbf), liner kernel, polynomial kernel, and sigmoid kernel. The computational formula of these kernel functions is defined as follows:

$$\text{RBF Kernel}: K(x_i,\ x_j)=\exp(\text{-}\gamma\|x_i\text{-}x_j\|^2);\gamma>0 \tag{10}$$

$$\text{Linear Kernel}: K(x_i,\ x_j)=x_i.x_j \tag{11}$$

$$\text{Polynomial Kernel}: K(x_i,\ x_j)=(1+x_i.\ x_i)^d;d>1 \tag{12}$$

$$\text{Sigmoid Kernel}: K(x_i,\ x_j)=tanh\ (kx_i.x_j+c) \tag{13}$$

In the case of this work, we applied the following steps:

Step 1: Divide the dataset into two phases: a training phase and a testing phase. In each iteration, one subject is used in the testing phase and the remaining (N-1) subjects are used for the training phase.

Step 2: The hyperparameters of each kernel parameter with cost (C) are tuned in the training phase. This involves a grid search method to obtain the optimal values of these parameters that yield the highest classification accuracy.

Step 3: SVM model is trained using four kernels with LOOCV on the training set.

Step 4: The trained SVM model is then utilized to predict the class label (ADHD-CT vs. Healthy Children) on the test set. Moreover, the probability of each predicted class label.

Step 5: Repeat Step 1 to Step 4 into N times (here N=32).

Step 6: Finally, compute the classification accuracy.

IV. Experimental Set Up and Performance Evaluations Metrics

In this work, we performed all experimental analyses using both the R-programming language and Python. The operating system used Windows 10 version 21H1 (build 19043.1151) 64-bit. In terms of hardware, an Intel (R) Core (TM) i5-10400 processor with 16 GB of RAM was used. We used the LOOCV protocol during performing the SVM model. The efficiency of the SVM model was assessed using accuracy (ACC) which is computed using the following formula:

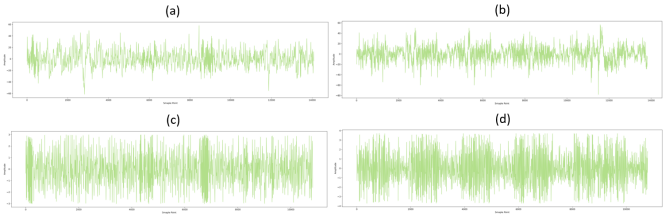


Fig. 3. Raw and preprocessed dataset of ADHD-CT and healthy children: (a)-(b) Raw EEG signals for ADHD-CT and healthy children; (c)-(d) preprocessed datasets of ADHD-CT and healthy children.

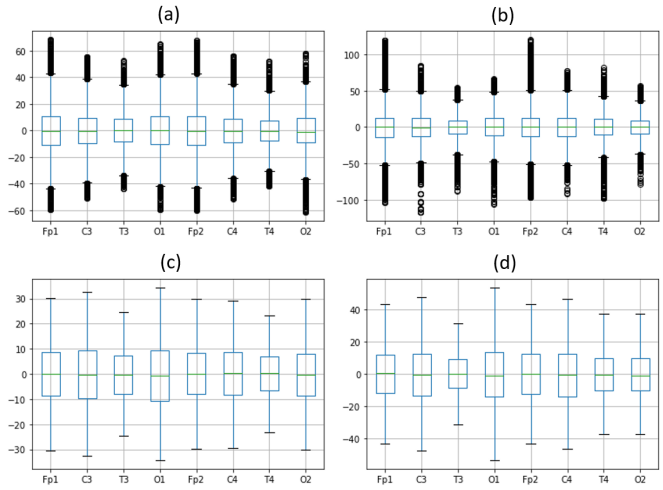


Fig. 4. Boxplot of one raw and preprocessed dataset for ADHD-CT and healthy children: (a)-(b) Raw EEG signals for ADHD-CT and healthy children; (c)-(d) preprocessed datasets of ADHD-CT and healthy children.

$$\mathrm{ACC}(\%) = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100 \qquad (14)$$

Here, TP: True positive; TN: True negative; FN: False positive (FP); and FN: False negative (FN).

V. Results

A. Statistical Analysis of EEG Dataset

The recorded EEG signals of ADHD-CT healthy children and healthy children are presented in Fig.3a and Fig.3b. Also, the boxplot of recorded EEG signals for ADHD-CT children and healthy children is illustrated in Fig.4a and Fig.4b. We noticed that the recorded raw EEG signals were non-stationary and contained outliers. In order to make EEG signals stationary, we used the 12th-order difference equation of the recorded EEG dataset to make the stationary EEG dataset. Moreover, we computed the z-score of each channel/signal, and then, we removed the rows (outliers), which had at least one z-score with an absolute value of more than 2. The preprocessed dataset of one ADHD-CT and healthy children is shown in Fig.3c-Fig.3d and Fig.4c-Fig.4d, respectively. These preprocessed were used for feature extraction and classification of ADHD-CT and healthy children.

TABLE I
Sets hyperparameters during training SVM model

Kernel Types	Set values of hyper-parameters
Radial	Cost (C): 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, length-scale= (1 to 5), alpha= (0.04, 0.05, 0.06), Sigma (σ): 0.00001, 0.0001, 0.001, 0.01, 0.1, 1
Linear	Cost (C): 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000
Polynomial	Cost (C): 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, Degree= 2, 3, 4
Sigmoid	Cost (C): 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000

B. Feature Extraction

In this work, we extracted five entropy-based features from each channel. As a result, 40 features (5x8) were extracted from 8 channels, which were used in SFFS with the SVM classification model for discriminating ADHD-CT children from healthy children.

C. Performance of SVM -based Classification Model

In this work, we implemented a support vector machine (SVM) for the classification of ADHD-CT children and healthy children. There were several kernel functions, used for training the SVM model. In this work, we used four kernels: RBF, linear, polynomial, and sigmoid. We performed three steps to conduct this experiment. First, the datasets were partitioned into two phases: the training phase and the test phase. We took one subject as the test phase and the remaining (32-1) =31 subjects were taken as the training phase. We selected the subset of relevant features using SFFS, which were used to train the SVM-based model and optimized the hyper-parameters of these four kernels on the basis of classification accuracy during the training phase. During the training model, we set various parameters (See in Table I) and chose the hyper-parameters, at which points yield the highest classification accuracy.

After optimizing hyper-parameters, one subject (test set) was fed into a trained SVM model to predict ADHD-CT children. These procedures were repeated in 32 trails and we computed the predicted class label of each trail. Finally, we compared these predicted class labels with actual class labels and then computed the classification accuracy of SVM over four kernels, which are presented in Table II.

TABLE II
Classification accuracy (in %) of seven classifiers of individual task for optimal features

Kernel Types	Conditions	
	Eyes Open	Eyes Closed
RBF	93.75	81.25
Linear	71.88	65.62
Polynomial	96.87	87.50
Sigmoid	68.75	68.65

We observed that SVM with a polynomial (degree=4) kernel provided the highest classification accuracy. Especially, SVM with polynomial kernel obtained a classification accuracy of 96.87% and 87.50% for eyes open and eyes close condition. Finally, it may be concluded that SVM with a polynomial

kernel may have more capable of differentiating ADHD-CT children and healthy children.

VI. Conclusions

This study developed an automated ADHD-CT detection system to help physicians diagnose ADHD-CT children at an early stage. Five entropy-based parameters were extracted from EEG signals and the optimal discriminative features were identified using the SFFS algorithm, which was fed to the SVM algorithm. In contrast, SVM with a polynomial kernel provided the highest classification accuracy of 96.87%. This proposed system can be extended to identify at early stages of ADHD with other overlapping coexisting disorders.

References

[1] A. Parashar, N. Kalra, J. Singh, and R. K. Goyal, "Machine learning based framework for classification of children with adhd and healthy controls," *Intell. Autom. Soft Comput.*, vol. 28, no. 3, pp. 669–682, 2021.

[2] S. N. Visser, M. L. Danielson, R. H. Bitsko, J. R. Holbrook, M. D. Kogan, R. M. Ghandour, R. Perou, and S. J. Blumberg, "Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United states, 2003–2011," *J. Am. Acad. Child Adolesc. Psychiatry.*, vol. 53, no. 1, pp. 34–46, 2014.

[3] M. Altınkaynak, N. Dolu, A. Güven, F. Pektaş, S. Özmen, E. Demirci, and M. İzzetoğlu, "Diagnosis of attention deficit hyperactivity disorder with combined time and frequency features," *Biocybern. Biomed. Eng.*, vol. 40, no. 3, pp. 927–937, 2020.

[4] D. L. Segal, "Diagnostic and statistical manual of mental disorders (dsm-iv-tr)," *The corsini encyclopedia of psychology*, pp. 1–3, 2010.

[5] A. Stickley, A. Koyanagi, V. Ruchkin, and Y. Kamio, "Attention-deficit/hyperactivity disorder symptoms and suicide ideation and attempts: Findings from the adult psychiatric morbidity survey 2007," *J. Affect. Disord.*, vol. 189, pp. 321–328, 2016.

[6] E. M. Derks, J. J. Hudziak, and D. I. Boomsma, "Why more boys than girls with adhd receive treatment: a study of dutch twins," *Twin Res. Hum. Genet.*, vol. 10, no. 5, pp. 765–770, 2007.

[7] H. T. Tor, C. P. Ooi, N. S. Lim-Ashworth, J. K. E. Wei, V. Jahmunah, S. L. Oh, U. R. Acharya, and D. S. S. Fung, "Automated detection of conduct disorder and attention deficit hyperactivity disorder using decomposition and nonlinear techniques with eeg signals," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105941, 2021.

[8] M. Maniruzzaman, J. Shin, M. A. M. Hasan, and A. Yasumura, "Efficient feature selection and machine learning based adhd detection using eeg signal," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 72, no. 3, pp. 5179–5195, 2022.

[9] A. Khaleghi, P. M. Birgani, M. F. Fooladi, and M. R. Mohammadi, "Applicable features of electroencephalogram for adhd diagnosis," *Research on Biomedical Engineering*, vol. 36, pp. 1–11, 2020.

[10] M. Maniruzzaman, M. A. M. Hasan, N. Asai, and J. Shin, "Optimal channels and features selection based adhd detection from eeg signal using statistical and machine learning techniques," *IEEE Access*, vol. 11, pp. 33 570–33 583, 2023.

[11] E. Pereda, M. García-Torres, B. Melián-Batista, S. Mañas, L. Méndez, and J. J. González, "The blessing of dimensionality: Feature selection outperforms functional connectivity-based feature transformation to classify adhd subjects from eeg patterns of phase synchronisation," *PloS one*, vol. 13, no. 8, p. e0201660, 2018.

[12] S. Kim, J. H. Baek, Y. J. Kwon, H. Y. Lee, J. H. Yoo, S.-h. Shim, and J. S. Kim, "Machine-learning-based diagnosis of drug-naive adult patients with attention-deficit hyperactivity disorder using mismatch negativity," *Translational Psychiatry*, vol. 11, no. 1, p. 484, 2021.

[13] E. Ghasemi, M. Ebrahimi, and E. Ebrahimie, "Machine learning models effectively distinguish attention-deficit/hyperactivity disorder using event-related potentials," *Cognitive Neurodynamics*, vol. 16, no. 6, pp. 1335–1349, 2022.

[14] A. Tenev, S. Markovska-Simoska, L. Kocarev, J. Pop-Jordanov, A. Müller, and G. Candrian, "Machine learning approach for classification of adhd adults," *Int. J. Psychophysiol.*, vol. 93, no. 1, pp. 162–166, 2014.

[15] S. Khoshnoud, M. A. Nazari, and M. Shamsi, "Functional brain dynamic analysis of adhd and control children using nonlinear dynamical features of eeg signals," *J. Integr. Neurosci.*, vol. 17, no. 1, pp. 17–30, 2018.

[16] S. Kaur, S. Singh, P. Arun, D. Kaur, and M. Bajaj, "Phase space reconstruction of eeg signals for classification of adhd and control adults," *Clinical EEG and neuroscience*, vol. 51, no. 2, pp. 102–113, 2020.

[17] J. C. Chow, C.-S. Ouyang, C.-T. Chiang, R.-C. Yang, R.-C. Wu, H.-C. Wu, and L.-C. Lin, "Novel method using hjorth mobility analysis for diagnosing attention-deficit hyperactivity disorder in girls," *Brain and Development*, vol. 41, no. 4, pp. 334–340, 2019.

[18] E. Pereda, "Controls. figshare. dataset," *https://doi.org/10.6084/m9.figshare.6810707.v1*, 2018.

[19] E. . Pereda, "Adhd. figshare. dataset," *https://doi.org/10.6084/m9.figshare.6812480.v1*, 2018.

[20] . W. H. Organization, "Who—icd-10 classification of mental and behavioural disorders," *Available from: http://www.who.int/substance$_a$buse/terminology/icd$_1$0/en*, 2002.

[21] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, "A regularity statistic for medical data analysis," *Journal of clinical monitoring*, vol. 7, pp. 335–345, 1991.

[22] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano, "Quantification of eeg irregularity by use of the entropy of the power spectrum," *Electroencephalography and clinical neurophysiology*, vol. 79, no. 3, pp. 204–210, 1991.

[23] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical review letters*, vol. 88, no. 17, p. 174102, 2002.

[24] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–49, 2000.

[25] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

Evaluating the effect of sparse convolutions on point cloud compression

Davi Lazzarotto, Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
davi.nachtigalllazzarotto@epfl.ch, touradj.ebrahimi@epfl.ch

***Abstract*—The use of point clouds as an imaging modality has been rapidly growing, motivating research on compression methods to enable efficient transmission and storage for many applications. While compression standards relying on conventional techniques such as planar projection and octree-based representation have been standardized by MPEG, recent research has demonstrated the potential of neural networks in achieving better rate-distortion performance for point cloud geometry coding. Early attempts in learning-based point cloud coding mostly relied on autoencoder architectures using dense convolutional layers, but the majority of recent research has shifted towards the use of sparse convolutions, which are applied only to occupied positions rather than the entire space. Since points are usually distributed on underlying surfaces rather than volumes, such operations allow to reduce the computational complexity required to compress and decompress point clouds. Moreover, recent solutions also achieve better compression efficiency, allocating fewer bits at similar levels of geometric distortion. However, it is not clear to which extent this gain in performance is due to the use of sparse convolutions, if any at all, since the architecture of the model is often modified. In this paper, we conduct an evaluation of the effect of replacing dense convolutions with sparse convolutions on the rate-distortion performance of the JPEG Pleno Point Cloud Verification Model. Results show that the use of sparse convolutions allows for an average BD-rate reduction of approximately 9% for both D1 and D2 PSNR metrics based on similar training procedures, with an even bigger reduction in point clouds featuring reduced point density.**

***Index Terms*—Point cloud compression, sparse convolutions**

I. Introduction

Immersive imaging modalities have grown in popularity in recent years due to their potential to offer more natural ways to interact with visual content. Applications such as virtual and augmented reality have become increasingly accessible and widely employed in numerous industries such as entertainment, education, training, and healthcare. These modalities enable users to explore and interact with virtual landscapes and objects in ways previously impossible, resulting in more impactful experiences. Because they allow for the correct collection and representation of 3D geometry in a scene, point clouds are a key component of many immersive imaging modalities. Point clouds provide a detailed and high-fidelity representation of the geometry that may be utilized for a range of applications such as autonomous navigation and visualization, by expressing the surface of objects as a collection of 3D points in space.

The vast amount of data needed to represent point clouds is however a major drawback for their use in mainstream applications. For that reason, effective solutions for compression have been heavily researched in recent years. Such efforts have led to the standardisation of two compression algorithms by MPEG, namely geometry-based point cloud compression (G-PCC) [1] and video-based point cloud compression (V-PCC) [2]. While the latter obtain planar projections of both point cloud attributes and geometry as color, depth, as well as occupancy maps and compresses them with conventional video codecs, G-PCC uses an octree to encode voxel occupancy and encodes color with either a region-adaptive hierarchical transform (RAHT) or a lifting transform.

Despite the usefulness demonstrated by conventional techniques, deep learning is receiving increased attention as an alternative approach for point cloud compression. Learning-based solutions have displayed even better rate-distortion performance for geometry data when compared to conventional techniques, allowing for better compression efficiency while maintaining a similar reconstruction quality. Early works in this direction were inspired by architectures previously used for the compression of 2D images, which relied on an autoencoder composed of convolutional layers. The input tensor is downsampled multiple times at the encoder, entropy coded using a probability distribution learned during training, and finally upsampled back to the original resolution at the decoder side. Initial efforts [3], [4] to adapt this algorithm for point clouds represented blocks as dense occupancy maps where all spatial positions are processed by dense 3D convolutions that operate similarly to their 2D counterparts.

However, these approaches fail to take into consideration the nature of most point clouds, which contain points sampled from an underlying surface and therefore occupy a small fraction of the space. Sparse convolutions, on the other hand, allow to better take advantage of these characteristics. In particular, sparse convolutions convolve a 3D kernel over a set of coordinates and apply the weights only at the occupied voxel positions from the input set of coordinates, differently from dense convolutions that convolve the kernel over all

This work was supported by the Swiss National Foundation for Scientific Research (SNSF) under the grant number 200020_207918.

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

indices of a three-dimensional grid and also consider all input positions to produce the input value. Sparse convolutions were leveraged in later compression methods [5], showing not only reduced computational complexity, but also increased rate-distortion performance, and have replaced dense convolutions on recent learning-based compression methods [6], [7].

The majority of recently proposed methods also include modifications to the architecture being used by previous models based on dense convolutions. Therefore, despite the rapid adoption of sparse convolutions as a de facto standard for learning-based voxelized point cloud compression, it is not possible to conclude to which extent the obtained improvements in rate-distortion performance are due to the use of sparse convolutions. The goal of this paper is therefore to assess the isolated impact on compression performance of replacing dense convolutions by these operations. The geometry-only pipeline of the JPEG Pleno Point Cloud verification model [8] is used as a baseline, and the evaluation is conducted using a test set composed of point clouds with different sparsity levels.

While early designs of the verification model also contained joint coding for both geometry and color, recent versions use a separate method for color coding. Moreover, the state of the art contains a much larger number of compression algorithms based both on dense and sparse convolutions for geometry-only coding than for color coding. For those reasons, sparse convolutions are evaluated only for geometry coding in this paper. Since the architecture of the baseline model is in many ways similar to a significant number of works in the state of the art [3]–[5], it is also considered that the results presented in this paper could be similar if other compression methods based on autoencoders were used.

II. Related work

Early works on point cloud geometry compression used an octree representation as data structure rather than a list of coordinates [9]. The octree became later prevalent with the addition of similar compression algorithms in widespread open-source libraries such as the Point Cloud Library (PCL) [10]. Later works explored pruning the octree, and then representing the leaf nodes as triangular primitives rather than singular points [11]. Both techniques were adopted in the geometry coding module of the G-PCC standard [1]. Other methods aimed to take advantage of the progress made in video compression during the last decades, and projected points onto multiple planes to represent point cloud geometry as two-dimensional maps. Such methods were later explored by the V-PCC standard [2], achieving high rate-distortion performance for dense point clouds, but struggling to effectively compress models with smaller point density.

Point cloud compression algorithms using neural networks were later introduced, with the first works [3], [4] mainly adapting a previous method designed for image compression [12] for three-dimensional representation. Several additional techniques were later studied [13] using previous works as baseline, such as entropy modeling using a hyperprior, adding

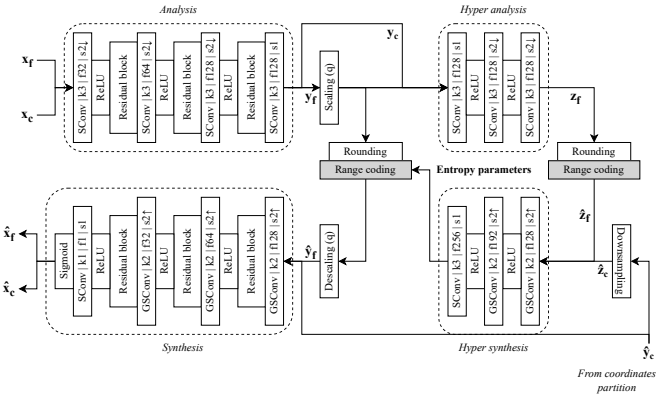


Fig. 1: Block diagram of the model using sparse convolutions. *SConv* denominates regular sparse convolutions and *GSConv* denominate generative sparse transposed convolutions. k denominates kernel size, f the number of features and s the convolution stride. The residual block was implemented with the same parameters as [8]. Scaling consists of the division of the latent features by the quantization step q, provided as an encoding parameter and added to the bitstream. Blocks highlighted in gray produce compressed representation added to the bitstream.

residual convolutional layers to the encoder and decoder, employing an adaptive threshold selection to translate output occupancy probability into voxels, as well as a sequential training method to allow for coding at different bitrates at reduced training time. Similar techniques were also employed by other authors [14], with results surpassing the rate-distortion performance by G-PCC for dense point cloud models. An autoregressive entropy coding model was also explored with comparable architecture [15], producing even better results that outperformed V-PCC for the evaluated test set. Other techniques, such as block prediction [16] and residual coding [17] explored further extension of similar techniques. Recently, the JPEG standardisation committee launched a call for proposals for learning-based point cloud coding, and a compression method based on dense convolutions was selected as the starting point known by the term verification model [8].

Sparse convolutions were first adopted [5] with an architecture similar to the same authors' previous work [14], with the addition of classification and pruning layers for progressive decoding. Another method, denominated as SparsePCGCv1 [6], improved upon the architecture by exploiting cross-scale and same-scale redundancies, allowing for both lossless and lossy compression. Moreover, GRASP-Net [7] proposed a heterogeneous architecture combining sparse convolutions with point-based MLP layers to recover fine details during decoding.

III. Evaluation conditions

The evaluated compression model is created by replacing all dense convolutions from the baseline model with sparse

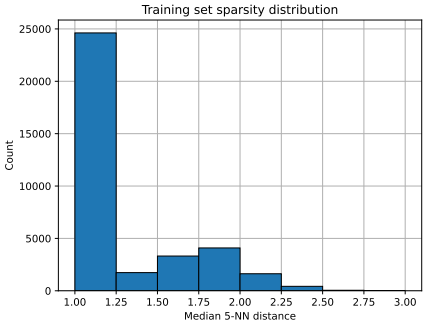


Fig. 2: Histogram of sparsity values of the training set

(a) *Annibal* from CfP original
Median 5-NN distance = 1.17

(b) *kinfudesk* from CfP supplemental
Median 5-NN distance = 2.26

(c) *Lausanne* from swissSURFACE3D — Median 5-NN distance = 3.36

Fig. 3: Point clouds from test set

convolutions, and its diagram block can be observed in Figure 1. The framework used in [8] includes downsampling prior to compression as a strategy to achieve lower bitrates, with learning-based upsampling being applied as a post-processing method. The present evaluation focuses only on the end-to-end autoencoder architecture, therefore setting the sampling factor to 1 and ignoring the advanced block upsampling network.

Prior to compression, the input point cloud is first partitioned into blocks. In the baseline model, the blocks are represented as dense tensors x of dimension $K \times K \times K \times N$, where K corresponds to the block size and N is the number of channels, which is set to 1 for geometry-only coding. The value at a given coordinate of x is set to 1 if the coordinate is present at the input block and to 0 otherwise. On the other hand, the sparse tensors used in the evaluated model are represented by a coordinate tensor x_c of dimensions $N_i \times 3$ set to the coordinates of the input point cloud block, and by a feature tensor x_f of size $N_i \times 1$ with all values set to 1, with N_i being the number of points in the input block.

The output of the analysis transform is a tensor y with its coordinates y_c being equivalent to x_c downsampled three times by a factor of 2. While the features y_f are encoded to the bitstream in a lossy manner by the range coding module and serve to build the input of the synthesis transform $\hat{y}$, the coordinates y_c are losslessly encoder and retrieved at the decoder side. The proposed compression algorithm downsamples the input point cloud geometry prior to block partition by a factor of 8 and compresses it using the lossless settings of the G-PCC codec. During decompression, the G-PCC bitstream is decoded and the obtained coordinates are partitioned into blocks in order to obtain $\hat{y}_c$, which is equivalent to y_c.

The synthesis transform takes $\hat{y}$ as input and passes it through three upsampling layers and residual blocks prior to a final sparse convolutional layer that produces a reconstructed tensor $\hat{x}$. Generative layers are employed when upsampling in order to generate new points, which, with kernel size 2, creates 8 output coordinates for each input coordinate corresponding to all possible positions that could have generated the point at the corresponding downsampling layer at the analysis transform.

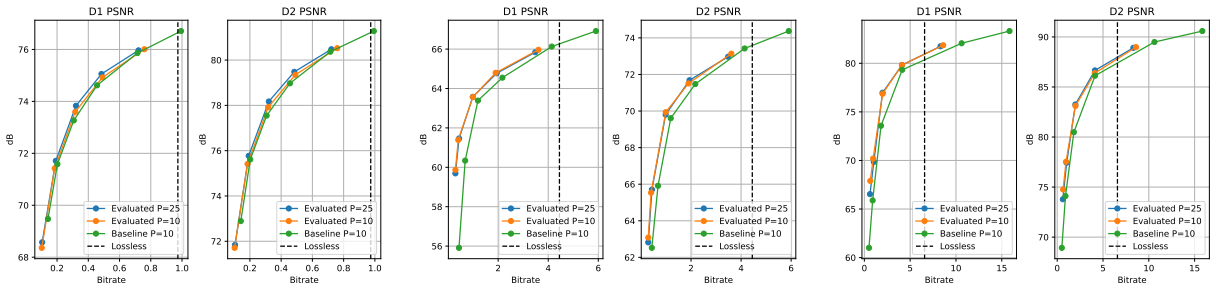
As a result, the decoded tensor $\hat{x}$ usually contains many additional points when compared to the input x. Similarly to the baseline model, the coordinates $\hat{x}_c$ are sorted according to the values of $\hat{x}_f$, which represent the estimated probability of occupancy for each coordinate. The decoded point cloud block will then contain the N_o points with the highest occupancy probability, with N_o being defined during compression as the value that maximizes a similarity metric between decoded and input block. During this experiment, the D1 PSNR metric is employed for this purpose.

This model is trained end-to-end to minimize a loss function equivalent to a weighted sum between the estimated bitrate of the compressed features and the distortion between the decoded point cloud block and the input. The rate R is estimated by the sum between the entropy of $\hat{y}_f$ and the entropy of $\hat{z}_f$, while the estimated distortion D is given by the sparse focal loss between $\hat{x}$ and x. The latter term is represented in Equation 1, where $\hat{x}_{f_j}$ and $\hat{x}_{c_j}$ correspond to the j^{th} row from $\hat{x}_f$ and $\hat{x}_c$, respectively.

$$\text{FL} = \begin{cases} -\alpha(1-\hat{x}_{f_j})^{\gamma}\log(\hat{x}_{f_j}), & \text{if } \hat{x}_{c_j} \in x_c \\ -(1-\alpha)\hat{x}_{f_j}^{\gamma}\log(1-\hat{x}_{f_j}), & \text{if } \hat{x}_{c_j} \notin x_c \end{cases} \quad (1)$$

The hyperparameter α can be configured to control the weight given to unoccupied voxels relative to occupied ones, while assigning a higher value of γ increases the importance given to voxels difficult to classify. The final loss value is given by $L = \lambda R + D$, with the hyperparameter λ setting the trade-off between rate and distortion.

The sequential procedure proposed by [13] was used to train the evaluated models in order to obtain different quality levels. In particular, the model with the lowest λ is first



(a) *Annibal* (b) *kinfudesk* (c) *Lausanne*

Fig. 4: Rate-distortion plots

trained from scratch, and the obtained weights are used to initialize the training for higher λ values. The baseline model was trained with λ values following the sequence $\{0.00025, 0.0005, 0.001, 0.002, 0.004, 0.008\}$. For the model with sparse convolutions, higher λ values had to be selected to allow for similar bitrates, since the focal loss is computed only on voxels in the neighborhoods of the input points rather than the entire block, given how voxels are generated at the decoder. Such voxels are harder to classify than the vastly empty zones of dense tensors, driving up the relative importance of the D term in the final loss value. Therefore, the sequence $\{0.0025, 0.005, 0.01, 0.025, 0.05\}$ was adopted.

Moreover, a patience parameter P was used to detect convergence of the model during training: if the loss function on the validation set does not decrease after P epochs, then training is stopped and the weights yielding the lowest loss value are selected. Both the baseline and the evaluated models were trained with $P = 10$ for all λ values. An additional model with sparse convolutions was trained using a learning rate scheduler, which decreased the learning rate by a factor of 10 whenever the validation loss was not reduced after 10 epochs. In this case, a patience value of $P = 25$ was set. All compression models are coded in PyTorch and were trained with an initial learning rate of 10^{-4} using the Adam optimizer, with values of $\alpha = 0.7$ and $\gamma = 2$ in the focal loss.

IV. Training and Testing Datasets

In order to train both the baseline and the evaluated compression models, the training and validation datasets presented in [8] were used, containing 35861 blocks of size $64 \times 64 \times 64$ obtained from 24 point clouds for training and 3822 blocks from 4 point clouds for validation. Recent works indicate that the performance of compression models depends heavily on the sparsity of the point clouds being compressed. One possible reason for this difference in performance is the distribution of the density values of the training set. In order to evaluate the impact of this factor, a sparsity metric is computed for each block in the training set by measuring the average distance from each point to its 5 nearest neighbors. The median value across all points is selected, denominated as the median 5-NN distance.

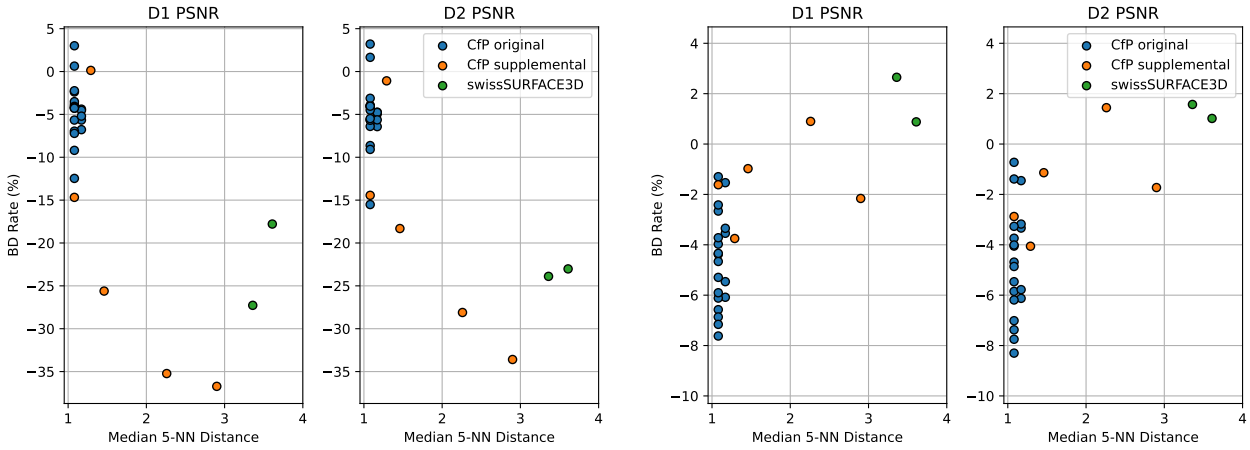
The distribution of this metric for the training set is illustrated in Figure 2. The majority of the point cloud blocks are highly dense, with more than 68% having a median 5-NN distance under 1.25. On the opposite side, less than 6% of the training set is between 2 and 3 in the histogram, and no block with sparsity higher than 3 was employed in the training set. Although the use of a more diverse training set would probably improve the performance of the model at higher sparsity levels, the same dataset was kept to ensure a fair comparison.

In order to test both the baseline and the evaluated compression models, the original test set from the JPEG Pleno Call for Proposals on Point Cloud Coding (CfP) [18] was used. The 20 point clouds were sampled with high density from meshes generated from the acquisition of real-world objects, all of them presenting a median 5-NN distance under 1.25. Additionally, the five point clouds from the supplemental dataset were employed, some of them presenting considerably smaller point densities. Finally, two point clouds obtained from swissSURFACE3D [19] were also included in the test set. This dataset was obtained with airborne LiDAR, with the entire set currently covering more than half of Switzerland. Two geographical regions were selected and the coordinates were voxelized with precision of 13 bits, leading to median 5-NN distances larger than 3. The entire test dataset contained a total of 27 point clouds, three examples of which are presented in Figure 3.

The test set was compressed and decompressed with the baseline model and the two versions of the evaluated model, using a block size of 128 and a latent quantization step of 1. They were additionally compressed using G-PCC software version 21 with lossless settings for comparison. Both point-to-point PSNR (D1 PSNR) and point-to-plane (D2 PSNR) metrics were computed on the decompressed models.

V. Results and Discussion

The metric values for the evaluated and baseline models were plotted against their bitrates, with a dashed vertical line indicating the bitrate for lossless encoding with G-PCC. The plots for the point clouds illustrated in Figure 3 are presented in Figure 4. For the *Annibal* point cloud, modest gains in performance from the use of sparse convolutions are observed



(a) Comparison between evaluated model with $P = 10$ and the baseline

(b) Difference between values obtained with $P = 25$ and $P = 10$

Fig. 5: BD-Rate reductions for D1 PSNR and D2 PSNR

across the evaluated range. Also, a difference in performance is observed when training the model at a higher patience value, with longer training leading to higher quality. The plots for *kinfudesk* indicate that even larger gains can be obtained for sparser point clouds. Moreover, it is also observed that the difference between different patience levels is reduced, possibly due to the lack of point cloud blocks with similar sparsity levels in the training set. It is also observed that the highest rate of the plot is significantly above the lossless line. Finally, the results obtained for *Lausanne* demonstrate that the model with sparse convolutions outperforms again the baseline, with a higher difference in performance for *Annibal*, but lower than for *kinfudesk*. Moreover, performance between models trained with $P = 10$ and $P = 25$ is again very similar, with even a slight advantage of the former at lower bitrates. Since this point cloud is even sparser than *kinfudesk*, these findings reveal that decreasing point density does not necessarily lead to an increased advantage of the sparse convolutions.

In order to better evaluate the effect of point cloud sparsity on the rate-distortion performance, the BD-Rate between the evaluated model trained with $P = 10$ and the baseline was obtained for each point cloud, ignoring quality levels above the lossless rate for G-PCC. The evaluated model achieved an average bitrate saving of approximately 9%. A consistent increase in performance from the use of sparse convolutions is observed while using the exact same training strategy.

The BD-Rate results were plotted against the median 5-NN distance of each point cloud and are displayed in Figure 5a, the color of each point of the plot indicating the original source of each point cloud. The use of sparse convolutions allowed for lower bitrates for the majority of test set. All point clouds of the original test set of the CfP are grouped at the left of the plot, achieving a BD-rate difference varying between 3% and -15%. Since BD-rate values can vary at a range of approximately 18 percentage points at almost identical point density levels, these results show that sparsity is not the only factor that determines the performance of sparse convolutions. However, the analysis of the CfP supplemental set indicates that sparsity is indeed among the most influential factors, with a high correlation between BD-Rate reduction and median 5-NN distance. In particular, the point exhibiting the highest rate reduction is the most sparse model from this set, achieving more than 36 % savings for the D1 PSNR metric at 2.9 median 5-NN distance. While this strong correlation would suggest that even higher savings should be possible on the sparser models from swissSURFACE3D, this trend was not observed, with overall rate difference remaining between -17% and -27%. Such results indicate that other factors such as homogeneity or voxelization precision can affect the compression performance as well.

The BD-Rate between the evaluated model trained with $P = 25$ and the baseline model was also computed for the entire set. The difference between these values and those from the previous comparison is presented in Figure 5b. It is observed that including a mechanism for progressively decreasing the learning rate and waiting more epochs prior to stop of the training induced an increase in performance for the majority of the tested point clouds. Indeed, the model trained with $P = 25$ achieved an average BD-Rate difference of approximately -12.5% when compared to the baseline. In particular, point clouds with higher density were more favored by the higher patience value. However, a longer training process was slightly detrimental to the efficiency of the compression of sparser point cloud models, likely because it caused the neural network to specialize for the sparsity values better represented in the training set. Since blocks with a median 5-NN distance higher than 2 account for only a small portion of the training data, higher patience values are not beneficial. Rather than encouraging earlier stops of the training process, these results

indicate the importance of including point cloud models with a wider range of sparsity values in the training set.

Naturally, the obtained results depend also on the evaluation conditions. For instance, the lack of sparse point cloud blocks in the training set probably hinders the performance of both the baseline and the evaluated model. As a matter of fact, no blocks with the same sparsity as *kinfubooks*, nor any models from swissSURFACE3D were used for the optimization of the models. Yet, the models are still capable of encoding such point clouds at rates below lossless with acceptable quality. While these results show the generalization capacity of the neural network to unseen examples, using a more diverse training set would probably increase the rate-distortion performance of such learning-based methods.

Aside from the rate reduction in sparse convolutions, a reduction in computational complexity is also inherently obtained since the convolution operations need to compute at fewer spatial locations. While this feature is already an advantage in itself, it would also allow the use of larger blocks both during training and testing. Indeed, one major limitation of using dense convolutions is their memory usage, which restricts the size of the point cloud blocks that can feed the neural network. While the point cloud size used by compression models based on sparse convolutions is not limitless, higher dimensions could certainly be used due to their smaller memory footprint, likely enabling better performance as previously demonstrated for models using dense convolutions.

Moreover, the hyperparameters for the loss function selected for the training of both the baseline and evaluated model were established by [8], considering only the characteristics of dense convolutions. In particular, the α parameter is set to 0.7 in order to give a higher weight on the correct classification of occupied voxels due to the fact that most spatial positions of the dense input tensor x are empty. Giving instead the same weight to both occupied and empty voxels would skew the network into producing lower probabilities for the occupancy of most positions due to class imbalance. However, the decoder of the evaluated model produces sparse tensors with coordinates only in the neighborhoods of occupied voxels, not considering regions that are totally empty. Therefore, the optimal α value is likely different from that of the baseline, and adapting this hyperparameter could lead to even better results. These experiments were considered out of the scope of this paper and are deferred to future work.

VI. Conclusion

In this paper, an evaluation of the performance of sparse convolutions for point cloud geometry compression is conducted by replacing the dense convolutions of an existing compression model with sparse layers, with minimal additional changes. Aside from the intrinsic complexity reduction, an increase in the rate-distortion performance is also observed, with an average BD-Rate reduction of approximately 9% in the evaluated test set. An improved training process also allowed to increase the rate savings to nearly 12.5%. While the improvement of rate-distortion performance is observed for the majority of the point clouds, the sparse convolutions are particularly effective for test models with lower density, with rate savings going up to 35%. The fact that such results were obtained without major adaptations indicates that sparse convolutions are more suitable for point cloud compression in most cases, corroborating the recent shift in research trends.

References

[1] MPEG Systems, "Text of ISO/IEC DIS 23090-18 Carriage of Geometry-based Point Cloud Compression Data," ISO/IEC JTC1/SC29/WG03 Doc. N0075, Nov. 2020.

[2] MPEG 3D Graphics Coding, "Text of ISO/IEC CD 23090-5 Visual Volumetric Video-based Coding and Video-based Point Cloud Compression 2nd Edition," ISO/IEC JTC1/SC29/WG07 Doc. N0003, Nov. 2020.

[3] A. F. Guarda, N. M. Rodrigues, and F. Pereira, "Deep learning-based point cloud coding: A behavior and performance study," in *2019 8th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2019, pp. 34–39.

[4] M. Quach, G. Valenzise, and F. Dufaux, "Learning convolutional transforms for lossy point cloud geometry compression," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 4320–4324.

[5] J. Wang, D. Ding, Z. Li, and Z. Ma, "Multiscale point cloud geometry compression," in *2021 Data Compression Conference (DCC)*. IEEE, 2021, pp. 73–82.

[6] J. Wang, D. Ding, Z. Li, X. Feng, C. Cao, and Z. Ma, "Sparse tensor-based multiscale representation for point cloud geometry compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[7] J. Pang, M. A. Lodhi, and D. Tian, "Grasp-net: Geometric residual analysis and synthesis for point cloud compression," in *Proceedings of the 1st International Workshop on Advances in Point Cloud Compression, Processing and Analysis*, 2022, pp. 11–19.

[8] A. F. Guarda, N. M. Rodrigues, M. Ruivo, L. Coelho, A. Seleem, and F. Pereira, "It/ist/ipleiria response to the call for proposals on jpeg pleno point cloud coding," *arXiv preprint arXiv:2208.02716*, 2022.

[9] R. Schnabel and R. Klein, "Octree-based point-cloud compression," in *Symposium on Point-Based Graphics 2006*, Jul. 2006.

[10] J. Kammerl, N. Blodow, R. B. Rusu, S. Gedikli, M. Beetz, and E. Steinbach, "Real-time compression of point cloud streams," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 778–785.

[11] E. Pavez, P. A. Chou, R. L. de Queiroz, and A. Ortega, "Dynamic polygon clouds: representation and compression for VR/AR," *APSIPA Transactions on Signal and Information Processing*, vol. 7, p. e15, 2018.

[12] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[13] M. Quach, G. Valenzise, and F. Dufaux, "Improved deep point cloud geometry compression," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[14] J. Wang, H. Zhu, H. Liu, and Z. Ma, "Lossy point cloud geometry compression via end-to-end learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4909–4923, 2021.

[15] N. Frank, D. Lazzarotto, and T. Ebrahimi, "Latent space slicing for enhanced entropy modeling in learning-based point cloud geometry compression," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4878–4882.

[16] D. Lazzarotto, E. Alexiou, and T. Ebrahimi, "On block prediction for learning-based point cloud compression," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3378–3382.

[17] D. Lazzarotto and T. Ebrahimi, "Learning residual coding for point clouds," in *Applications of Digital Image Processing XLIV*, vol. 11842. SPIE, 2021, pp. 223–235.

[18] WG1, "Final Call for Proposals on JPEG Pleno Point Cloud Coding," ISO/IEC JTC1/SC29/WG1 Doc. N100097, Jan 2022.

[19] "swisssurface3d," [Accessed: 2022-05-11] https://www.swisstopo.admin.ch/en/geodata/height/surface3d.html.

Study on Viewpoint-Dependent Time-Multiplexing for Weighted Optimization of 3D Layered Displays

Armand Losfeld, Daniele Bonatto, Gauthier Lafruit, Mehrdad Teratani
Laboratories of Image Synthesis and Analysis (LISA)
Université Libre de Bruxelles (ULB)
Brussels, Belgium
{firstname}.{lastname}@ulb.be

***Abstract*—3D Layered displays are a type of 3D display that stacks LCD panels to reproduce different viewpoints of a scene without glasses. However, these displays fail at reproducing high-parallax scenes, thereby limiting their Field of View (FoV). To enhance the FoV, various strategies have been employed, including the multiplexing of distinct sets of layered images, i.e. frames. Despite achieving improved quality, this multiplexing method let no control to adjust the quality of the frames depending on its application. By introducing distinct weights for each frame during the optimization process, we expect to improve the frames' optimization based on the input data, and thus, the quality of the display. This weighted-multiplexing method motivates us to investigate the use of a first weighted-multiplexing method to study exhaustively the impact of the weights on the multiplexing optimization process. The proposed method involves viewpoint-dependent time-multiplexing where each frame is tailored to optimize a specific viewing region within the FoV. To define the weights of each frame, three weighted approaches are then proposed. Three objective evaluations and one subjective comparison are presented in the study.**

***Index Terms*—3D Layered Displays, Time-Multiplexing, Viewpoint-Dependent, Light Field, Virtual Reality**

I. Introduction

In recent years, virtual reality and 3D videos regained popularity primarily driven by the *Metaverse* [1] concept in commercial applications of multinational corporations. However, these commercial applications rely on the immersive experience provided by current 3D displays, which use is limited by the cyber-sickness effect caused by the lack of depth cues. Notably, head-mounted displays [2] provide a great immersive experience, but they suffer from their discomfort and lack adequate eye accommodation and eye vergence. Therefore, holographic displays [2]–[6] have emerged as a potential solution, aiming to reproject multiple glasses-free viewpoints with better depth cues at an expensive cost.

3D layered display [7]–[9] represents an affordable alternative that enables a glasses-free 3D viewing experience. These displays consist of a backlight and a stack of n display panels, typically LCD panels. The emitted light from the backlight traverses through each panel, and its final color and intensity are determined by the intersected pixels present in the display panels. Through precise control of the pixels' colors on each panel, multiple viewpoints of a scene can be reproduced. Consequently, the pixel values for the layers are determined by solving an optimization problem. An iterative factorization method [7] is used in the case of an *attenuation-based* layered display where each LCD panel is surrounded by two polarizing filters. For a *polarization-based* display, where only two polarization filters are used, an iterative projection method [8] is commonly used. Recently, a quasi-Newton method [10] was used for both display designs. In this paper, the quasi-Newton method and only *polarization-based* displays [8] are considered since both display types always lead to comparable conclusions [10] though *polarization-based* displays give slightly superior outcomes compared to *attenuation-based* ones.

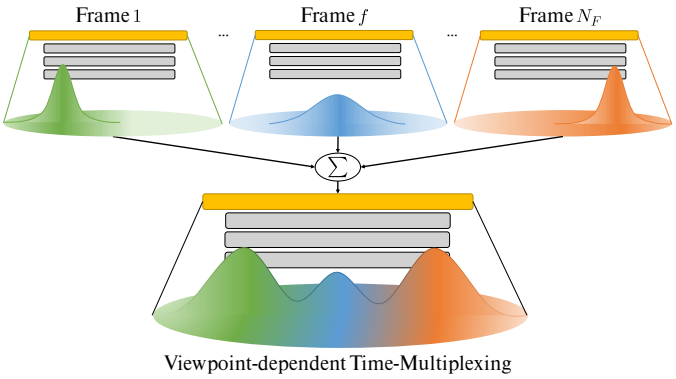


Fig. 1. Our Viewpoint-dependent Time-Multiplexing method uses different weight distributions depending on the frame of the multiplexing.

Compared to other displays, layered displays generally exhibit a limited Field of View (FoV) which is related to the amount of parallax [7], [11]–[13] present in the target scene. To address this limitation, a weighted method [14]–[16] was proposed to dynamically adjust the narrow light field [17], [18], *i.e.* a set of narrowly rendered viewpoints, based on the introduction of weights during the optimization process allowing the use of only a subset of the input light field.

Alternatively, *time-multiplexing* [7], [10] was proposed. It formulates the problem as the optimization of N_F different sets of multi-layer images, *i.e.* N_F *frames*, that are displayed simultaneously. Although the quality increases, there is no control over the frames' optimization process and a more sophisticated multiplexing process based on the light field

Armand Losfeld is funded by the *Ecole Polytechnique de Bruxelles* (EPB) from the *Université Libre de Bruxelles* (ULB).

979-8-3503-4218-5/23/$31.00 © 2023 European Union

information might outperform the conventional approach.

We investigate the potential of such a method by exploring the impact of the weights on the multiplexing process. For the study, we propose a first weighted-multiplexing method, called *Viewpoint-dependent Time-Multiplexing*, which uses distinct weights for each frame to force each on reproducing in high quality a distinct subset of the light field, cf. Fig. 1. To define the weights of each frame, three weighted approaches based on the light field spatial resolution and the frame count are proposed. Two datasets are used to evaluate the three approaches, identify the optimal weighted parameterization, and study the differences compared to the conventional multiplexing method. Even if no significant improvements are expected due to the frame averaging process of the multiplexing, the exhaustive study of the proposed method will help define a more sophisticated weighted-multiplexing method.

II. Proposed Method

A. Layered Display Model

The light field formalism [17], [18] is commonly employed to describe 3D layered displays, cf. Fig. 2, in particular the two-planes parametrization [17]. This formalism provides a framework for representing the radiance of all light rays captured by two distinct planes. The first plane (u, v) is the camera plane while the second plane (s, t) is the focal plane. By considering one of the LCD layers as the focal plane and placing the camera plane at an arbitrary distance from the display, the light rays of a layered display of N_L layers are expressed in the light field formalism as follows:

$$\tilde{l}(u, v, s, t) = (g_{N_L} \circ \cdots \circ g_1)(s', t'; \tilde{l_0}) \tag{1}$$

where $g_i(s', t'; x)$ is the response of the pixel (s', t') in the layer i to the input x, $\tilde{l_0}$ is the backlight intensity, and (s', t') are defined as the intersection of the light ray with the layer i. Time-multiplexing is then introduced as an average of N_F different sets of multi-layer images displayed simultaneously.

$$\tilde{l}(u, v, s, t) = \frac{1}{N_F} \sum_{f=1}^{N_F} \tilde{l}^f(u, v, s, t) \tag{2}$$

where $\tilde{l}^f$ is the reproduced light field of the frame f. More generally for the rest of the paper, the notation $\cdot^f$ will refer to an element $\cdot$ of the frame f.

By denoting the target light field with $l(u, v, s, t)$, the problem of computing the frames displayed for the 3D reproduction can be formulated as an optimization problem where we minimize a cost function, such as the mean-squared norm, between the reproduced and target light fields. Following the approach in [14]–[16], we introduce the weights $w_{u,v}$ within the cost function C. However, our method uses distinct weights for each frame of the time-multiplexing process, *i.e.* $w^f_{u,v}$. Indeed, with this strategy, each frame can be forced to be optimized on a subset of the input light field, corresponding to particular viewpoints of the scene. In this section, only the incorporation of the weights in the optimization problem is presented, the weights will be defined in the next section. The cost function is

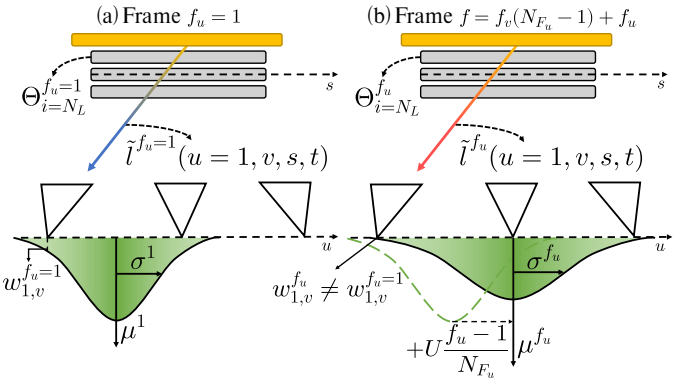


Fig. 2. An example of the *Centered-Gaussian* approach used for two viewpoint-dependent weighted frames. (a) For the frame $f_u = 1$, the weight distribution is centered on the left side of the FoV and is narrow. (b) For an arbitrary horizontal frame f_u, the distribution is displaced on the right. Here, the distribution is also flatter since its position is at the FoV center, and thus, the weight for the view $u = 1$ is much less than it was for the frame $f_u = 1$.

$$C = \left|\left| \sum_f w^f_{u,v} \left(\tilde{l}^f(u, v, s, t) - l(u, v, s, t) \right) \right|\right|^2 . \tag{3}$$

For *polarization-based* displays, the composition of the g_i functions [8] is then defined by:

$$(g_{N_L} \circ \cdots \circ g_1)(s', t'; I) = I \sin^2 \left(\sum_{i=1}^{N_L} \theta_i(s', t') \right) \tag{4}$$

where $\theta_i(s', t')$ denotes the light wave phase shifts of the pixel (s', t') of the layer i, and I the initial intensity. Since LCD layers have a fixed pixel number N_P, the cost in (3) can be expressed in a matrix formalism similarly as in [10]:

$$C = \left|\left| \sum_{u,v} h(\Theta) \right|\right|^2 \quad h(\Theta) = \mathbf{E}_{u,v} \phi(\mathbf{A}_{u,v} \Theta) - w_{u,v} L_{u,v} \tag{5}$$

where $w_{u,v}$ is a scalar obtained by summing the weights of the multiplexed frames for the view (u, v). The weighted multiplexing matrix $\mathbf{E}_{u,v}$, of dimensions $N_P \times (N_F \cdot N_P)$, and the orthographic projection matrix $\mathbf{A}_{u,v}$, of dimensions $(N_F \cdot N_P) \times (N_F \cdot N_L \cdot N_P)$, are in bold. The vector Θ contains $N_F \cdot N_L \cdot N_P$ elements and represents the phase shifts of all pixels' layers of all multiplexed frames. The target view vector $L_{u,v}$ for view (u, v) has size N_P. The element-wise operator ϕ is defined, for *polarization-based* displays, as $\phi = \sin^2(\cdot)$. The weighted multiplexing matrix $\mathbf{E}_{u,v}$ is used to sum the frames and scaled each of them with a factor $w^f_{u,v}/N_F$.

The gradient of C is therefore given by:

$$\nabla C = 2 \sum_{u,v} \left(\mathbf{A}^T_{u,v} \mathrm{diag} \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \phi(\mathbf{A}_{u,v} \Theta) \right) \mathbf{E}^T_{u,v} h(\Theta) \right) \tag{6}$$

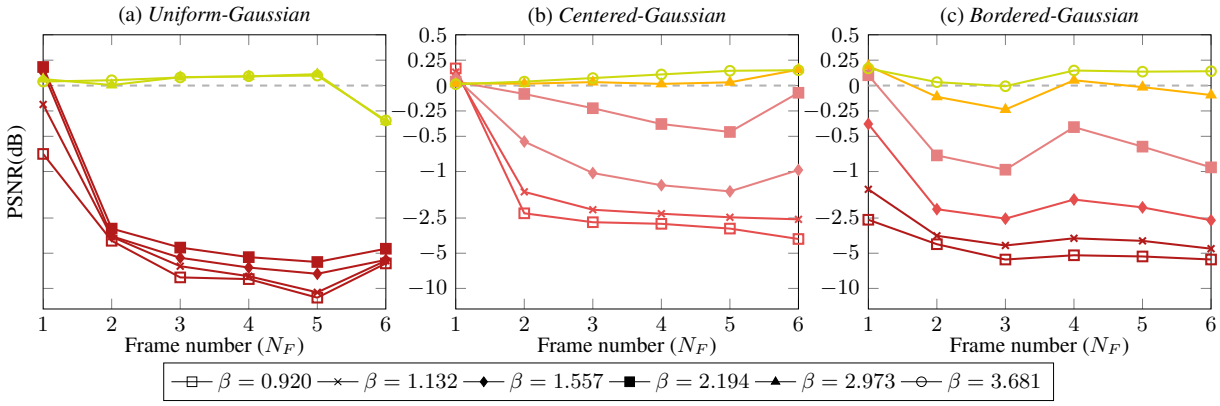


Fig. 3. *Experiment 1*: Difference of the mean PSNR of our method with the conventional method in the function of the frame number N_F (up to 6 frames), the three weighted approaches, and the parameter β. For the weighted approaches, the first five frames are only horizontal frames (*i.e.* $N_{F_v} = 1$) while the sixth frame is divided into 3 horizontal and 2 vertical frames. The iteration number of the L-BFGS algorithm was set to 300 and the parameter c to $100/2\pi$.

As in [10], a solution is found by giving the cost function and its gradient to the L-BFGS algorithm [19] with the same transformation variable, *i.e.* $x = \frac{\pi}{2}\sin^2(y)$.

B. Viewpoint-dependency with multiplexing

The effectiveness of our viewpoint-dependent time-multiplexing method critically relies on the appropriate selection of weights, as random choices yield a wrong optimization. Moreover, there exists an infinite number of choices for partitioning the Field of View (FoV) into distinct viewing zones depending on the number of frames and the light field spatial resolution. In this section, we present three solutions: one naive and two alternative approaches which take into account the small amount of information on the extremity of the light field.

The **Uniform-Gaussian** approach is the *naive* strategy. It divides equally the FoV into N_F viewing zones with two parameters: the horizontal frame number N_{F_u} and the vertical one N_{F_v} with $N_F = N_{F_u}N_{F_v}$. Gaussian distributions are then positioned at the center of each viewing zone to determine the weights assigned to each view (u, v) of each frame f. Also, the amplitude of the distributions depends on the light field resolution (*i.e.* the number of views), the frame number, and the amplitude parameter c. For every frame, the weights assigned to each viewpoint (u, v) are given by:

$$w^f_{u,v} = \alpha e^{\lambda^{f_u}_u + \lambda^{f_v}_v} \quad \lambda^{f_x}_x = -\frac{(x-\mu^{f_x})^2}{2(\sigma^{f_x})^2} \quad \alpha = c\frac{UV}{\sigma^{f_u}\sigma^{f_v}} \tag{7}$$

where UV is the spatial resolution of the light field, x is either u or v, and c is the amplitude parameter. While μ^{f_x} only depends on the frame number N_{F_x}, the frame f_x, and X the resolution of the coordinate x (*e.g.* for u, $X = U$); σ^{f_x} also depends on the weight-control parameter β.

$$\sigma^{f_x} = \beta\frac{X-1}{4+2N_{F_x}} \quad \mu^{f_x} = \frac{X(1+f_x)}{(2+N_{F_x})} \tag{8}$$

The **Centered-Gaussian** is the second approach. It introduces a variable width for the Gaussian distributions, which is based on the proximity of the zone center with the borders. If the center is near one of the borders of the light field, then the distribution is narrower. Conversely, if the center is located at the center of the FoV, the distribution becomes flatter. From (8), σ^{f_x} becomes

$$\sigma^{f_x} = \beta \cdot \begin{cases} 1.0 + \mu^{f_x} & \text{if } \mu^{f_x} \leq \frac{X}{2} \\ 1.0 + X - \mu^{f_x} & \text{if not.} \end{cases} \tag{9}$$

The **Bordered-Gaussian** is the third approach and is defined as the opposite of the second approach. The distributions are narrow at the center of the FoV but flatter at the FoV extremities.

$$\sigma^{f_x} = \beta\left(1.0 + \left|\frac{X}{2} - \mu^{f_x}\right|\right) \tag{10}$$

For the Centered-Gaussian and Bordered-Gaussian strategies, an initial width of 1.0 was introduced to avoid zero width when μ^{f_x} approaches 0, X, or $X/2$.

For all the proposed approaches, the weights $w^f_{u,v}$ are appropriately scaled to ensure that their sum remains similar, *i.e* $\forall f \sum_{u,v} w^f_{u,v}$ is the same. Without this condition, some frames could contribute more than others to the light field reproduction and thus yield wrong results.

III. Experiments and Results

Three qualitative experiments and one subjective evaluation are presented in this research. For each qualitative experiment, the Peak Signal-to-Noise Ratio (PSNR) metric was used to assess the quality. The first experiment studies the efficiency of our method under various parameter configurations, and the second and third experiments study the method's impact on a wider FoV. Lastly, subjective comparisons are presented to evaluate the perceptual aspect of the proposed method. For all experiments, the maximal number of frames is set to six because using more frames highly increases the chance of

TABLE I

EXPERIMENT 2: COMPARISON BETWEEN THE CONVENTIONAL MULTIPLEXING AND OUR VIEWPOINT-DEPENDENT MULTIPLEXING ON A WIDE FOV FOR DIFFERENT PARAMETRIZATIONS. IN ORANGE AND IN BLUE, THE WORST AND BEST RESULTS IN A ROW; IN YELLOW, THE RESULT BEING COMPARED.

Method	**Conventional**			**Uniform-Gaussian**			**Centered-Gaussian**			**Bordered-Gaussian**		
Frame number	1 ***(a)***	6 ***(b)***	10 ***(c)***	6			6			6		
Thickness β				1.557	3.681	5.097	1.557	3.681	5.097	1.557	3.681	5.097
Min PSNR (dB)	25.236	27.613	28.180	22.638	25.086	26.335	25.403	27.546	27.731	23.615	26.171	26.926
Max PSNR (dB)	31.601	34.418	35.262	35.167	36.613	35.882	36.208	34.994	34.766	34.782	34.409	34.730
Mean PSNR (dB)	28.594	31.519	32.124	27.327	30.099	31.046	30.325	31.692	31.692	28.424	30.931	31.406
Diff w/ ***(a)*** (dB)	0	2.924	3.530	−1.266	1.505	2.452	1.731	3.098	3.098	−0.170	2.337	2.812
Diff w/ ***(b)*** (dB)	−2.924	0	0.605	−4.191	−1.419	−0.472	−1.193	0.173	0.173	−3.095	−0.587	−0.112
Diff w/ ***(c)*** (dB)	−3.530	−0.605	0	−4.796	−2.024	−1.077	−1.798	−0.431	−0.431	−3.700	−1.193	−0.717

flicker effects due to LCD hardware limitations. But in the second and fourth experiments, the result of the conventional method utilizing ten frames is presented for completeness.

The software implementation uses C++17 and the open-source libraries OpenCV [20], Eigen [21], and LBFGSpp [22]. The program was executed on a Windows 10 operating system equipped with an Intel $i9-10920X$ processor running at 3.50 GHz. For all experiments, the iteration number of the L-BFGS algorithm was set to 300 to assure its convergence, and the parameter c was set to $100/2\pi$ to avoid divergent results due to single-floating numerical precision.

Two datasets were used: *Dice* [23] and *SauceDino* [24]. The *Dice* dataset [23] has 7×7 orthographic viewpoints of size 512×384 for an FoV of $10°$ and was used in the first experiment. The dataset is composed of a texture-less background and 5 dice of different colors and depths which result in a reproduction of high quality ($\approx$ 30 dB) without time-multiplexing. The *SauceDino* has 15×15 perspective viewpoints of size 512×300 placed at 4 meters from the wall and focused on the dinosaur's body for an FoV of $6°$. Three scanned objects from the collection *Scanned Objects by Google Research* [25], two user-modeled objects, and a brick texture background compose the scene. This dataset, used for the second experiment, is generally harder to reproduce due to the high-texture objects. Note that, the FoV of the second dataset is smaller due to the perspective acquisition.

A. Parameters' analysis

This experiment studies the impact of the parameters of our method on a small popular dataset (*i.e. Dice* [23]). Our three weighted approaches are compared to the conventional method by varying the frame number N_F, up to six frames, and the parameter β, cf. Fig. 3. Since the weighted approaches use horizontal and vertical frame numbers, the first five frames are only set horizontally (*i.e.* $N_{F_v} = 1$ and $N_F = N_{F_u}$) but the sixth one is set to three horizontal frames and two vertical frames.

It is observed that the naive approach, *i.e. Uniform-Gaussian*, fails to outperform the conventional method for low values of β for any frame number. When β exceeds 2.5, it yields slightly higher or similar results. These observations are also valid for the other two approaches even if their PSNR loss is significantly lower and more spread. The fact that

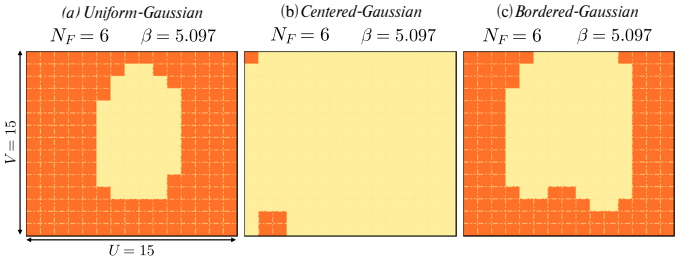


Fig. 4. Difference of PSNR quality for all viewpoints of our weighted approaches using 6 frames and $\beta = 5.097$ with the conventional method using 6 frames. Gains are highlighted in yellow while losses are in orange.

only high β values yield similar or minor improved results would mean that the optimization process is preferable to be used with a good amount of information. Indeed, if β is low, most Gaussian distributions are narrow and the information contained in some viewpoints is not used anymore during the optimization process. Therefore, carefully choosing β is crucial.

It is important to note that our method gives slightly higher results when utilizing only one frame, *i.e.* absence of multiplexing, certainly due to the low-textured background, dice, and low parallax of the *Dice* dataset. When employing a single frame, our method focuses the optimization process on the central view, resulting in very high PSNR values($\approx 40-50$ dB) for that particular view. For border views, reasonable values ($\approx 24-27$ dB) are achieved. Hence, with our method, we noted that the average PSNR is biased when only one frame is employed.

B. Analysis on large Field of View

The *second* experiment analyzes the impact of the method on a wide FoV. To accomplish this, a wider dataset was used, *SauceDino* [24], to compare the mean quality of (a) a single conventional frame, (b) employing six conventional frames, and *(c)* utilizing ten conventional frames with our approaches using different parametrizations. Table I summarizes the minimum, maximum, and mean PSNR of the reproduced light field optimized with the different configurations. Additionally, the differences in the mean PSNR between the three conventional configurations and the others using our method are presented. The worst and best results are highlighted with different colors.

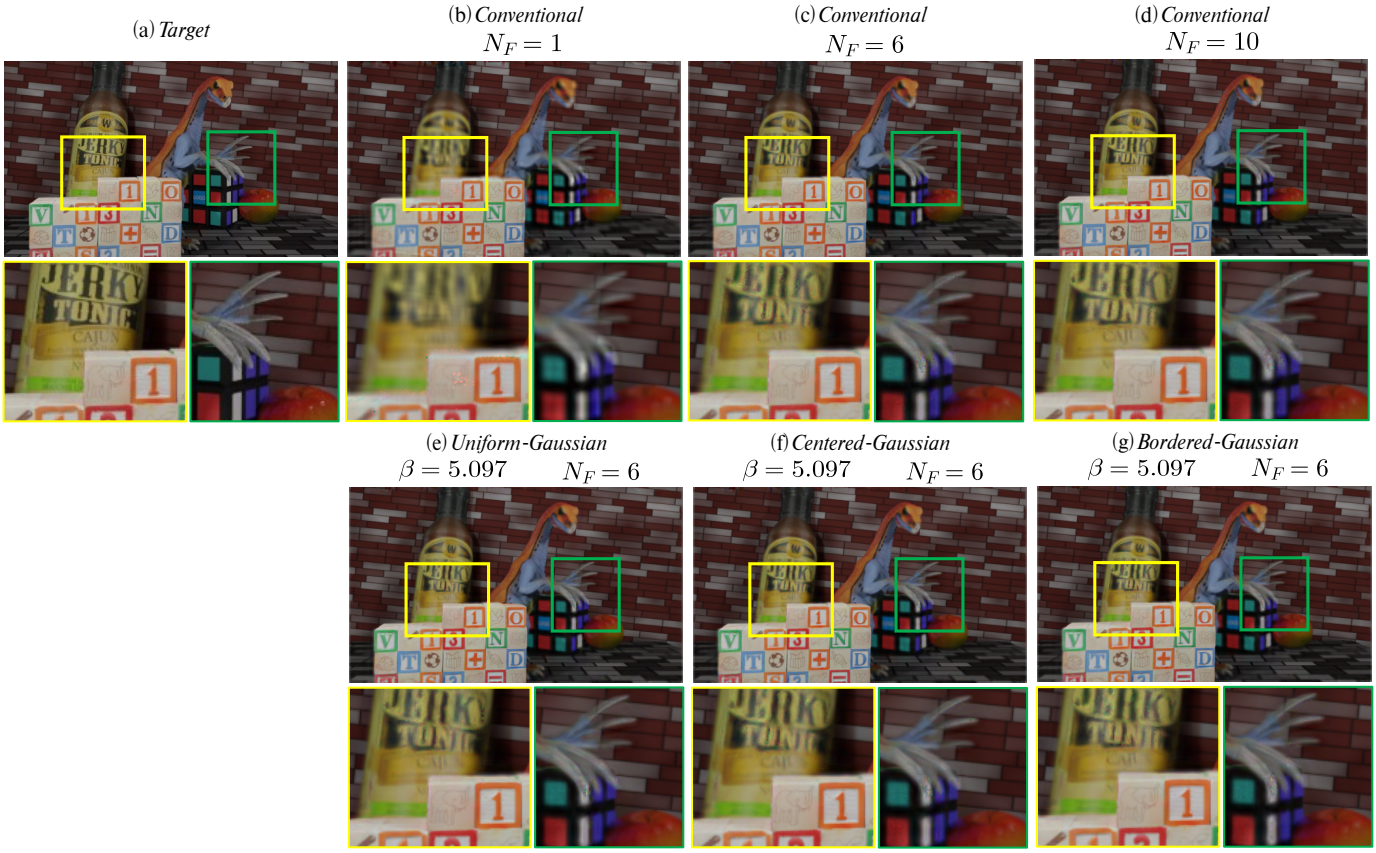


Fig. 5. Subjective comparison of one of the top-left views from (a) the target light field and (b-g) the light fields reproduced with different optimizations.

From the analysis of the three last rows of results, it is evident that the poorest outcomes are always given by the *Uniform-Gaussian* approach with a low β value. Even when optimizing six frames, the layer-wise optimization without multiplexing still outperforms it by 1.2dB. For all our weighted approaches, using a low β value always failed at giving similar or better results than the conventional method using the same frame number. These observations confirmed the sensitivity of our method to β and that using a low β might lose some essential information needed during the optimization process. Also, as in the first experiment, our method gives a minor improvement of 0.17 dB when higher β values are used. This minor improvement is studied carefully by subjectively comparing one reproduced view in the last experiment.

When more information is available to reproduce the light field, as the conventional using ten frames, we observe that our method failed at giving similar results.

C. Gain/loss distribution of multiplexing methods

In this experiment, we use the results collected in the second experiment to focus on the distribution of the quality in the reproduced light fields. Three multiplexing methods are compared with the conventional method using six frames, cf. Fig. 4.

The first and third approaches, *Uniform-Gaussian* and *Bordered-Gaussian* with $\beta = 5.097$ and six frames, exhibit better performance in the center of the light field while the *Centered-Gaussian* is better in 97.7% of the light field. Even if minor changes are observed on average, the quality is not distributed randomly in the light field. This would possibly mean that very slight improvements are achieved when the average is slightly superior.

It is interesting to note that top and right viewpoints appear to be easier to reproduce due to the scene composition. Indeed, the complexity of the scene is not uniformly distributed and more textured objects are present on the bottom-left side, such as the sauce and the letter cubes.

D. Subjective comparison of one viewpoint

Finally, a subjective comparison of one top-left viewpoint of the light field is presented, cf. Fig. 5. We compare the views reproduced by our weighted approaches using $\beta = 5.097$ and six frames to the conventional method using one, six, and ten frame(s). Overall, only small differences can be observed between the views reproduced with a multiplexing method, such as the noise amount in the bottle sticker. However, since only the noise amount is reduced, it is difficult to conclude that any approach yields improvements.

In zoom-in regions, as stated in Table I, the configuration with ten conventional frames gives the best results while the method without multiplexing gives the worst. Ranking the

remaining subjective results is challenging due to their high similarity.

IV. Conclusion

In this study, we introduced a novel weighted-multiplexing method consisting of three weighted approaches that can be adjusted using two parameters: the weight-control β and the amplitude parameter c. Our method divides the FoV evenly among the frames and employs weights to guide the optimization process of each frame on specific viewpoints in the input light field. Furthermore, we conducted experiments using two datasets to study the effects of incorporating weights in the multiplexing procedure.

As expected, our method does not yield considerable improvement compared to the conventional method, and in fact, yields inferior results when the weight-control parameter was set to a low value. In such cases, the Gaussian distributions, defining the weights, become narrow, resulting in some viewpoints being disregarded. Because the optimization process is done considering less information, viewpoint-dependent weights are undoubtedly sub-optimal for a weighted-multiplexing method. Therefore, we foresee the use of pixel-dependent weights for a further weighted-multiplexing method. By defining weights based on pixel information or groups of pixels, it may be possible to achieve adjustable frame optimization based on the input data while avoiding the removal of important information during the optimization process.

Acknowledgment

This work was supported in part by the Emile DEFAY 2021 project (grant N° 4R00H000236), Belgium; in part by the FER 2021 project (grant N° 1060H000066-FAISAN), Belgium; and in part by the FER 2023 project (grant N° 1060H000075), Belgium.

References

[1] G. D. Ritterbusch and M. R. Teichmann, "Defining the metaverse: A systematic literature review," *IEEE Access*, vol. 11, pp. 12 368–12 377, 2023.

[2] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister, *Depth cues in human visual perception and their realization in 3D displays*, B. Javidi, J.-Y. Son, J. T. Thomas, and D. D. Desjardins, Eds. SPIE, 2010, vol. 7690. [Online]. Available: https://doi.org/10.1117/12.850094

[3] T. Balogh, T. Forgács, T. Agács, O. Balet, E. Bouvier, F. Bettio, E. Gobbetti, and G. Zanetti, "A scalable hardware and software system for the holographic display of interactive graphics applications." *Eurographics (Short Presentations)*, pp. 109–112, 2005. [Online]. Available: http://dx.doi.org/10.2312/egs.20051036

[4] J. Hua, E. Hua, F. Zhou, J. Shi, C. Wang, H. Duan, Y. Hu, W. Qiao, and L. Chen, "Foveated glasses-free 3d display with ultrawide field of view via a large-scale 2d-metagrating complex," *Light: Science & Applications*, vol. 10, no. 1, p. 213, Oct 2021. [Online]. Available: https://doi.org/10.1038/s41377-021-00651-1

[5] P. Chakravarthula, E. Tseng, T. Srivastava, H. Fuchs, and F. Heide, "Learned hardware-in-the-loop phase retrieval for holographic near-eye displays," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, p. 186, 2020.

[6] F. Yaraş, H. Kang, and L. Onural, "State of the art in holographic displays: A survey," *Journal of Display Technology*, vol. 6, no. 10, pp. 443–454, 2010.

[7] W. Gordon, L. Douglas, H. Matthew, and R. Ramesh, "Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 31, no. 4, pp. 1–11, 2012.

[8] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, and R. Raskar, "Polarization Fields: Dynamic Light Field Display Using Multi-Layer LCDs," in *Proceedings of the 2011 SIGGRAPH Asia Conference*, ser. SA '11. New York, NY, USA: Association for Computing Machinery, 2011. [Online]. Available: https://doi.org/10.1145/2024156.2024220

[9] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar, "Layered 3d: Tomographic image synthesis for attenuation-based light field and high dynamic range displays," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: Association for Computing Machinery, 2011. [Online]. Available: https://doi.org/10.1145/1964921.1964990

[10] J. Zhang, Z. Fan, D. Sun, and H. Liao, "Unified Mathematical Model for Multilayer-Multiframe Compressive Light Field Displays Using LCDs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 3, pp. 1603–1614, 2019.

[11] K. Takahashi, Y. Kobayashi, and T. Fujii, "From Focal Stack to Tensor Light-Field Display," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4571–4584, 2018.

[12] A. Losfeld, E. Soetens, D. Bonatto, S. Fachada, L. Van Bogaert, G. Lafruit, and M. Teratani, "3D Tensor Display for Non-Lambertian Content," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2022, pp. 1–5.

[13] E. Soetens, A. Losfeld, D. Bonatto, S. Fachada, L. Bogaert, G. Lafruit, and M. Teratani, "Towards non-lambertian scenes for tensor displays," *London Imaging Meeting*, vol. 3, pp. 44–48, 07 2022.

[14] C. Gao, L. Dong, L. Xu, X. Liu, and H. Li, "Weighted Simultaneous Algebra Reconstruction Technique (wSART) for Additive Light Field Synthesis," *SID Symposium Digest of Technical Papers*, vol. 53, no. S1, pp. 243–246, 2022. [Online]. Available: https://sid.onlinelibrary.wiley.com/doi/abs/10.1002/sdtp.15905

[15] D. Chen, X. Sang, X. Yu, X. Zeng, S. Xie, and N. Guo, "Performance improvement of compressive light field display with the viewing-position-dependent weight distribution," *Optics Express*, vol. 24, no. 26, pp. 29 781–29 793, 2016. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?uri=oe-24-26-29781

[16] C. Gao, L. Dong, L. Xu, X. Liu, and H. Li, "Performance improvement for additive light field displays with weighted simultaneous algebra reconstruction technique and tracked views," *Journal of the Society for Information Display*, vol. 31, 02 2023.

[17] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 31–42. [Online]. Available: https://doi.org/10.1145/237170.237199

[18] T. Fujii, T. Kimoto, and M. Tanimoto, "Ray space representation for 3D image processing," in *Stereoscopic Displays and Virtual Reality Systems IV*, S. S. Fisher, J. O. Merritt, and M. T. Bolas, Eds., vol. 3012, International Society for Optics and Photonics. SPIE, 1997, pp. 330 – 336. [Online]. Available: https://doi.org/10.1117/12.274476

[19] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[20] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[21] G. Guennebaud, B. Jacob *et al.*, "Eigen v3," http://eigen.tuxfamily.org, 2010.

[22] Y. Qiu, "LBFGS++," https://https://github.com/yixuan/LBFGSpp, 2022.

[23] G. Wetzstein, "Synthetic Light Field Archive," https://web.media.mit.edu/gordonw/SyntheticLightFields/, Feb 2003.

[24] A. Losfeld, L. Van Bogaert, G. Lafruit, and M. Teratani, "ULB SauceDino," May 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7950729

[25] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items," 2022. [Online]. Available: https://arxiv.org/abs/2204.11918

Self-Supervised Super-Resolution Approach for Isotropic Reconstruction of 3D Electron Microscopy Images from Anisotropic Acquisition

Mohammad Khateri, Morteza Ghahremani, Alejandra Sierra, and Jussi Tohka
A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Finland
{mohammad.khateri, morteza.ghahremani, alejandra.sierralopez, jussi.tohka}@uef.fi

***Abstract*—Three-dimensional electron microscopy (3DEM) is an essential technique to investigate volumetric tissue ultrastructure. Due to technical limitations and high imaging costs, samples are often imaged anisotropically, where resolution in the axial direction (z) is lower than in the lateral directions (x, y). This anisotropy 3DEM can hamper subsequent analysis and visualization tasks. To overcome this limitation, we propose a novel deep-learning (DL)-based self-supervised super-resolution approach that computationally reconstructs isotropic 3DEM from the anisotropic acquisition. The proposed DL-based framework is built upon the U-shape architecture incorporating vision-transformer (ViT) blocks, enabling high-capability learning of local and global multi-scale image dependencies. To train the tailored network, we employ a self-supervised approach. Specifically, we generate pairs of anisotropic and isotropic training datasets from the given anisotropic 3DEM data. By feeding the given anisotropic 3DEM dataset in the trained network through our proposed framework, the isotropic 3DEM is obtained. Importantly, this isotropic reconstruction approach relies solely on the given anisotropic 3DEM dataset and does not require pairs of co-registered anisotropic and isotropic 3DEM training datasets. To evaluate the effectiveness of the proposed method, we conducted experiments using three 3DEM datasets acquired from brain. The experimental results demonstrated that our proposed framework could successfully reconstruct isotropic 3DEM from the anisotropic acquisition.**

***Index Terms*—self-supervised, super-resolution, electron microscopy, isotropic reconstruction, deep learning.**

I. Introduction

Three-dimensional electron microscopy (3DEM) enables the visualization and analysis of volumetric tissue ultrastructure at nanometer resolution. Achieving isotropic acquisition, where resolution is consistent in all dimensions, can assist downstream image analysis and visualization tasks. However, practical limitations, such as the constraints of EM techniques and imaging time and costs, often lead to achieving the resolution in the axial (z) direction lower than lateral (x, y) directions. Focused ion beam scanning EM (FIB-SEM) is one EM technique that can obtain isotropic 3DEM images with sub-10nm resolution in all directions; however, FIB-SEM is low-throughput. On the other hand, serial section transmission EM (ssTEM) or serial block-face scanning EM (SBEM) offers higher throughput and cost-effectiveness compared to FIB-SEM but cannot achieve the required axial resolution [1]. Image super-resolution (SR) is a computational approach that can increase the axial resolution to match lateral resolutions, enabling the reconstruction of isotropic 3DEM from anisotropic acquisitions.

Traditional SR approaches rely on interpolation methods, which can increase axial resolution. However, these methods have limitations in recovering fine missing details in low-resolution (LR) axial planes (xz/yz). To overcome these limitations, learning-based methods have been proposed that leverage prior knowledge about the latent data to the interpolation. One such method is sparse representation over learned dictionaries, which has been used in various SR applications [2], [3]. However, since dictionaries are learned from small image patches, they may not reconstruct high-quality EM images with large field-of-view. Authors in [4] proposed a dictionary-learning-based approach to reconstruct isotropic 3DEM by combining anisotropic 3DEM with sparse tomographic views of the same sample acquired at a finer axial resolution. While this approach offered a promising solution for isotropic reconstruction of 3DEM, it relies on the availability of both anisotropic and sparse tomographic views, which may not always be feasible.

Deep learning (DL) has emerged as a promising approach for SR in computer vision [7], medical [8], and biomedical [9] applications. DL-based methods follow an end-to-end learning procedure, enabling them to effectively learn the mappings from LR to high-resolution (HR) spaces when abundant LR and HR training datasets are available. The DL-based approach for isotropic 3DEM reconstruction from the anisotropic acquisition was introduced in [10], in which authors adopted a 3D convolutional neural network (CNN) architectures, then trained it using pairs of down-sampled isotropic 3DEM (synthetic anisotropic) and isotropic 3DEM acquired from FIB-SEM and tested on images obtained from the same technology. However, this approach has some limitations. Importantly, it requires the availability of isotropic 3DEM images at the desired resolution, which is often not feasible – especially in ssTEM and SBEM techniques. Additionally, when the network is fed with anisotropic 3DEM images acquired from a different technology, severe performance drops, and artifacts may occur due to the domain gap between EM imaging techniques.

This work was in part supported by the Academy of Finland (#323385), the Erkko Foundation, and the Doctoral Programme in Molecular Medicine at the University of Eastern Finland.

979-8-3503-4218-5/23$31.00©2023 European Union

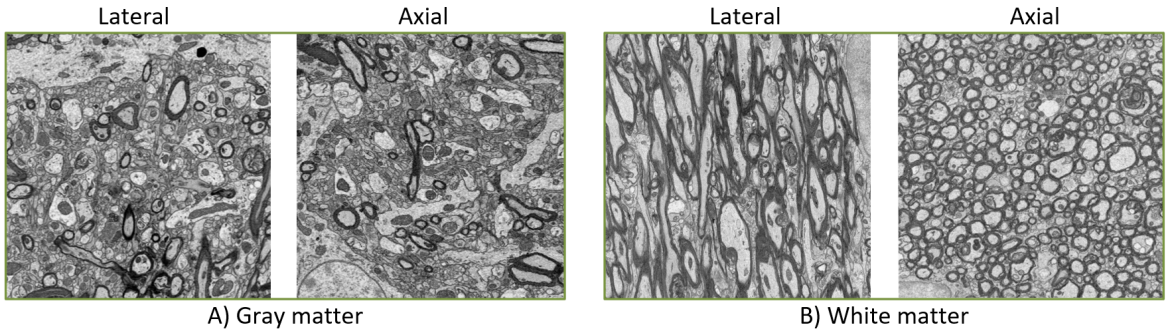


Fig. 1. Ultra-structural self-similarity in 3DEM datasets from the rat brain. A) Gray matter [5] demonstrates the ultra-structural self-similarity across a wide range of sizes, while B) white matter [6] predominantly exhibits this self-similarity in smaller structures.

Self-supervised super-resolution learning is a powerful technique that can eliminate the need for training datasets and address the domain gap between training and test datasets. It involves training super-resolution algorithms solely on the given LR image, using synthetically generated LR-HR training pairs derived from the LR image itself. Authors in [11] introduced the concept of self-supervised super-resolution learning, where they harnessed the internal recurrence of information inside a given LR natural image across different resolution scales to generate synthetic pairs of LR and HR image datasets. When the network is trained, the given LR image is fed to the network to produce the corresponding HR image. This approach has been employed within studies in the biomedical [12] and medical [13] domains to produce 3D isotropic images from the anisotropic acquisition, respectively, with the focus on the optical fluorescence microscopy and magnetic resonance imaging.

Motivated by the remarkable self-similarity observed in ultra-structures within brain 3DEM datasets, we present an efficient self-supervised super-resolution framework specifically designed to transform anisotropic 3DEM data into isotropic 3DEM, named A2I-3DEM. The key contributions of our work are as follows:

- We propose a framework for reconstructing isotropic 3DEM data from anisotropic acquisition while mitigating the inherent noise-like artifacts present in electron microscopy.
- We introduce a novel DL architecture based on the vision transformer, which effectively captures multi-scale local and global image dependencies, helping in enhanced reconstruction.
- We employ an efficient training strategy by simulating the distortions commonly observed in 3DEM imaging.

II. Method

Let $\boldsymbol{x} \in \mathbb{R}^{W\times W\times W}$ and $\boldsymbol{y} \in \mathbb{R}^{W\times W\times C}$ denote respectively isotropic and anisotropic 3DEM, where $\rho = W/C$ indicates the resolution ratio between isotropic and anisotropic acquisitions in the axial direction (z), i.e., super-resolution ratio. In this section, we introduce our ViT-empowered self-supervised super-resolution approach to reconstruct isotropic 3DEM, $\boldsymbol{x}$, from the anisotropic acquisition $\boldsymbol{y}$.

A. Self-Supervised Super-Resolution

Self-similarity of ultra-structures between lateral and axial planes in 3DEM data, especially in the brain gray matter, allows for self-supervised learning upon the anisotropic 3DEM data, see Fig 1. Leveraging such a structural self-similarity, we can synthesize training image pairs from the isotropic $xy-$lateral plane. To synthesize the training pairs, large patches that adequately represent the ultrastructural features of interest are extracted from the lateral plane $P^i_{xy} \in \mathbb{R}^{M\times M}$. These patches are then subjected to various degradations such as noise, artifacts, distortions, and anisotropic undersampling resolution with ratio $1/\rho$ to generate corresponding synthesized axial patches $P^i_{xz/yz} \in \mathbb{R}^{(M/\rho)\times M}$. The synthesized pairs $\{(P^i_{xz/yz}, P^i_{xy})\}_{i=1}^N$ are then used to train network $f_\theta(\cdot) : \mathbb{R}^{(M/\rho)\times M} \rightarrow \mathbb{R}^{M\times M}$ parameterized with θ to learn the mapping from axial to lateral planes. In practice, ρ may not always be an integer, which poses challenges to determining the mapping. To overcome this issue, we first employ interpolation to resize the anisotropic data to match the desired isotropic data size. This interpolated data is then utilized as the LR image. The network's parameters θ are obtained by optimizing the following empirical loss:

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(P^i_{xz/yz}), P^i_{xy}), \tag{1}$$

where $\mathcal{L}$ is the loss function between network prediction $f_\theta(P_{xz/yz})$ and ground truth P_{xy}. The trained network $f_\theta(\cdot)$ is then used to super-resolve the real axial planes to the desired resolution. Finally, by stacking the super-resolved axial planes in the perpendicular direction, the isotropic 3DEM is reconstructed. The proposed self-supervised super-resolution framework is illustrated in Fig.2.

B. Network Architecture

1) Overall Pipeline: The proposed network architecture is a hierarchical U-shaped design of the encoder-decoder equipped with ViT blocks, as illustrated in Fig.3. The input is a low-resolution axial plane image, $\mathbf{I} \in \mathbb{R}^{1\times H\times W}$, which is first

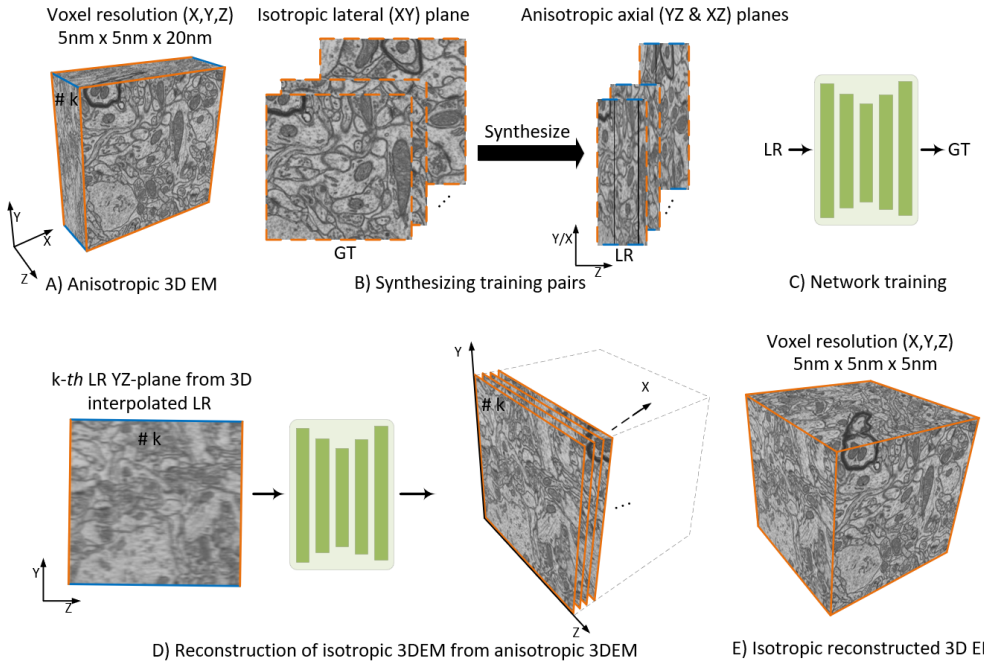


Fig. 2. The workflow of the proposed self-supervised super-resolution framework for the isotropic reconstruction of 3DEM from the anisotropic acquisition. A) Input 3DEM data is anisotropic, with high resolution in the lateral (x, y) directions and inferior resolution in the axial direction (z). B) Training pairs are synthesized from the anisotropic 3DEM data. The isotropic $xy-$lateral plane undergoes under-sampling and distortions to synthesize the $xz-$ and $yz-$anisotropic axial planes. C) The proposed network is trained using synthesized training pairs, where the interpolated synthesized axial plane is employed as LR input, while the isotropic lateral plane is regarded as GT. D) The trained network sequentially takes each axial plane as input, and the resultant outputs are stacked together to obtain isotropic 3DEM, involving two steps: Initially, 3D interpolation is employed to resize the anisotropic 3DEM data, aligning it with the size of the desire 3D isotropic data. Subsequently, the trained network is consecutively fed with each slice from the interpolated data's axial plane, and the resultant outputs are stacked together to generate isotropic 3DEM with an improved resolution in the axial direction. E) The output is an isotropic 3DEM with the improved resolution ratio ρ in the axial direction.

fed through convolutional layers to extract low-level features, $\mathbf{X_0} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W respectively indicate the number of channels, height, and width. Afterward, the feature map is passed through a symmetric encoder-decoder with K levels. Starting from the first encoder, the encoder hierarchically reduces the spatial resolution $(H \times W)$ while increasing the channel size, leading to the bottleneck feature map, $\mathbf{F}_\ell \in \mathbb{R}^{2^{K-1}C\times \frac{H}{2^{K-1}}\times \frac{W}{2^{K-1}}}$. The feature maps from the bottleneck and encoders are then passed to the decoders to progressively produce the high-resolution representation. Finally, the low-level features are added to the output from the last decoder, and fed with to the feature projection block, producing the super-resolved image.

2) Vision Transformer: The ViTs partition an image into a sequence of small patches, i.e., local windows, and learn relationships between them. By learning these relationships, the ViT can learn a wide range of image dependencies, which is crucial for achieving high performance in low-level vision tasks like image super-resolution. To capture both global and local image dependencies while keeping computational costs low, we employ the window-based multi-head attention (W-MSA) approach [14], [15]. The extracted attention maps using W-MSA are then passed through the novel gating mechanism, called the gated locally-enhanced feed-forward network (GLEN), to enhance the important features while suppressing the less important ones. These W-MSA and GLEN are embedded into a ViT block illustrated in Fig.3, and the corresponding computation is as follows:

$$\begin{aligned}\mathbf{X}' &= \textbf{W-MSA}(\mathbf{LN}(\mathbf{X})),\\ \mathbf{X}'' &= \mathbf{GLEN}(\mathbf{LN}(\mathbf{X}')) + \mathbf{X}',\end{aligned} \tag{2}$$

where, $\mathbf{LN}$ is layer normalization and $\mathbf{X}$ is the input feature map.

a) W-MSA: The input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ is firstly partitioned into $N = HW/M^2$ non-overlapping $M\times M$ local windows, leading to the local feature map $\mathbf{X}^i \in \mathbb{R}^{M^2\times C}$. The standard self-attention mechanism is then applied to each local feature map. The W-MSA, when there is k head with the dimension of $d_k = C/k$, is obtained by concatenating attention heads $\hat{\mathbf{X}}_k = \{\mathbf{Y}_k^i\}_{i=1}^N$, where $\mathbf{Y}_k^i$ is k-th head attention related to i-th local window calculated as below:

$$\mathbf{Y}_k^i = \textbf{Attention}(\mathbf{X}^i\mathbf{W}_k^Q, \mathbf{X}^i\mathbf{W}_k^K, \mathbf{X}^i\mathbf{W}_k^V), i = 1, \ldots, N, \tag{3}$$

where $\mathbf{W}_k^Q$, $\mathbf{W}_k^K$, $\mathbf{W}_k^V \in \mathbb{R}^{C\times d_k}$ are projection metrices of queries $(\mathbf{Q})$, keys $(\mathbf{K})$, and values $(\mathbf{V})$ for the k-th head, respectively. The attention is obtained as follows:

$$\textbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \textbf{SoftMax}(\frac{\mathbf{QK}^T}{\sqrt{d_k}} + \mathbf{B})\mathbf{V}, \tag{4}$$

where $\mathbf{B}$ is the relative position bias [16].

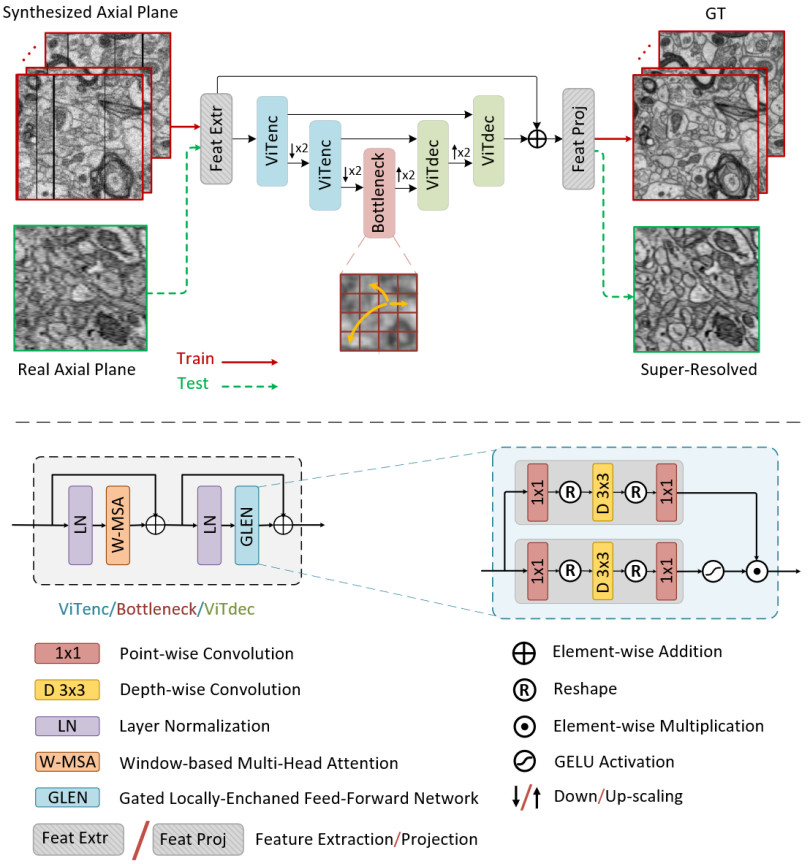


Fig. 3. The proposed U-shaped architecture based on the vision transformer. Training and testing are illustrated in the upper part of the figure, marked respectively in red and green. The bottom part of the figure visualizes the component of the proposed architecture.

b) GLEN: This block processes attention maps through two components: depth-wise convolution, which learns contextual image dependencies required for SR, and a gating mechanism, which highlights informative features while suppressing non-informative ones. As shown in Fig.3, the gating mechanism is implemented as the element-wise product of two parallel paths of linear transformation layers.

3) Loss Function: To optimize the network's parameters, we utilize the $\mathcal{L}_{\ell_1} = \frac{1}{N}\sum_{i=1}^{N}\|x_i - \hat{x}_i\|_1$ and projected distribution loss (PDL) [17], which respectively penalize pixel value and distribution mismatch between restored image $\hat{x}$ and ground truth x, ensuring both pixel-level accuracy and distribution-level fidelity. The total loss is given by:

$$\mathcal{L}_{Total} = \mathcal{L}_{\ell 1} + \alpha \mathcal{L}_{PDL}, \tag{5}$$

where α is a hyperparameter governing the trade-off between loss functions, which was empirically set to 0.01. For optimization, we employed the Adam algorithm [18] with an initial learning rate of 10^{-4}. The implementation was done using PyTorch framework.

III. Experiments and Results

A. Datasets

1) Synthetic Data: We synthesized an anisotropic 3DEM dataset by under-sampling an isotropic FIB-SEM dataset [19]. In the first step, to reduce noise and artifacts in data, we isotropically downscaled the original data– with voxel resolution $5\times5\times5\ nm^3$ and image size of $1530\times1530\times1053$– by a factor of three, resulting in a voxel resolution of $15\times15\times15\ nm^3$. Subsequently, we applied anisotropic downsampling to achieve a voxel resolution of $15\times15\times45\ nm^3$. These synthetic pairs of anisotropic and isotropic 3DEM datasets were utilized in our experiments.

2) Real Data: We used two anisotropic 3DEM datasets acquired from rat brains through the SBEM technique. The first dataset was acquired from the gray matter in the visual cortex [5] with the size of $1024\times1024\times540$, while the second was acquired from the white matter at the corpus callosum [6], with the size of $1024\times1024\times490$. Both datasets had a voxel resolution of $15\times15\times50\ nm^3$.

B. Results

We compared the proposed super-resolution method, A2I-3DEM, with several established techniques, including the standard cubic interpolation approach as well as two CNN-based methods: SRMD [20] and PSSR [9]. Additionally, we considered a transformer-based method, SwinIR [15]. For synthetic data, we utilized PSNR and SSIM [21] for quantitative assessments and visually compared the super-resolved

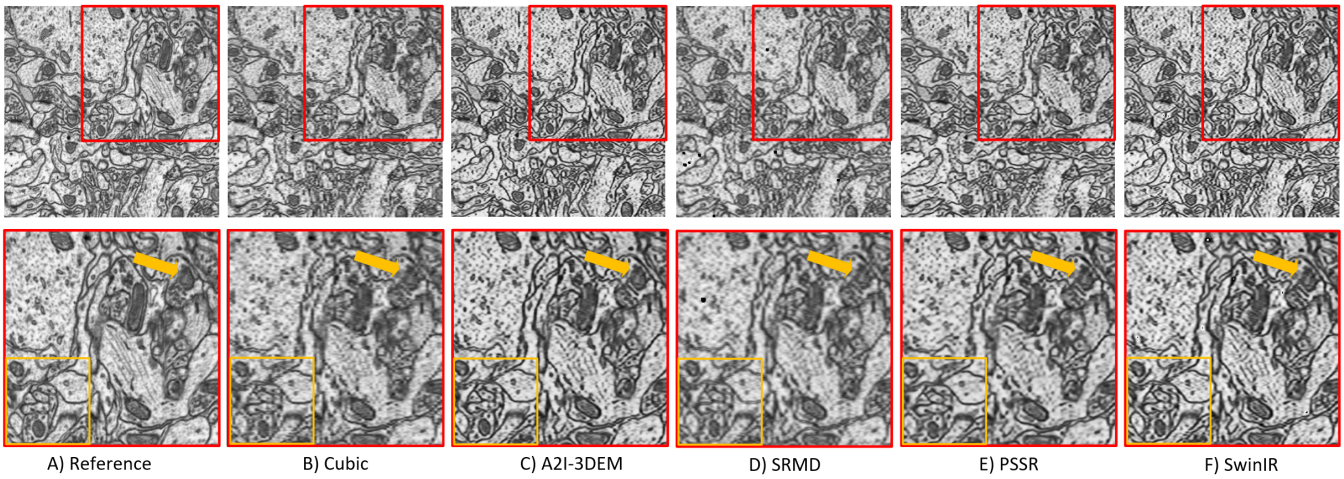


Fig. 4. Visual comparison of isotropic 3DEM reconstruction results using various methods on the synthetic dataset: $xz-$axial plane perspective.

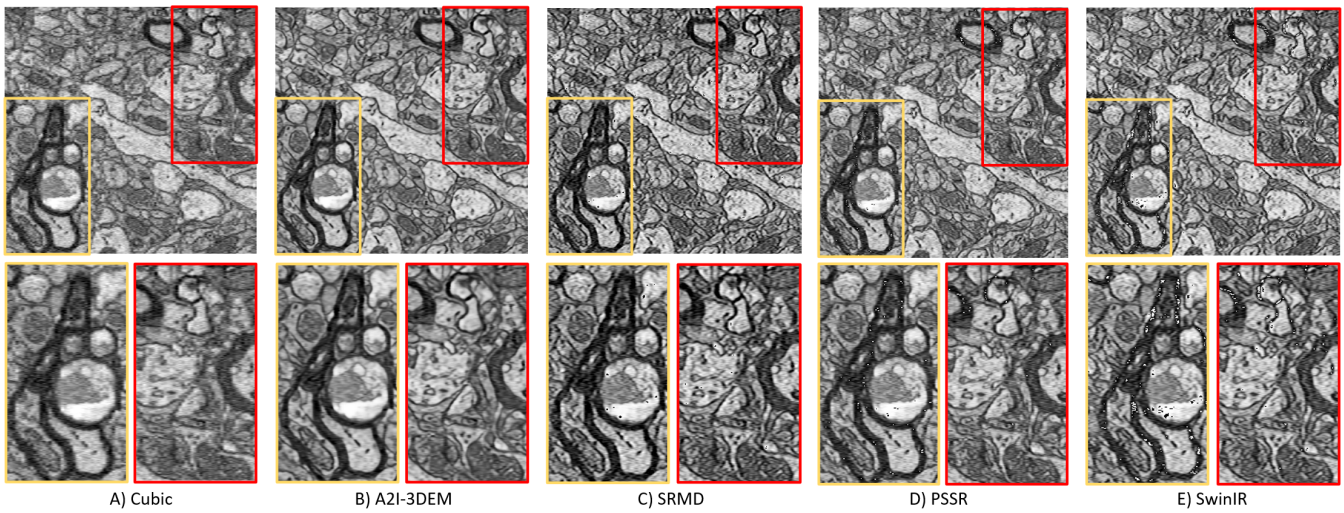


Fig. 5. Visual comparison of isotropic 3DEM reconstruction results using various methods on the real dataset from gray matter: $xz-$axial plane perspective.

volumes with the reference. For real data, lacking a reference, we visually compared the results with the Cubic interpolated data, the initial point for all competitors, to assess resolution enhancement and consistency of details.

For the synthetic dataset, where we have the reference, a visual comparison with competitors is drawn in Fig. 4, and corresponding quantitative results were tabulated in TableI. In Fig. 4, orange restricted areas show that cubic and SRMD led to severely blurred results. Among other methods, A2I-3DEM and SwinIR could produce images with better contrast and distinguishable membranes. Notably, as pointed out by the arrows, A2I-3DEM outperforms SwinIR by producing outputs with reduced blurriness. The superiority of A2I-3DEM is in agreement with the PSNR value reported in TableI. However, SSIM values contradict the visual outcomes, as the cubic interpolation method appears to outperform all other competitors according to SSIM. This discrepancy calls for an alternative image quality assessment metric.

TABLE I
QUANTITATIVE COMPARISONS OF ISOTROPIC 3DEM RECONSTRUCTION ON THE SYNTHETIC DATASET. THE BEST METRIC VALUE FOR EACH METHOD IS MARKED IN BOLD.

Metric	**Method**				
	Cubic	SRMD	PSSR	SwinIR	A2I-3DEM
PSNR	28.15	29.16	29.11	30.57	**30.61**
SSIM	**0.698**	0.644	0.675	0.632	0.645

Visual comparison of the first real dataset, pertaining to brain gray matter, is presented in Fig. 5. Consistent with expectations, DL-based methods demonstrate enhanced detail compared to cubic interpolation. Zooming in on specific regions in Fig. 5 (B-E), artifacts such as black point artifacts in white areas or white point artifacts in black areas are evident in the results of SRMD, PSSR, and SwinIR. In contrast, A2I-3DEM not only avoids these artifacts but also successfully reduces noise compared to the other methods.

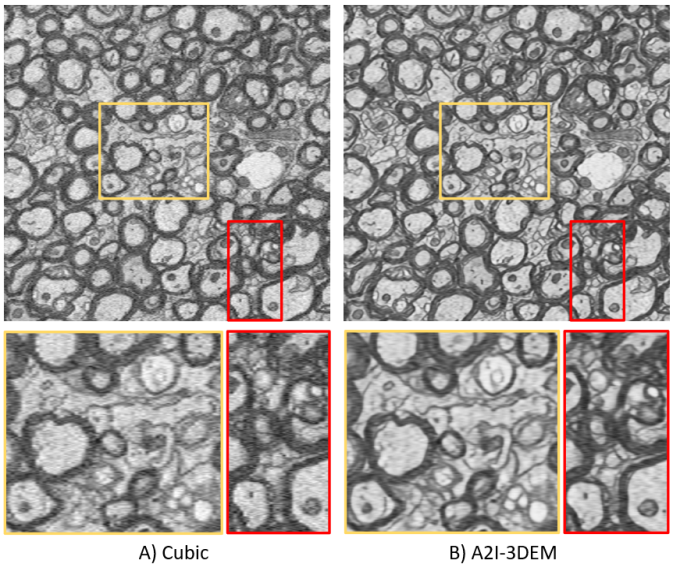


Fig. 6. Isotropic 3DEM reconstruction of real dataset from white matter: $xz-$axial plane perspective.

A subset of visual results from the second real dataset, related to brain white matter, is depicted in Fig. 6. These results highlight the success of our proposed self-supervised method in enhancing the resolution of the given LR image while effectively mitigating noise.

IV. Conclusion

This paper introduced a deep-learning-based self-supervised super-resolution framework to overcome the challenge of acquiring isotropic 3DEM. The framework's ability to generate training datasets directly from the provided anisotropic 3DEM data makes it a practical preprocessing tool for downstream visualization and processing tasks. The incorporation of simulated distortions within the efficient training strategy not only improved the model's generalizability but also enabled the network to learn to mitigate noise that exists in the given LR EM image. Furthermore, the proposed U-shaped architecture, equipped with ViT blocks, effectively captures multi-scale local and global image dependencies, leading to enhanced reconstruction performance. Experimental evaluations conducted on 3DEM datasets of brain tissue demonstrated the network's proficiency in recovering fine details while effectively mitigating noise.

Acknowledgment

We thank CVLab at École Polytechnique Fédérale de Lausanne for sharing their 3DEM dataset, Electron Microscopy Unit of the Institute of Biotechnology at University of Helsinki for rat datasets, and the Bioinformatics Center at University of Eastern Finland, for providing computational resources.

References

[1] S. Mikula, "Progress towards mammalian whole-brain cellular connectomics," *Frontiers in neuroanatomy*, vol. 10, p. 62, 2016.

[2] P. Song, X. Deng, J. F. Mota, N. Deligiannis, P. L. Dragotti, and M. R. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 57–72, 2019.

[3] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[4] T. Hu, J. Nunez-Iglesias, S. Vitaladevuni, L. Scheffer, S. Xu, M. Bolorizadeh, H. Hess, R. Fetter, and D. Chklovskii, "Super-resolution using sparse representations over learned dictionaries: Reconstruction of brain structure using electron microscopy," *arXiv preprint arXiv:1210.0564*, 2012.

[5] R. A. Salo, I. Belevich, E. Manninen, E. Jokitalo, O. Gröhn, and A. Sierra, "Quantification of anisotropy and orientation in 3d electron microscopy and diffusion tensor imaging in injured rat brain," *Neuroimage*, vol. 172, pp. 404–414, 2018.

[6] A. Abdollahzadeh, I. Belevich, E. Jokitalo, A. Sierra, and J. Tohka, "Deepacson automated segmentation of white matter in 3d electron microscopy," *Communications biology*, vol. 4, no. 1, p. 179, 2021.

[7] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, "Real-world single image super-resolution: A brief review," *Information Fusion*, vol. 79, pp. 124–145, 2022.

[8] Y. Sui, O. Afacan, C. Jaimes, A. Gholipour, and S. K. Warfield, "Scan-specific generative neural network for mri super-resolution reconstruction," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1383–1399, 2022.

[9] L. Fang, F. Monroe, S. W. Novak, L. Kirk, C. R. Schiavon, S. B. Yu, T. Zhang, M. Wu, K. Kastner, A. A. Latif, *et al.*, "Deep learning-based point-scanning super-resolution imaging," *Nature methods*, vol. 18, no. 4, pp. 406–416, 2021.

[10] L. Heinrich, J. A. Bogovic, and S. Saalfeld, "Deep learning for isotropic super-resolution from non-isotropic 3d electron microscopy," in *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pp. 135–143, Springer, 2017.

[11] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3118–3126, 2018.

[12] M. Weigert, L. Royer, F. Jug, and G. Myers, "Isotropic reconstruction of 3d fluorescence microscopy images using convolutional neural networks," in *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pp. 126–134, Springer, 2017.

[13] C. Zhao, B. E. Dewey, D. L. Pham, P. A. Calabresi, D. S. Reich, and J. L. Prince, "Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 805–817, 2020.

[14] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693, 2022.

[15] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[17] M. Delbracio, H. Talebi, and P. Milanfar, "Projected distribution loss for image enhancement," *arXiv preprint arXiv:2012.09289*, 2020.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] https://www.epfl.ch/labs/cvlab/data/data-em/.

[20] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3262–3271, 2018.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

Using Deep Generative Models for Glossy Appearance Synthesis and Exploration

Abhinav Reddy Nimma
Colourlab, Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
abhinavn@stud.ntnu.no

Davit Gigilashvili
Colourlab, Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
davit.gigilashvili@ntnu.no

***Abstract*—Generating images with realistic material appearance using a physically-based renderer demands significant time and human labor. The images are used in psychophysical experiments to study human perception of material appearance attributes, such as glossiness. Recently, deep learning-based image synthesis models have emerged as a promising approach for generating realistic images with less human supervision. Deep Generative Models are deep learning-based models that learn to generate unique and novel images based on a given training data distribution. Using them for image synthesis is fast and manually less tiresome. An additional benefit these Deep Generative Models offer is latent space encodings that may help to better understand the feature space of gloss and its perception. In this study, we propose to explore the possibility of using Deep Generative Models for realistic image synthesis, focusing on gloss appearance and evaluating the efficiency of such gloss generation process using psychophysical experiments. Additionally, we build tools to extract the latent space of generative models to use them as a feature space representation of gloss appearance and perception. Finally, we analyse the trends and patterns in the learnt feature space to aid gloss appearance modelling.**

***Index Terms*—Gloss Perception, Image Synthesis, Material Appearance Modelling, Learning a Feature Space Representation**

I. Introduction

Perception of material appearance and its properties is fundamental to humans for interacting with the environment. The human visual system (HVS) has complex and sophisticated mechanisms for appearance perception that are a product of millions of years of evolution and remain poorly understood [1], [2]. Gloss – together with color, texture, and translucency – is one of the fundamental attributes of how objects and materials look [3]. Although gloss is primarily understood as a surface reflectance property, the link between instrumentally measured and human perceived gloss is complex and non-monotonic [4], [5]. Multiple handcrafted features have been proposed to predict gloss appearance from image statistics [6]–[8], but handcrafted features are rarely robust enough to account for complex influences from shape, illumination, and observation geometry [9]–[11].

Perceptual studies often involve computer graphics to generate the experimental stimuli. The process of rendering images with glossy surfaces involves understanding the complex interactions between all the intrinsic (optical properties) and extrinsic (environmental) factors. Most images generated using physically-based renderers are labelled using the physical parameter values. This does not help us to understand how the human visual system deciphers gloss appearances and how each factor influences gloss perception in humans. We need a better representation for navigating the gloss appearance space. It is not easy to handcraft features for human gloss perception as it is not fully understood how the human visual system deciphers gloss appearance into individual factors [2], and more efficient feature space is needed. Apart from that, using a physically-based renderer (such as Mitsuba [12]) is both very time-consuming and human labor-intensive. It would be desirable to develop a way to render or generate images with a realistic gloss appearance that requires minimal supervision.

Deep Generative Models have shown promising results in generating realistic images. Image synthesis in deep learning refers to generating images using neural networks. Deep Generative Models are based on deep learning. They learn to generate novel images based on a training data distribution. They first learn to model the distribution in the images in the training data and then use the learnt patterns to generate novel images that are not part of the training dataset. Deep Generative Models are considered unsupervised as they neither need manual supervision during training nor annotations for the data they are being trained on. The learning process is data-driven, i.e., the models learn to form the given data without needing any target labels for the given data. They have demonstrated capabilities in generating realistic novel images that are not part of the training data. If we can generate realistic material appearance using Deep Generative Models, it would save us significant amounts of time and labor. Deep Generative Models try to develop an understanding of the statistical structure in the data distributions. In developing this understanding, Deep Generative Models develop a latent space representation for the data distribution. Thus, apart from aiding in generating images, they also help us encode images into a new latent space. The latent space of these models can be used as a representational space for material appearance attributes.

The models encode the input image into its internal latent space and then decode the latent vector from its internal latent space into output images. During training, the model optimises this encoding and decoding process and learns to model the

979-8-3503-4218-5/23/$31.00 © 2023 IEEE

statistical structure in the data distribution of the input images in its internal latent space. This way, in an unsupervised manner, we end up with a new feature space representation of the images in the training dataset. We can use this new feature space to better understand the dataset. It is believed that the HVS exploits statistical structure and regularities in the environment to derive information about our surroundings and develop perception and awareness of the world [13]. The development of latent space in Deep Generative Models is similar, and it is hypothesized that such feature space can eventually be used to model the perception of the HVS.

In this work, we trained a Deep Generative Model with low number of physically-based renderings of glossy objects and synthesized novel images with this model to check whether it can produce realistic images. We report the results of a psychophysical experiment that we conducted to assess the convincingness of the synthesized images. Afterward, we explore the latent space to understand the feature space of gloss and navigate through it in a meaningful manner.

II. RELATED WORK

Several attempts have been made in developing a feature representation for material appearance for surface gloss [7], [14], surface roughness [15], [16], transparency [17], [18], and translucency [19]–[21]. The studies use an analytical approach to find diagnostic image features for material perception. There is a significant challenge in this approach, since the features may not be stable across a broad range of intrinsic and extrinsic factors [1], [19]. An alternative approach in the diagnosis of features for material appearance is a data-driven one [22], [23]. These approaches attempt to extract features of material appearance by modeling the statistical distribution of material appearance across image samples. This approach has demonstrated great potential in modeling human perception [24]. Especially with the rapid progress of deep neural networks to learn patterns from enormous and diverse datasets, data-driven approaches show a significant potential in perception modeling [25]–[27]. Convolutional neural networks can be used to extract features from the images.

For long, deep learning-based techniques were used to analyse images for content objects etc. Recently, with the advancements in deep learning-based techniques, neural networks can generate images from random noise [28], seed [29], or text inputs [30], with remarkable realism. These networks can learn an image generation procedure from the training dataset's images. During training, they model the statistical structure in the distribution of images in the training set and construct an internal latent space representation for all the images in the training dataset. With models that generate accurate, realistic images, the internal latent space can be extracted and used as an efficient and compact feature representation of the distribution of images in the training dataset.

Generative Adversarial Networks (GANs) [31] is a breakthrough architecture on which most of the state-of-the-art Deep Generative Models are based. GANs consist of two deep neural networks: a discriminator and a generator. The task of the generator is to generate images from random input vectors, similar to the training data distribution. Discriminator judges whether the image presented is from the training data distribution or the generator generates it. This way, the generator is forced to get better at synthetic image generation.

StyleGANs can generate various styles at high-resolution [32] and also be able to control the styles in the generated images. For instance, Celeb-A dataset is a collection of high-resolution images of the faces of celebrities. StyleGAN was trained on this dataset. One can fine-tune the faces generated by the model as one wishes. Using the learned inputs to the network, one could control the face's sharpness, the eyebrows' width, and the hair's color. This way, StyleGANs were able to perform high-resolution image synthesis. However, StyleGANs still suffered from multiple issues, like water droplet artefacts and shift-invariance. Blob artefacts have been found in images generated by StyleGANs. StyleGAN2 [29] and StyleGAN2-ADA [33] propose some improvements to tackle these issues. Although StyleGAN2 has solved the issue of high-resolution image synthesis, the problem of requiring enormous-sized datasets to train GANs persists. StyleGAN2-ADA solves the issue of having large datasets and provides a way to train deep generative models on little data [33]. ADA stands for Adaptive Discriminator Augmentation. StyleGAN2 makes use of Adaptive Discriminator Augmentation instead of Stochastic Discriminator Augmentation. This way, StyleGAN2-ADA provides a way to train image synthesis models with limited data.

Some attempts have been made to construct a feature space for material appearance based on deep learning-based models' internal latent space embedding. Storrs *et al.* [24] used Variational Autoencoder (VAE) to model the distribution in images with gloss and matte surfaces . The study has shown that the image features from the internal latent space encoding of trained VAE models correlate well with human gloss perception and even mimic the mistakes that humans make in gloss judgments.

Generative Adversarial Networks (GANs) show improvements over VAEs in realistic image synthesis. Liao *et al.* [34] have generated realistic images of translucent objects with GANs and noticed that structured perceptual attributes emerge in the model's representation. They suggest that Deep Generative Models can discover an efficient and compact feature representation space for material appearance and can be potentially used to mimic the perception model of the HVS.

III. METHODOLOGY

Building upon the literature, we propose to train StyleGAN2-ADA [33] on physically-based renderings of glossy objects. We then evaluate the realism of images generated by the trained model, build tools to encode images into the latent space of the trained model and vice versa, build tools to traverse and analyse the feature space representation to check for gloss appearance attributes and analyse the usability of such feature space in aiding understanding of gloss appearance.

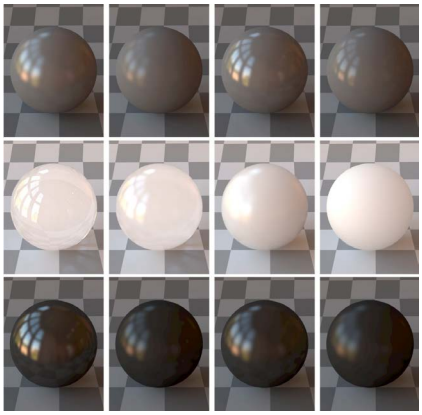


Fig. 1. Some samples from the training dataset.

A. StyleGAN2-ADA

StyleGAN2-ADA [33] is a Generative Adversarial Network designed by researchers at NVIDIA. The implementation provided by NVIDIA in the official GitHub repository is used for all the experiments (https://github.com/NVlabs/stylegan2-ada-pytorch). No specific changes have been made to the network architecture and training procedures. StyleGANs do not use the latent space directly. They first map these latent vectors into an extended latent space before generating an image. In the latent space Z, z is a 512 feature vector. Seed is the number used to generate this 512 feature vector. Then this latent vector is mapped into the extended latent space *W*. A vector w ($w \in W$) is of dimensions $1 \times 14 \times 512$. StyleGAN2-ADA applies data augmentation after the input component for both the generator and the discriminator. StyleGAN2-ADA solves the issue of collecting images to create large-scale datasets. It involves flipping the images, rotating them by a small angle, and zooming in on the image, among others.

B. Dataset

We used 132 physically-based renderings of glossy spherical objects rendered with Mitsuba [12] (can be accessed at https://github.com/davitgigilashvili/GANs4GlossEUVIP). The objects vary in surface roughness, lightness, and translucency – covering a broad range of gloss appearances. To increase the size of the dataset, we performed the augmentations by rotating the image by 90, 180 and 270 degrees, thus quadrupling the size of the dataset to 528 images. The examples of the images that were used for training are shown in Fig. 1.

C. Training

We use model weights from the pre-trained model on the (Flickr-Faces-HQ) FFHQ dataset [29] and transfer learning to train StyleGAN2-ADA to generate images with a realistic gloss appearance. We train the model for 5000kimg (i.e. how many images are evaluated; 528 $\times$ number of epochs). Training such an advanced GAN like StyleGAN2-ADA requires much computational power. We have used two NVIDIA TITAN RTX GPUs to run all our experiments. We train the model to generate images with a resolution of 256×256 pixels. The batch size used for training the model is 32, parallelised over two GPUs. A learning rate of 0.0025 is used for the transfer learning process. It took one day, 17 hours and 42 minutes to train the StyleGAN2-ADA model for 5000 kimg.

D. Image Synthesis

In StyleGAN-based architectures, a mapping network is used to map vectors from latent space Z to extended latent space W. These latent vectors w are directly plugged into the various layers of the network, thus giving us direct control to alter the styles in the images being generated. Since we do not have any understanding of the latent space of the model, to explore this latent space, we need to sample the feature space randomly. To do this, we randomly generate latent vectors from the space. Most random number generators are built on algorithms that start with a base value as an input known as a seed. For the same seed, we always get the same output random value. This helps us to lock random vectors across the experiments. We use seed values from 0 to 2000 and generate corresponding images using the trained StyleGAN2-ADA network. The first step in generating images from the seed involves generating latent vector z from the seed. Later, the latent vector z (1×512 feature space) is mapped into the extended latent space W. The resulting vector w ($w \in W$) is fed to the generator of StyleGAN2-ADA to generate images.

E. Evaluation

We evaluate the images using two methods. The first one involves using an image quality metric called Frechet Inception Distance (FID), which is a popular method to compare real and synthetic images [35]. We calculate FID after every 400 epochs, 50k images are generated from randomly sampling the latent space. FID is calculated on these 50k images by comparing them to the images in the training set.

The second method to evaluate performance was psychophysical experiment, which was hosted at the online QuickEval [36] platform. 19 observers participated in the experiment – mostly researchers and graduate students with substantial knowledge of graphics and appearance. In total, the observers were shown 60 images, 30 real images and 30 synthetic images. The real images were selected from the training set. Some of the synthetic images were those that were trying to mimic the respective real ones, while others corresponded to the random vectors from the latent space. The observers were asked to judge whether the image was real or synthetic. We explained to them that *Real* means that the images were generated using physically-based rendering with human supervision, while *Synthetic* ones were produced by GANs without human supervision. They were instructed to judge the realism of the images solely based on the realism of the gloss on the surface of the sphere.

F. Latent Space Exploration

We use the algorithm discussed above to generate W space latent vectors for all the images in the training dataset. The latent vector z ($z \in Z$) is of size 1×512, and the extended

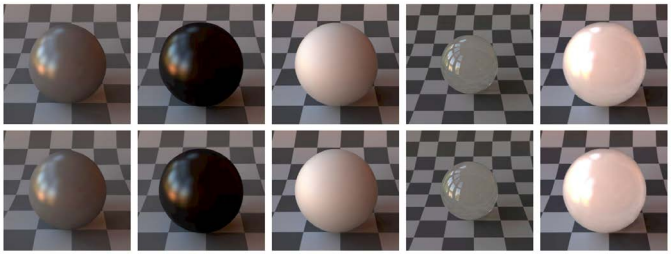


Fig. 2. The first objective is the synthesis of the realistic images. The original images are shown in the top row. They are projected into the extended latent space W. Synthetic images generated from the corresponding w latent vectors are shown in the bottom row that look highly similar to those in the top row.

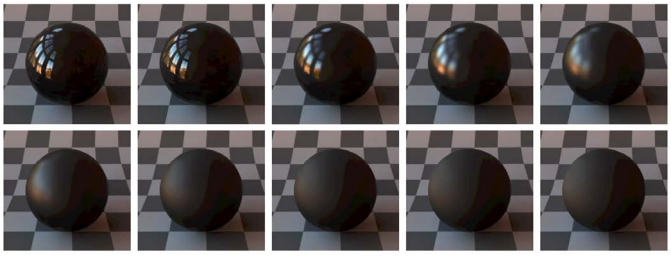


Fig. 3. Interpolations of images (performed in the latent W space) generated between the two target images shown on the left and right of each row.

latent space vectors w are of the size $1 \times 14 \times 512$. We have generated the corresponding latent vector w for every input image in the training dataset. We then use this latent vector w to generate the image. This generated image is referred to as a fake image. The original image is referred to as a real image. The examples are illustrated in Fig. 2.

We perform linear interpolation between the latent space encodings in the W space. To generate interpolations between *Image A* and *Image B*, we first find the latent space encodings of these two images in GAN's latent space. We then perform linear interpolation between the two corresponding latent codes generating a set of new latent codes. We then generate images from these interpolated latent codes. In other words, we can morph between two target images to generate interpolations between these two images. Fig. 3 demonstrates that the interpolations in the latent space W look perceptually meaningful, which indicates that the space is well-developed.

We also explored the directions in the latent space. Exploring directions in the latent space means moving along a specific dimension of the feature space and seeing how it affects the resulting images generated. In this experiment, we limit the directions to primary dimensions in the data, i.e. if the latent space has 512 dimensions, we explore along these 512 directions only. This is a simple algorithm developed from scratch by us to traverse through the latent space of the models. However, there is a significant limitation here. We are only exploring the directions along the primary dimensions. What about the direction with a slope of 45 degrees with the two primary directions? The possible directions are infinite in the data. This can be addressed in future works.

Shen *et al.* [37] propose closed form factorisation, a simple and efficient way to explore latent semantics in GANs to identify interpretable dimensions in the latent space of GANs and to extract the underlying patterns. The algorithm identifies semantically meaningful directions in the latent space by decomposition on the model weights. The output of closed-form factorisation is eigenvectors corresponding to the largest eigenvalues that maximise the objective function. The objective function is to find the directions in the latent space of GANs that reveal explanatory factors. Once we have extracted the interpretable directions in the latent space, the next step is to traverse through these directions to check how each direction impacts the style of the generated images.

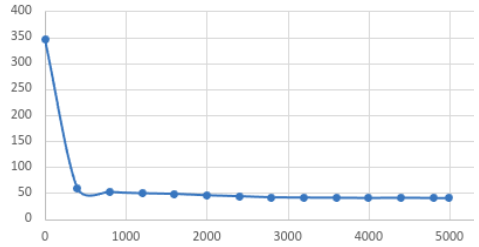


Fig. 4. FID score of images generated (vertical) vs epochs trained (horizontal).

IV. RESULTS

A. Evaluation

Fig. 4 shows how the FID score changes across epochs. As mentioned earlier, a smaller FID score implies that the images generated are closer to the images used for training and thus more realistic. This is a decent score, considering that it is evaluated on 50,000 images randomly sampled from the latent space. By increasing the number of images used for training, we can lower the FID score and thus improve the realism in the images generated. The results of the psychophysical experiments are shown in Table I. 69.02 % of the times observers judged real images as real and 30.98 % of the times observers judged real images as synthetic. When it came to synthetic images, 53.53 % of the times observers judged synthetic images as synthetic and 46.48 % of the times observers judged synthetic images as real. This implies that it was difficult for observers to assess if the images shown were real or synthetic and shows the potential of our models to generate realistic images that can trick humans.

B. Interpretable Directions

We have extracted 512 directions from the latent space and traverse through them. In total, for images generated from seeds 0 to 2000, we have generated the images by moving 5, 10, -5, -10 steps in each of the 512 directions exploited from the latent space. It is not manually possible to analyse all the images extracted, neither fits it within the scope of this paper. Hence, we show some of the significant directions extracted from closed form factorisation. From Fig. 5, we can see that by moving in the direction of the first interpretable direction, we can control the surface roughness and hence, glossiness on the sphere. This way by extracting interpretable

TABLE I
THE RESULTS OF THE PSYCHOPHYSICAL EXPERIMENT. OBSERVERS FOUND IT CHALLENGING TO DISTINGUISH REAL AND SYNTHETIC IMAGES.

	Judged Correctly	Judged Incorrectly
Real	69.02%	30.98%
Synthetic	53.52%	46.48%

directions, we can control the styles in images generated by our StyleGAN2-ADA model. We can see that, the surface roughness changes, making the spheres appear less glossy and more translucent. As the surface becomes smoother, we see that the spheres appear more glossy and less translucent. This is an interesting interaction between translucency and glossiness that automatically appears in the latent space of the model without any human supervision. From Fig. 6 we can see that when moving in the direction of the second extracted direction, we alter the style of translucency in the resulting images. The level of glossiness is more or less constant, but the level of translucency changes. This is very interesting, cause moving in the first direction altered both gloss and translucency in an inversely proportional relation, but moving in the second direction only alters translucency without altering gloss. From Fig. 7 we can see that when moving in the third interpretable direction, we alter the size of the sphere in resulting images. Specular highlights also change slightly, but the change in size is more apparent. Thus, using the extracted directions, we can alter the desired styles like glossiness, translucency or size of the sphere in resulting images. Analysing more directions would give us more control over the appearance attributes and style in synthesized images.

This is a baseline study to demonstrate that the approach can produce realistic images with very limited training set and to make first steps toward explainability. The work has limitations that will be addressed in future works. While fine tuning works for many cases, future work can explore potential changes in the architecture as well as training from scratch on a more specific dataset. Currently we have 512 dimensions that are perceptually non-uniform and exhibit cross-contaminations among perceptual attributes (e.g. size and gloss can change in the same dimension). Dimensionality reduction techniques, such as PCA, can be used to reduce dimensionality of the space from 512 to more manageable and perceptually meaningful dimensions, and psychophysical experiments will be needed to scale each dimension. Besides, we can use differentiable rendering to map the latent space back to the optical properties [38]. In addition to FID, future works can use perceptual loss-based methods for evaluating the results. Finally, although the approach is generalizable, the generated images are limited by the training dataset that the model was exposed to (e.g. single shape and environment map). Future works should include more diverse training datasets with more shapes, materials, and lighting conditions.

V. CONCLUSION

In this study, we have explored two things: 1) the potential of Deep Generative Models for generating images with realistic glossy surfaces from a limited training dataset; and 2) the usability of internal latent space of Deep Generative Models as a compact feature representation space for gloss appearance and perception. We trained StyleGAN2-ADA model to generate images of spheres with realistic glossy surfaces. We built the tools to generate the images from seeds, from z and

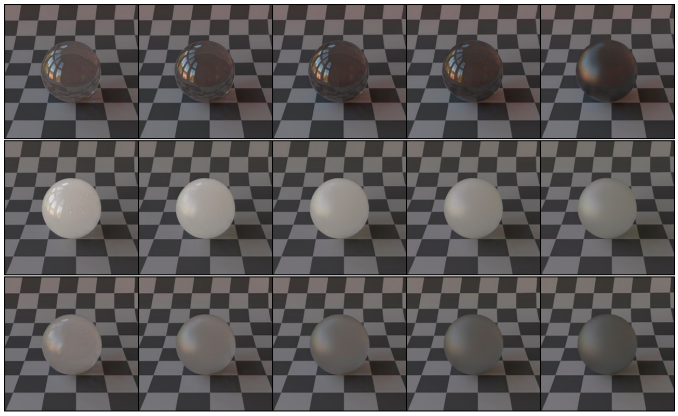


Fig. 5. Seed 6, 7, and 10 (from top to down, respectively). Moving in the direction of first interpretable direction (the direction with largest eigen value). From left to right, 10 steps in positive direction, 5 steps in positive direction, image from seed, 5 steps in negative direction, 10 steps in negative direction.

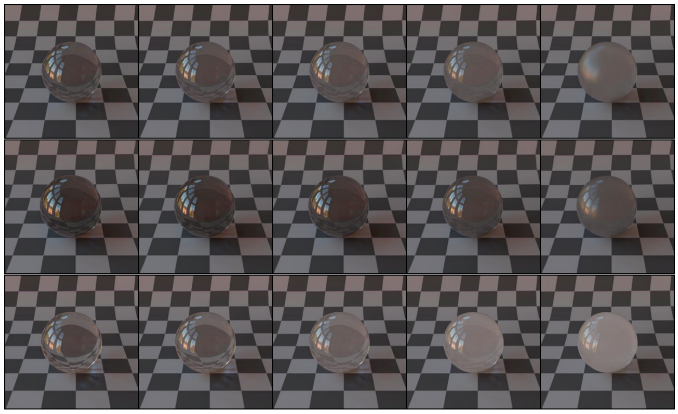


Fig. 6. Seed 1, 6, 13. Moving in the direction of second interpretable direction.

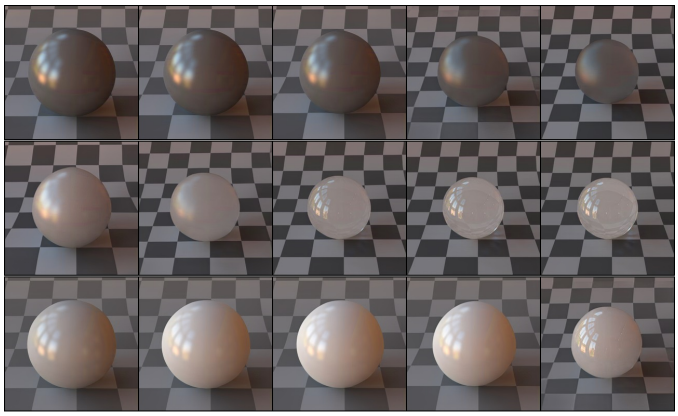


Fig. 7. Seed 1, 15, 16. Moving in the direction of third interpretable direction.

w latent vectors. We have also built the tools to map images to and from the internal latent space of StyleGAN2-ADA. We then analysed usability of this latent space as a feature space for gloss appearance and perception by extracted interpretable directions from the latent space and moving in these directions. It can be seen from our experiments and results that the images generated by StyleGAN2-ADA trick human observers into thinking that these were actually generated by human supervision in a physically based renderer. The results also show that interesting interactions between gloss and translucency emerge in the latent space of the trained model. This space can be used to find relevant features for visual perception of gloss. From linear interpolations between images, we can also see that the latent space is quite well developed. However, there are some limitations – some visual artifacts emerge due to a small dataset size. This implies that the latent space of the model contains some information gaps. Nevertheless, this shows the potential of using Deep Generative Models to generate images with realistic glossy surfaces even with a limited training set and also the potential of latent space of these models to be used as an efficient feature space for gloss appearance. It is known that in neural networks, the initial layers of the model are responsible for constructing low level features, and the final layers of the model are responsible for constructing higher level features. As a future work, the feature space can be further studied to understand which layers of the model influence what parameters of gloss in the synthesized images. Also, psychophyscial experiments need to be conducted to study how human perception correlates with the trends and patterns emerged in the latent space. Overall, using Deep Generative Models for realistic glossy image synthesis shows promising results and certainly merits future research.

REFERENCES

[1] R. W. Fleming, "Visual perception of materials and their properties," *Vision Research*, vol. 94, pp. 62–75, 2014.

[2] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?" *J. Vis.*, vol. 9, pp. 784–784, 2010.

[3] CIE, *CIE 175:2006 A framework for the measurement of visual appearance*. International Commission on Illumination., 2006.

[4] A. C. Chadwick and R. Kentridge, "The perception of gloss: A review," *Vision research*, vol. 109, pp. 221–235, 2015.

[5] F. B. Leloup, G. Obein, M. R. Pointer, and P. Hanselaer, "Toward the soft metrology of surface gloss: A review," *Color Research & Application*, vol. 39, no. 6, pp. 559–570, 2014.

[6] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson, "Image statistics and the perception of surface qualities," *Nature*, vol. 447, no. 7141, pp. 206–209, 2007.

[7] F. Pellacini, J. A. Ferwerda, and D. P. Greenberg, "Toward a psychophysically-based light reflection model for image synthesis," in *Proceedings of the ACM SIGGRAPH 2000*, 2000, pp. 55–64.

[8] J.-B. Thomas, J. Y. Hardeberg, and G. Simone, "Image contrast measure as a gloss material descriptor," in *Computational Color Imaging: 6th International Workshop*. Springer, 2017, pp. 233–245.

[9] M. Lagunas, A. Serrano, D. Gutierrez, and B. Masia, "The joint role of geometry and illumination on material recognition," *Journal of Vision*, vol. 21, no. 2:, pp. 1–18, feb 2021.

[10] M. Olkkonen and D. Brainard, "Joint effects of illumination geometry and object shape in the perception of surface reflectance," *i-Perception*, vol. 2, pp. 1014–34, 12 2011.

[11] D. Gigilashvili and A. J. Islam, "The role of shape in modeling gloss," *30th Color and Imaging Conference (CIC30)*, pp. 271–276, 2022.

[12] W. Jakob, "Mitsuba Renderer," 2010, http://www.mitsuba-renderer.org.

[13] K. Storrs and R. Fleming, "Learning about the world by learning about images," *Curr. Dir. Psychol. Sci.*, vol. 30, pp. 120–128, 2021.

[14] P. J. Marlow, J. Kim, and B. L. Anderson, "The perception and misperception of specular surface reflectance," *Current Biology*, vol. 22, no. 20, pp. 1909–1913, 2012.

[15] S. C. Pont and J. J. Koenderink, "Shape, surface roughness and human perception," in *Handbook of Texture Analysis*. World Scientific, 2008, pp. 197–222.

[16] Y.-X. Ho, M. S. Landy, and L. T. Maloney, "How direction of illumination affects visually perceived surface roughness," *Journal of Vision*, vol. 6, no. 5, p. 634–648, 2006.

[17] R. Fleming, F. Jäkel, and L. Maloney, "Visual perception of thick transparent materials," *Psychol. Sci.*, vol. 22, pp. 812–20, 06 2011.

[18] T. Kawabe, K. Maruya, and S. Nishida, "Perceptual transparency from image deformation," *Proc. Natl. Acad. Sci. USA*, vol. 112, 08 2015.

[19] D. Gigilashvili, J.-B. Thomas, J. Y. Hardeberg, and M. Pedersen, "Translucency perception: A review," *Journal of Vision*, vol. 21, no. 8:4, pp. 1–41, 2021.

[20] I. Motoyoshi, "Highlight-shading relationship as a cue for the perception of translucent and transparent materials," *Journal of Vision*, vol. 10:6, pp. 1–11, 07 2010.

[21] B. Xiao, S. Zhao, I. Gkioulekas, W. Bi, and K. Bala, "Effect of geometric sharpness on translucent material perception," *Journal of Vision*, vol. 20:10, pp. 1–17, 07 2020.

[22] H. Tamura, K. E. Prokott, and R. W. Fleming, "Distinguishing mirror from glass: A "big data" approach to material perception," *Journal of Vision*, vol. 22, no. 4:4, pp. 1–22, 2022.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[24] K. Storrs, B. Anderson, and R. Fleming, "Unsupervised learning predicts human perception and misperception of gloss," *Nature Human Behaviour*, vol. 5, pp. 1–16, 10 2021.

[25] K. Prokott, H. Tamura, and R. Fleming, "Gloss perception: Searching for a deep neural network that behaves like humans," *Journal of Vision*, vol. 21:14, pp. 1–20, 11 2021.

[26] A. O'Toole and C. Castillo, "Face recognition by humans and machines: Three fundamental advances from deep learning," *Annual Review of Vision Science*, vol. 7, 08 2021.

[27] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing." *Annual Review of Vision Science*, vol. 1, pp. 417–446, 2015.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE CVPR*, 2022, pp. 10 684–10 695.

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE CVPR*, 2020, pp. 8110–8119.

[30] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36 479–36 494, 2022.

[31] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[32] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[33] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," 2020. [Online]. Available: https://arxiv.org/abs/2006.06676

[34] C. Liao, M. Sawayama, and B. Xiao, "Translucency perception emerges in deep generative representations for natural image synthesis," 08 2022.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," 2017. [Online]. Available: https://arxiv.org/abs/1706.08500

[36] K. Van Ngo, J. J. Storvik, C. A. Dokkeberg, I. Farup, and M. Pedersen, "Quickeval: a web application for psychometric scaling experiments," in *IQSP XII*, vol. 9396. SCIA, 2015, pp. 1–13.

[37] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," 2020. [Online]. Available: https://arxiv.org/abs/2007.06600

[38] W. Chen, J. Litalien, J. Gao, Z. Wang *et al.*, "DIB-R++: learning to predict lighting and material with a hybrid differentiable renderer," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 22 834–22 848, 2021.

ConvNeXt-ChARM: ConvNeXt-based Transform for Efficient Neural Image Compression

Ahmed Ghorbel[1], Wassim Hamidouche[1,2], Luce Morin[1]
[1]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
[2]Technology Innovation Institute P.O.Box: 9639, Masdar City Abu Dhabi, UAE

***Abstract*—In recent years, neural image compression has garnered considerable attention from both research and industry. It has shown great promise in surpassing traditional methods in terms of rate-distortion performance through the development of end-to-end deep neural codecs. Despite these advancements, there is still room for improvement, particularly in reducing the coding rate while maintaining high reconstruction fidelity, especially in non-homogeneous textured image areas. Current models, including attention-based transform coding, also tend to have a higher number of parameters and longer decoding times. To address these challenges, we propose ConvNeXt-ChARM, an efficient ConvNeXt-based transform coding framework. It is coupled with a compute-efficient channel-wise auto-regressive prior that captures both global and local contexts from the hyper and quantized latent representations. Our architecture can be optimized end-to-end, fully leveraging context information to extract compact latent representations and achieve higher-quality image reconstructions. Experimental results conducted on four widely-used datasets demonstrate the effectiveness of ConvNeXt-ChARM. It consistently delivers significant BD-rate (PSNR) reductions, averaging 5.24% over the VVC reference encoder (VTM-18.0) and 1.22% over the state-of-the-art learned image compression method SwinT-ChARM. Additionally, we conduct model scaling studies to verify the computational efficiency of our approach. Furthermore, we perform objective and subjective analyses to highlight the performance gap between ConvNeXt, the next-generation ConvNet, and the Swin Transformer. Overall, our proposed ConvNeXt-ChARM framework showcases improved compression efficiency and reconstruction quality, establishing itself as a promising solution in the field of neural image compression.**

I. Introduction

Visual information is crucial in human development, communication, and engagement, and its compression is necessary for effective storage and transmission over constrained wireless/wireline channels. Thus, thinking about new lossy image compression approaches is a goldmine for scientific research. The goal is to reduce an image file size by permanently removing less critical information, particularly redundant data and high frequencies, to obtain the most compact bit-stream representation while preserving a certain level of visual fidelity. Nevertheless, the high compress rate and low distortion are fundamentally opposing objectives involving optimizing the rate-distortion tradeoff.

Conventional image and video compression standards including JPEG [1], JPEG2000 [2], H.265/high-efficiency video coding (HEVC) [3], and H.266/versatile video coding (VVC) [4], rely on hand-crafted creativity to present module-based encoder/decoder block diagram. In addition,

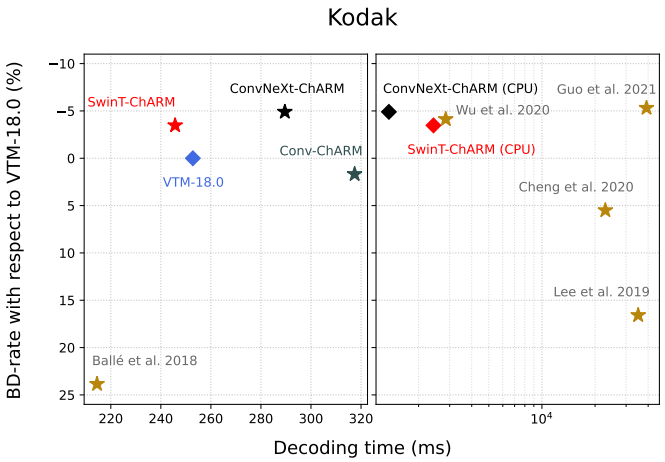


Fig. 1. BD-rate (%) versus decoding time (ms) on the Kodak dataset. Left-top is better. Star and diamond markers refer to decoding on GPU and CPU, respectively.

these codecs employ intra-prediction, fixed transform matrices, quantization, context-adaptive arithmetic coders, and various in-loop filters to reduce spatial and statistical redundancies, and alleviate coding artifacts. However, it has taken several years to standardize a conventional codec. Moreover, existing image compression standards are not anticipated to be an ideal and global solution for all types of image content due to the rapid development of new image formats and the growth of high-resolution mobile devices.

Lossy image compression consists of three modular parts: transform, quantization, and entropy coding. Each of these components can be represented as follows: i) autoencoders as flexible nonlinear transforms where the encoder (i.e., analysis transform) extracts latent representation from an input image and the decoder (i.e., synthesis transform) reconstructs the image from the decoded latent, ii) various differentiable quantization approaches which encode the latent into bitstream through arithmetic coding algorithms, iii) deep generative models as potent learnable entropy models estimating the conditional probability distribution of the latent to reduce the rate. Moreover, these three components can be optimized with end-to-end training by reducing the joint loss of the distortion between the original image and its reconstruction and the rate needed to transmit the bitstream of latent representation.

Thanks to recent advances in deep learning, we have seen many works exploring the potential of artificial neural

networks (ANNs) to form various learned image and video compression frameworks. Over the past two years, the performance of neural compression has steadily improved thanks to the prior line of study, reaching or outperforming state-of-the-art conventional codecs. Some previous works use local context [5]–[7], or additional side information [8]–[10] to capture short-range spatial dependencies, and others use non-local mechanism [11]–[14] as long-range spatial dependencies. Recently, Toderici *et al*. [15] proposed a generative compression method achieving high-quality reconstructions, Minnen *et al*. [16] introduced channel-conditioning and latent residual prediction taking advantage of an entropy-constrained model that uses both forward and backward adaptations, and Zhu *et al*. [17] replaced all convolutions in the channel-wise auto-regressive model (ChARM) prior approach [16] with Swin Transformer [18] blocks, Zou *et al*. [19] combined the local-aware attention mechanism with the global-related feature learning and proposed a window-based attention module, Koyuncu et al. [20] proposed a Transformer-based context model, which generalizes the standard attention mechanism to spatio-channel attention, Zhu *et al*. [21] proposed a probabilistic vector quantization with cascaded estimation under a multi-codebooks structure, Kim *et al*. [22] exploited the joint global and local hyperpriors information in a content-dependent manner using an attention mechanism, and He *et al*. [23] adopted stacked residual blocks as nonlinear transform and multi-dimension entropy estimation model.

One of the main challenges of learned transform coding is the ability to identify the crucial information necessary for the reconstruction, knowing that information overlooked during encoding is usually lost and unrecoverable for decoding. Another main challenge is the tradeoff between performance and decoding speed. While the existing approaches improve the transform and entropy coding accuracy, they remain limited by the higher decoding runtime and excessive model complexity leading to an ineffective real-world use. Finally, we found that attention-based networks taking advantage of attention mechanisms to capture global dependencies, such as Swin Transformer [18], have over-smoothed and contain undesirable artifacts at low bitrates. Furthermore, the global semantic information in image compression is less effective than in other computer vision tasks [19].

In this paper, we propose a nonlinear transform built on ConvNeXt blocks with additional down and up sampling layers and paired with a ChARM prior, namely ConvNeXt-ChARM. Recently proposed in [24], ConvNeXt is defined as a modernized ResNet architecture toward the design of a vision Transformer, which competes favorably with Transformers in terms of efficiency, achieving state-of-the-art on ImageNet classification task [25] and outperforming Swin Transformer on COCO detection [26] and ADE20K segmentation [27] challenges while maintaining the maturity and simplicity of convolutional neural networks (ConvNets) [24]. The contributions of this paper are summarized as follows:

- We propose a learned image compression model that leverages a stack of ConvNeXt blocks with down and up-sampling layers for extracting contextualized and non-linear information for effective latent decorrelation. We maintain the convolution strengths like sliding window strategy for computations sharing, translation equivariance as a built-in inductive bias, and the local nature of features, which are intrinsic to providing a better spatial representation.
- We apply ConvNeXt-based transform coding layers for generating and decoding both latent and hyper-latent to consciously and subtly balance the importance of feature compression through the end-to-end learning framework.
- We conduct experiments on four widely-used evaluation datasets to explore possible coding gain sources and demonstrate the effectiveness of ConvNeXt-ChARM. In addition, we carried out a model scaling analysis to compare the complexity of ConvNeXt and Swin Transformer.

Extensive experiments validate that the proposed ConvNeXt-ChARM achieves state-of-the-art compression performance, as illustrated in Figure 1, outperforming conventional and learned image compression methods in the tradeoff between coding efficiency and decoder complexity.

The rest of this paper is organized as follows. Section II presents our overall framework along with a detailed description of the proposed architecture. Next, we dedicate Section III to describe and analyze the experimental results. Finally, Section IV concludes the paper.

II. Proposed ConvNeXt-ChARM Model

A. Problem Formulation

The objective of learned image compression is to minimize the distortion between the original image and its reconstruction under a specific distortion-controlling hyper-parameter. Assuming an input image $\boldsymbol{x}$, the analysis transform g_a, with parameter ϕ_g, removes the image spatial redundancies and generates the latent representation $\boldsymbol{y}$. Then, this latent is quantized to the discrete code $\hat{\boldsymbol{y}}$ using the quantization operator $\lceil . \rfloor$, from which a synthesis transform g_s, with parameter θ_g, reconstructs the image denoted by $\hat{\boldsymbol{x}}$. The overall process can be formulated as follows:

$$\begin{aligned} \boldsymbol{y} &= g_a(\boldsymbol{x} \mid \phi_g), \\ \hat{\boldsymbol{y}} &= \lceil \boldsymbol{y} \rfloor, \\ \hat{\boldsymbol{x}} &= g_s(\hat{\boldsymbol{y}} \mid \theta_g). \end{aligned} \tag{1}$$

A hyperprior model composed of a hyper-analysis and hyper-synthesis transforms (h_a, h_s) with parameters (ϕ_h, θ_h) is usually used to reduce the statistical redundancy among latent variables. In particular, this hyperprior model assigns a few extra bits as side information to transmit some spatial structure information and helps to learn an accurate entropy model. The hyperprior generation can be summarized as follows:

$$\begin{aligned} \boldsymbol{z} &= h_a(\boldsymbol{y} \mid \phi_h), \\ \hat{\boldsymbol{z}} &= \lceil \boldsymbol{z} \rfloor, \\ p_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{z}}) &\leftarrow h_s(\hat{\boldsymbol{z}} \mid \theta_h). \end{aligned} \tag{2}$$

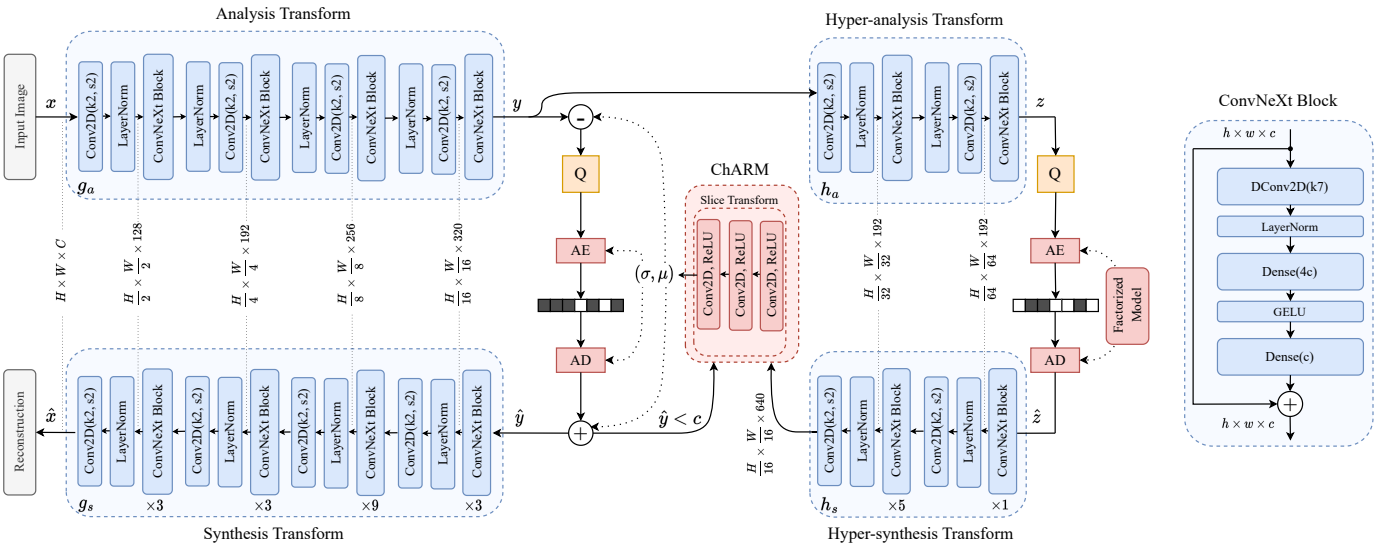


Fig. 2. Overall ConvNeXt-ChARM Framework. We illustrate the image compression diagram of our ConvNeXt-ChARM with hyperprior and channel-wise auto-regressive context model. We also present the ConvNeXt block used in both transform and hyper-transform coding for an end-to-end feature aggregation.

Transform and quantization introduce a distortion $D = MSE(\boldsymbol{x}, \hat{\boldsymbol{x}})$, for mean squared error (MSE) optimization that measures the reconstruction quality with an estimated bitrate R, corresponding to the expected rate of the quantized latents and hyper-latents, as described bellow:

$$R = \mathbb{E}\left[-\log_2(p_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{z}})) - \log_2(p_{\hat{\boldsymbol{z}}}(\hat{\boldsymbol{z}}))\right]. \tag{3}$$

Representing (g_a, g_s), (h_a, h_s), and entropy model by deep neural networks (DNNs) enables jointly optimizing the end-to-end model by minimizing the rate-distortion tradeoff $\mathcal{L}$, giving a rate-controlling hyper-parameter λ. This optimization problem can be presented as follows:

$$\begin{aligned}\mathcal{L} &= R + \lambda D, \\ &= \underbrace{\mathbb{H}(\hat{\boldsymbol{y}}) + \mathbb{H}(\hat{\boldsymbol{z}})}_{R} + \lambda\, MSE(\boldsymbol{x}, \hat{\boldsymbol{x}}),\end{aligned} \tag{4}$$

where $\mathbb{H}$ stands for the entropy.

B. ConvNeXt-ChARM network architecture

To better parameterize the distributions of the quantized latent features with a more accurate and flexible entropy model, we adopted the ChARM prior approach proposed in [16] to build an efficient ConvNeXt-based learning image compression model with strong compression performance. As shown in Figure 2, the analysis/synthesis transform (g_a, g_s) of our design consists of a combination of down and up-sampling blocks and ConvNeXt encoding/decoding blocks [24], respectively. Down and up-sampling blocks are performed using Conv2D and Normalisation layers sequentially. The architectures for hyper-transforms (h_a, h_s) are similar to (g_a, g_s) with different stages and configurations.

C. ConvNeXt design description

Globally, ConvNeXt incorporates a series of architectural choices from a Swin Transformer while maintaining the network's simplicity as a standard ConvNet without introducing any attention-based modules. These design decisions can be summarized as follows: macro design, ResNeXt's grouped convolution, inverted bottleneck, large kernel size, and various layer-wise micro designs. In Figure 2, we illustrates the ConvNeXt block, where the DConv2D(.) refers for the a depthwise 2D convolution, LayerNorm for the layer normalization, Dense(.) for the densely-connected NN layer, and GELU for the activation function.

Macro design: The depth per stages is adjusted from (3, 4, 6, 3) in ResNet-50 to (3, 3, 9, 3), which also aligns the FLOPs with Swin-T. In addition, the ResNet-style stem cell is replaced with a patchify layer implemented using a 2×2, stride two non-overlapping convolutional layers with an additional normalization layer to help stabilize the training. In ConvNeXt-ChARM diagram, we adopted the (3, 3, 9, 3) and (5, 1) as stage compute ratios for transforms and hyper-transforms, respectively.

Depthwise convolution: The ConvNeXt block uses a depthwise convolution, a special case of grouped convolution used in ResNeXt [28], where the number of groups is equal to the considered channels. This is similar to the weighted sum operation in self-attention, which operates by mixing information only in the spatial dimension.

Inverted bottleneck: Similar to Transformers, ConvNeXt is designed with an inverted bottleneck block, where the hidden dimension of the residual block is four times wider than the input dimension. As illustrated in the ConvNeXt block Figure 2, the first dense layer is 4 times wider then the second one.

large kernel: One of the most distinguishing aspects of Swin Transformers is their local window in the self-attention block. The information is propagated across windows, which enables each layer to have a global receptive field. The local window is at least 7×7 sized, which is still more extensive than the 3×3 ResNeXt kernel size. Therefore, ConvNeXt adopted large kernel-sized convolutions by using a 7×7 depthwise 2D convolution layer in each block. This allows our ConvNeXt-ChARM model to capture global contexts in both latents and hyper-latents, which are intrinsic to providing a better spatial representation.

Micro design: In ConvNeXt's micro-design, several per-layer enhancements are applied in each block, by using: a single Gaussian error linear unit (GELU) activation function (instead of numerous ReLU), using a single LayerNorm as normalization choice (instead of numerous BatchNorm), and using separate down-sampling layers between stages.

III. Results

First, we briefly describe used datasets with the implementation details. Then, we assess the compression efficiency of our method with a rate-distortion comparison and compute the average bitrate savings on four commonly-used evaluation datasets. We further elaborate a model scaling and complexity study to consistently examine the effectiveness of our proposed method against pioneering ones.

A. Experimental Setup

Datasets. The training set of the CLIC2020 dataset is used to train the proposed ConvNeXt-ChARM model. This dataset contains a mix of professional and user-generated content images in RGB color and grayscale formats. We evaluate image compression models on four datasets, including Kodak [29], Tecnick [29], JPEG-AI [29], and the testing set of CLIC21 [29]. For a fair comparison, all images are cropped to the highest possible multiples of 256 to avoid padding for neural codecs.

Implementation details. We implemented all models in TensorFlow using tensorFlow compression (TFC) library [30], and the experimental study was carried out on an RTX 5000 Ti GPU. All models were trained on the same CLIC2020 training set with 3.5M steps using the ADAM optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 10^{-4} and drops to 10^{-5} for another 100k iterations, and $L = R + \lambda D$ as loss function. The MSE is used as the distortion metric in RGB color space. Each batch contains eight random 256 × 256 crops from training images. To cover a wide range of rate and distortion, for our proposed method, we trained five models with $\lambda \in \{0.006, 0.009, 0.020, 0.050, 0.150\}$. Regarding the evaluation on CPU, we used an Intel(R) Xeon(R) W-2145 @ 3.70GHz.

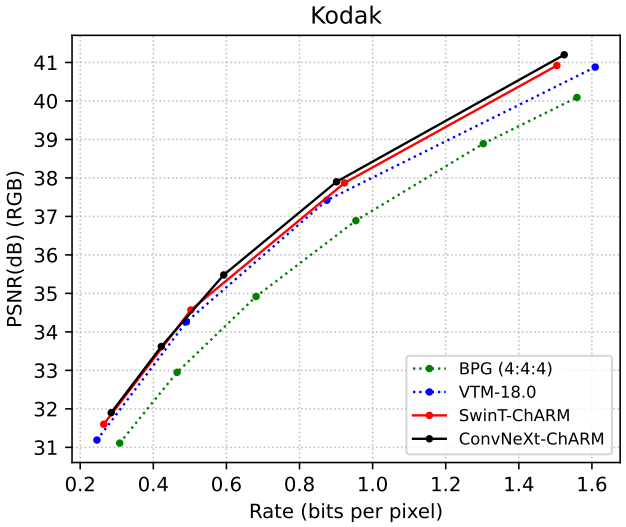


Fig. 3. Rate-distortion comparison on Kodak dataset.

Baselines.[1] We compare our approach with the state-of-art neural compression method SwinT-ChARM proposed by Zhu *et al.* [17], and non-neural compression methods, including better portable graphics (BPG)(4:4:4), and the most up-to-date VVC official Test Model VTM-18.0 in All-Intra profile configuration.

B. Rate-Distortion coding performance

To demonstrate the compression efficiency of our proposed approach, we visualize the rate-distortion curves of our model and the baselines on each of the considered datasets. Considering the Kodak dataset, Figure 3 shows that our ConvNeXt-ChARM outperforms the state-of-the-art learned approach SwinT-ChARM, as well as the BPG(4:4:4) and VTM-18.0 traditional codecs in terms of PSNR. Regarding rate savings over VTM-18.0, SwinT-ChARM has more compression abilities only for low PSNR values. Our model can be generalized to high resolution image datasets (Tecnick, JPEG-AI, and CLIC21), and can still outperform existing traditional and the learned image compression method SwinT-ChARM in terms of PSNR. Besides the rate-distortion curves, we also evaluate different models using Bjontegaard's metric [31], which computes the average bitrate savings (%) between two rate-distortion curves. In Table I, we summarize the BD-rate of image codecs across all four datasets compared to the VTM-18.0 as the anchor. On average, ConvNeXt-ChARM is able to achieve 5.24% rate reduction compared to VTM-18.0 and 1.22% relative gain from SwinT-ChARM. Figure 1 shows the BD-rate (with VTM-18.0 as an anchor) versus the decoding time of various approaches on the Kodak dataset. It can be seen from the figure that our ConvNeXt-ChARM achieves a good tradeoff between BD-rate performance and decoding time.

[1]For a fair comparison, we only considered SwinT-ChARM [17] from the state-of-the-art models [17], [19]–[23], due to the technical feasibility of models training and evaluation under the same conditions and in an adequate time.

TABLE I
BD-RATE↓ PERFORMANCE OF BPG (4:4:4), SWINT-CHARM, AND CONVNEXT-CHARM COMPARED TO THE VTM-18.0 FOR THE FOUR CONSIDERED DATASETS.

Dataset	BPG444	SwinT-ChARM	ConvNeXt-ChARM
Kodak	20.73%	-3.47%	**-4.90%**
Tecnick	27.03%	-6.52%	**-7.56%**
JPEG-AI	28.14%	-0.23%	**-1.17%**
CLIC21	26.54%	-5.86%	**-7.36%**
Average	25.61%	-4.02%	**-5.24%**

TABLE II
IMAGE CODEC COMPLEXITY. WE CALCULATED THE AVERAGE DECODING TIME ACROSS 7000 IMAGES AT 256×256 RESOLUTION, ENCODED AT 0.6 BPP. THE BEST SCORE IS HIGHLIGHTED IN BOLD.

Image Codec	Latency(ms)↓		GFLOPs↓	#params(M)↓
	GPU	CPU		
Conv-ChARM	124.32	967.43	**117**	123.84
SwinT-ChARM	**102.45**	1088.16	122	127.78
Ours	122.70	**834.42**	119	**122.33**

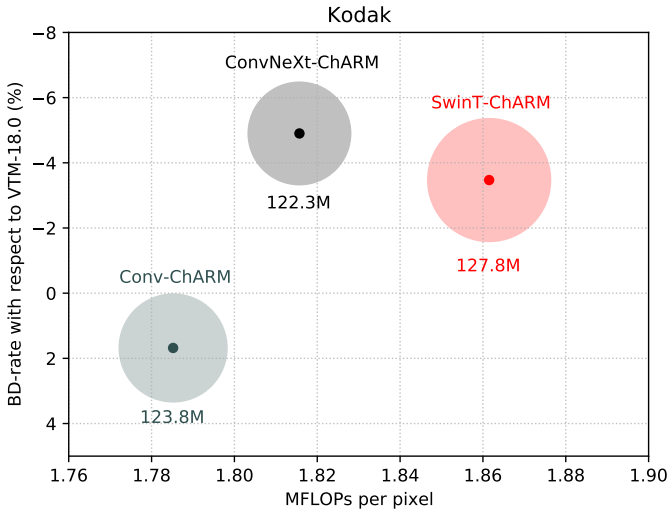


Fig. 4. Model size scaling. BD-Rate versus MFLOPs per pixel for our model ConvNeXt-ChARM compared to Conv-ChARM and SwinT-ChARM (for both encoding and decoding).

C. Models Scaling Study

We evaluated the decoding complexity of the three considered image codecs by averaging decoding time across 7000 images at 256×256 resolution, encoded at 0.6 bpp. We present the image codec complexity in Table II, including decoding time on GPU and CPU, floating point operations per second (GFLOPs), the memory required by model weights, and the total model parameters. The models run with Tensorflow 2.8 on a workstation with one RTX 5000 Ti GPU. The Conv-ChARM model refers to the Minnen *et al*. [16] architecture with a latent depth of 320 and a hyperprior depth of 192, and can be considered as ablation of our model without ConvNeXt blocks. We maintained the same slice transform configuration of the ChARM for the three considered models. The total decoding time of SwinT-ChARM decoder is less than ConvNets-based decoder on GPU but is the highest on CPU. Our ConvNeXt-ChARM is lighter than the Conv-ChARM in terms of the number of parameters, which proves the ConvNeXt block's well-engineered design. Compared with SwinT-ChARM, our ConvNeXt-ChARM shows lower complexity, requiring lower training time with less memory consumption. In addition, Figure 4 shows that our method is in an interesting area, achieving a good tradeoff between BD-rate score on Kodak, total model parameters, and MFLOPs per pixel, highlighting an efficient and hardware-friendly compression model.

D. Comparison with SwinT-ChARM

ConvNeXt-ChARM achieves good rate-distortion performance while significantly reducing the latency, which is potentially helpful to conduct, with further optimizations, high-quality real-time visual data transmission, as recently proposed in the first software-based neural video decoder running HD resolution video in real-time on a commercial smartphone [32]. Since fewer works attempt to explicitly compare Swin Transformer and ConvNet-based blocks, here, we compare our ConvNeXt-ChARM with SwinT-ChARM under the same conditions and configurations. We found that a well-designed ConvNet, without any additional attention modules, can outperform the highly coveted Swin Transformer in learned transform coding in terms of BD-rate, with more visually pleasing reconstructions and comparable decoding latency. In addition, ConvNeXt-ChARM maintains the efficiency and maturity of standard ConvNets and the fully-convolutional nature for both training and inference. There is no doubt that Transformers are excellent architectures with enormous potential for the future of various computer vision applications. However, their vast hunger for data and computational resources [33] poses a big challenge for the computer vision community. Taking SwinT-ChARM as an example, it needs, on average, ×1.33 more time than ConvNeXt-ChARM, to train on the same number of epochs.

IV. CONCLUSION

In this work, we reconcile compression efficiency with ConvNeXt-based transform coding paired with a ChARM prior and propose an up-and-coming learned image compression model ConvNeXt-ChARM. Furthermore, we inherit the advantages of pure ConvNets in the proposed method to improve both efficiency and effectiveness. The experimental results, conducted on four datasets, showed that our approach outperforms previously learned and conventional image compression methods, creating a new state-of-the-art rate-distortion performance with a significant decoding runtime decrease. Future work will further investigate efficient low-complexity entropy coding approaches to further enhance decoding latency. With the development of GPU chip technology and the further optimization of engineering, learning-based codecs will be the future of coding, achieving better

compression efficiency when compared with traditional codecs and aiming to bridge the gap to a real-time operation. We hope our study will challenge certain accepted notions and prompt people to reconsider the significance of convolutions in computer vision.

REFERENCES

[1] G.K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[2] Majid Rabbani and Rajan Joshi, "An overview of the jpeg 2000 still image compression standard," *Signal processing: Image communication*, vol. 17, no. 1, pp. 3–48, 2002.

[3] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] Gary Sullivan, "Versatile video coding (vvc) arrives," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 1–1.

[5] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[6] Jooyoung Lee, Seunghyun Cho, Seyoon Jeong, Hyoungjin Kwon, Hyunsuk Ko, Hui Yong Kim, and Jin Soo Choi, "Extended end-to-end optimized image compression method based on a context-adaptive entropy model.," in *CVPR Workshops*, 2019, p. 0.

[7] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool, "Conditional probability models for deep image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4394–4402.

[8] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[9] Yueyu Hu, Wenhan Yang, and Jiaying Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11013–11020.

[10] David Minnen, George Toderici, Saurabh Singh, Sung Jin Hwang, and Michele Covell, "Image-dependent local entropy models for learned image compression," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 430–434.

[11] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[12] Mu Li, Kai Zhang, Jinxing Li, Wangmeng Zuo, Radu Timofte, and David Zhang, "Learning context-based nonlocal entropy modeling for image compression," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[13] Yichen Qian, Zhiyu Tan, Xiuyu Sun, Ming Lin, Dongyang Li, Zhenhong Sun, Hao Li, and Rong Jin, "Learning accurate entropy model with global reference for image compression," *arXiv preprint arXiv:2010.08321*, 2020.

[14] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Transactions on Image Processing*, vol. 30, pp. 3179–3191, 2021.

[15] George Dan Toderici, Fabian Julius Mentzer, Eirikur Thor Agustsson, and Michael Tobias Tschannen, "High-fidelity generative image compression," June 2 2022, US Patent App. 17/107,684.

[16] David Minnen and Saurabh Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.

[17] Yinhao Zhu, Yang Yang, and Taco Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2021.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[19] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang, "The devil is in the details: Window-based attention for image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17492–17501.

[20] A Burakhan Koyuncu, Han Gao, Atanas Boev, Georgii Gaikov, Elena Alshina, and Eckehard Steinbach, "Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*. Springer, 2022, pp. 447–463.

[21] Xiaosu Zhu, Jingkuan Song, Lianli Gao, Feng Zheng, and Heng Tao Shen, "Unified multivariate gaussian mixture for efficient neural image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17612–17621.

[22] Jun-Hyuk Kim, Byeongho Heo, and Jong-Seok Lee, "Joint global and local hierarchical priors for learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5992–6001.

[23] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5718–5727.

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11976–11986.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[29] Benchmark datasets, "kodak testing set: http://r0k.us/graphics, tecnick testing set: https://testimages.org/, jpeg-ai testing set: https://jpegai.github.io/test_images/, and clic21 testing set: http://compression.cc/tasks/," .

[30] Johannes Ballé, Sung Jin Hwang, and Eirikur Agustsson, "TensorFlow Compression: Learned data compression," 2022.

[31] Gisle Bjontegaard, "Calculation of average psnr differences between rd-curves," *VCEG-M33*, 2001.

[32] Hoang Le, Liang Zhang, Amir Said, Guillaume Sautiere, Yang Yang, Pranav Shrestha, Fei Yin, Reza Pourreza, and Auke Wiggers, "Mobilecodec: neural inter-frame video compression on mobile devices," in *Proceedings of the 13th ACM Multimedia Systems Conference*, 2022, pp. 324–330.

[33] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

Image quality assessment for MRI: Is it up to the task?

Mohamed L Seghier
D*ep. of Biomedical Engineering*
Khalifa University
Abu Dhabi, UAE
mseghier@gmail.com

***Abstract*—** The accuracy of MRI-based diagnosis can be degraded by artifacts, a challenging problem for both radiologists and automated computer-aided systems. Assessment of MR images quality is thus of paramount importance for clinical and research purposes. In this brief review, I discuss image quality assessment (IQA) methods for MRI. After briefly describing some typical artifacts encountered in MRI, a succinct summary of popular IQA metrics is provided. A particular emphasis is put on the relevance of current IQA metrics, borrowed from existing IQA methods for natural images, when dealing with MR images with diverse contrasts and artifact types. In the IQA process, what matters for clinicians is not whether an MR image is beautiful or not but whether the clinically-relevant diagnostic pattern in the MR image is unaffected by artifacts. I then discuss the growing interest in AI-based as well as brain-inspired IQA methods, including their strengths and limitations. The possibility of integrating IQA metrics within AI image reconstruction and processing tools will have major ramifications on IQA for MRI. Challenges and promises are finally discussed in the light of the recent trends in scanning patients at ultrahigh (≥7 Tesla) or at ultralow (≤0.1 Tesla) magnetic fields with portable MRI devices.

***Keywords*—** *MRI, artifacts, image quality, contrast, signal, noise, distortions, motion, AI, MR sequence, field strength.*

I. INTRODUCTION

MRI is widely used in clinical and research settings for both diagnostic and prognostic purposes. With a strong static magnetic field, combined with radiofrequency pulses and spatially variable magnetic gradients, different contrast images can be generated of the soft tissue of the human body. High quality MR image acquisition requires highly homogenous magnetic fields. However, inhomogeneities caused by different sources yield visible artifacts that can lessen the quality of MR images. Such artifacts can lead to geometric distortions, inaccurate contrast distribution, variable signal intensities across the image, and signal loss. Such artifacts can negatively impact upon the diagnostic potential of MR images, and hence different correction techniques are commonly used to reduce artifacts during (prospective) or after (offline) acquisition. Indeed, there is a rich literature that is interested in the development of processing methods towards artifact-free MRI. One question of paramount importance in this literature is the possibility to objectively assess the quality of MR images, as discussed in this brief review. The examples provided here are mainly taken from the domain of neuroimaging, though they are also valid for MR images of other body parts.

Raw MRI data are typically collected in a spatial frequency-based space called the k-space (Figure 1). Depending on how the k-space is sampled and reconstructed, methods based on Fourier transform are used to transform the k-space data into a real MR image. Each pixel of an MR image is by construction a weighted sum of all the individual points in the k-space. Image contrast and low signal variations are coded in central regions of the k-space, whereas sharp intensity transitions and edges are coded in peripheral regions of the k-space (Figure 1). Accordingly, as each point of the k-space contributes to the entire MR image, any mislocalization of points in the k-space will translate into artifacts [1]. Strategies have been suggested to reduce artifacts directly in the k-space (e.g. [2, 3]. However, in the context of image quality assessment (IQA), methods have exclusively been developed and tested on real MR images.

One critical aspect for IQA concerns the estimation and modelling of MRI noise type and intensity. Noise in MRI in single-coil acquisitions is mainly governed by a Rician distribution at low signal-to-noise ratio (SNR) but behaves like Gaussian noise at high SNR [4]. Noise estimation for multiple-coil acquisitions and parallel imaging protocols uses the noncentral chi model [5]. Many parameters in the MRI acquisition sequence have direct impact on SNR [6]. For dynamic acquisitions in functional or perfusion MRI (i.e. 4D data), different metrics have been put forward, including the concept of temporal SNR as a useful metric of timeseries quality [7, 8].

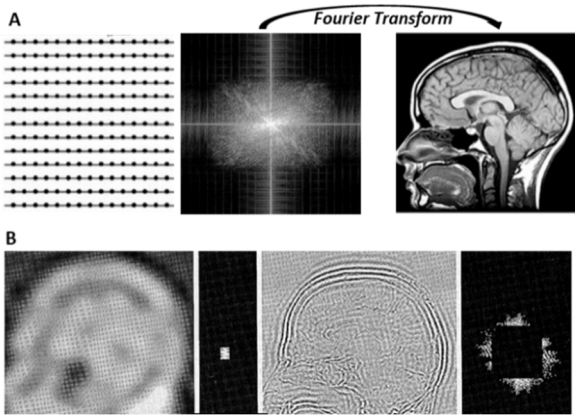


Figure 1: (A) k-space is an array of spatial frequencies. Each point of the k-space represents a collected intensity at a given spatial frequency. The points in k-space are acquired through frequency encoding and phase encoding steps. MR image of the scanned part is directly obtained with a Fourier transform. Inaccurate sampling of the k-space translates into artifacts in the real MR images. (B) the center of the k-space codes low frequencies (image contrast) and the periphery of the k-space codes high frequencies (image edges and sharp transitions).

II. ARTIFACTS IN MRI

Below, I succinctly describe ten main artifacts in MRI; more details can be found elsewhere [9-13]. These artifacts might be due to the scanner (inhomogeneity in the static

979-8-3503-4218-5/23/$31.00 ©2023 IEEE

magnetic field, software errors), hardware (suboptimal gradients, malfunctioning RF coils), sequence type (some sequences are more prone to artifacts than others), or other patient factors (tissue heterogeneity or movement) [14]. Figure 2 illustrates some common artifacts.

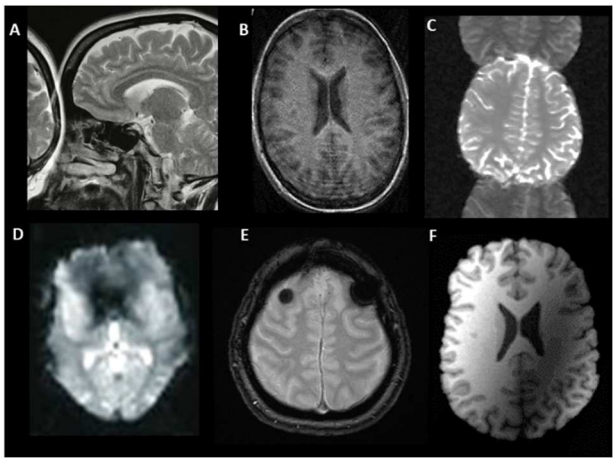


Figure 2: (A) aliasing artifact (adapted from Radiology_study); (B) motion artifacts (see the MR-ART dataset), (C) Nyquist N/2 ghost, (D) signal loss; (E) blooming effect (adapted from radiopaedia); (F) bias field artifact.

Aliasing artifacts: This artifact typically occurs when the dimensions of the imaged object exceed the selected field of view [15]. The part of the body that is outside the field of view is projected onto the other side of the image.

Truncation artifacts: also known as Gibbs artifacts, they manifest as multiple fine parallel lines adjacent to high-contrast interfaces [16]. They can be seen for example in spinal MRI as artifactual false widening of the edges at these high-contrast interfaces, or edge enhancement of the interface and distortion of tissues immediately adjacent to the high-contrast interface. They result from the application of truncated Fourier transform to reconstruct the MR signals.

Nyquist N/2 Ghost: This artifact is commonly seen with echo-planar imaging (EPI) sequences. Slight timing differences between peaks of odd and even echoes can lead to an aliased ghost halfway across the image [17]. Poor shimming, gradient coil heating and eddy currents can result in Nyquist ghosts.

Motion artifacts: they appear like blurring or ghosting caused by subject motion, breathing, or cardiac pulsations [18]. It is one of most frequent artifacts that can affect all sequence types in the presence of movement from the subject. They are very common in particular when scanning vulnerable clinical populations (epileptic patients, stroke patients, very young children, seniors with dementia…etc).

Susceptibility artifacts: they reflect distortions due to local magnetic field inhomogeneities [19]. One classic example is imaging near metallic orthopedic or dental implants. The magnetic field distortions created by susceptibility effects result in frequency changes that, in turn, produce a signal loss.

Blooming artifact: is a type of susceptibility artifact seen in with some MRI sequences in the presence of paramagnetic substances. For instance, blooming can be seen surrounding calcifications or hemosiderin from prior hemorrhage. This artifact may sometimes lead to tissue signal cancellation and loss of anatomical borders.

Geometric distortion: They can arise from a variety of sources, including tissue-dependent chemical shift, susceptibility differences, gradient field nonlinearity and the static field inhomogeneity. This artifact can reduce spatial fidelity, due to a geometric offset of the voxel's representation in the image space, which can directly introduce spatial inaccuracies in tissue localization and delineation [20].

Intensity inhomogeneity (bias field) artifact: This artifact can lead to undesirable intensity variations across the image in tissues having the same physical property, which may hinder the accuracy of tissue segmentation [21]. It commonly refers to a very smooth and/or low-frequency variation that can corrupt MR images.

Partial volume effects: these are notoriously difficult to assess or correct and are particularly challenging for tissue segmentation techniques [22]. They are present at the voxel level when more than one tissue type occurs in a given voxel. They occur when the voxel size is larger (low spatial resolution) than the size of tissue variation in the image.

Poor signal-to-noise ratio (SNR): SNR is an important factor that determines the quality of an MR image. Poor SNR can lead to poor sensitivity and low differentiation of tissue types. Factors such as acquisition duration and voxel size have direct impact on SNR [23]. For instance, poor SNR is a limiting factor in MRI at ultralow fields. As discussed below, noise estimation in MRI is not a straightforward question as many factors have direct impact on the exact modelling of noise in MR images, including the type of contrast, sequence parameters, single versus multi-coil acquisition, and type of imaging acceleration [6, 24, 25].

In sum, MRI offers a powerful framework to display soft tissue at different contrasts using tailored MR sequences. Figure 3 illustrates some of the most common MR contrasts used in clinical neuroimaging. These MR images at different contrasts are sensitive to different types of artifacts. Some artifacts are more severe in some MR contrasts than others. For example, geometric distortions might be more pronounced in diffusion imaging than other contrasts, whereas Nyquist N/2 artifacts is more common when collecting T2*-weighted images with EPI-type sequences. The choice of the optimal IQA protocol thus depends on which MR contrast (i.e. sequence) is of interest to the user.

III. IMAGE QUALITY ASSESSMENT (IQA)

In the clinical setting, it is not unusual that patients are asked to repeat scans when artifacts are detected in the acquired MR images. Those images affected by artifacts require extensive preprocessing to improve quality, making them very challenging for existing automated diagnostic tools. Hence, there is a need to assess, qualitatively or quantitatively, the quality of collected MR images before carrying out any subsequent processing or analysis. In this section, I will describe some popular IQA approaches that have been used on MR images. The exact mathematical formulations of IQA metrics are not covered in this brief review.

The gold standard IQA approach still relies on the subjective assessment made by experts to rate or rank the quality of MR images. For example, experts can label a given MR image as of good quality with less geometric distortions or with no apparent head motion artifacts. This is typically based on prior knowledge about how artifacts manifest themselves in MR images. However, subjective IQA has

many limitations [26, 27]: it is time consuming, expensive, operator-dependent, and not practical for the evaluation of large MRI datasets. Perhaps most importantly, although many IQA tools for natural images have been successfully introduced into MRI, what matters for clinicians is not necessarily whether an MR image looks beautiful or nice (i.e. a perceptual quality) but whether the image's features are clinically useful for accurate diagnosis. For instance, a diffusion-weighted image (DWI) might look of poor quality compared to a high-resolution T1-weighted image, but a DWI image is more useful for clinicians in the diagnosis of acute stroke lesions. This point is critical when it comes to the design of optimal IQA for MR images, given that quality varies with the MR contrast of interest. This point is central to any discussion about IQA for MRI because 'quality' here would basically mean that the relevant diagnostic pattern in the MR image is not altered by any present artifact. This issue is sometimes overlooked in current literature that borrowed IQA metrics from natural images to MR images.

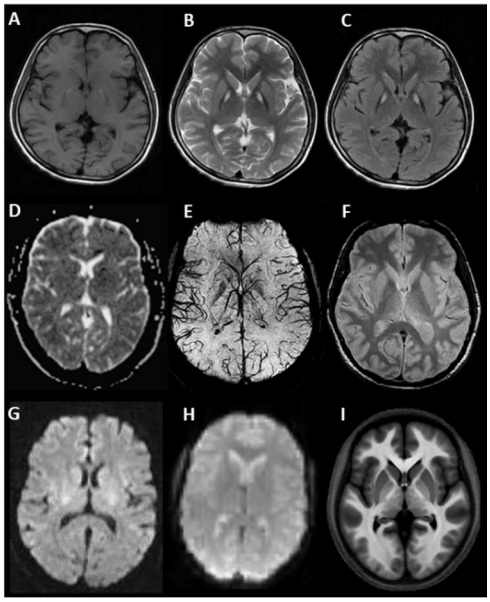


Figure 3. Some common MR contrasts. (A) T1-weighted image, (B) T2-weighted image, (C) fluid attenuated inversion recovery (FLAIR)-weighted image, (D) apparent diffusion coefficient (ADC)-weighted image, (E) susceptibility-weighted image, (F) proton density-weighted image, (G) diffusion-weighted image, (H) T2*-weighted image with EPI, (I) magnetization transfer-weighted image. Importantly, type and severity of artifacts vary with MR contrast.

In this context, there is a need to develop objective measures of IQA to automate screening and reduce reliance on human raters [28]. One of the early IQA frameworks in the context of medical imaging is what is known as model observers [29], used as objective alternatives to human observers. The models based on statistical decision theory, including for instance the ideal Bayesian observer and the optimum linear discriminator (Fisher-Hotelling) model, can predict performance for clinical classification and estimation tasks [30]. They rely on both mathematical modelling and psychophysical considerations in designing both optimal observers (for optimizing medical imaging systems) and anthropomorphic observers (for modeling human observers); for review see [31, 32]. As task-based approaches, these model observers are still attracting interest in the field for the development of optimal IQA protocols [33-35]. However, they usually lead to an overall outcome without the separation of the image quality parameters (see discussion in [36]).

Objective IQA can be divided into different categories depending on whether a reference (pristine) image is available or not: (i) full-reference IQA (FR-IQA) requires the availability of a reference image, (ii) reduced-reference IQA (RR-IQA) requires the availability of some features extracted from a reference image, and (iii) no-reference IQA (NR-IQA), sometimes referred to as blind IQA, which can assess image quality without a reference image. The vast majority of IQA algorithms belong to FR-IQA type. Others have suggested (iv) a relative IQA framework based on ranking quality scores instead of absolute quality scores *per se*, a framework that has been shown to be useful for MR images collected with susceptibility weighted imaging protocols (e.g. [37, 38]. Below, I provide a brief introduction of both FR-IQA and NR-IQA; more details can be found elsewhere (cf. [39-41]).

A. Phantom imaging for FR-IQA:

Phantom imaging is a common protocol to evaluate, calibrate, and tune the performance of MRI scanners, including the analysis of scanner-related artifacts. It is based on imaging a standard phantom that is stable and having well-defined properties to allow monitoring of scanner performance and accuracy of image-based measurements. For instance, imaging phantoms is used to assess SNR, size of geometric distortion, signal drift for 4D acquisitions as in functional or perfusion MRI, and/or the size of bias field with different types of coils [42]. Following the American College of Radiology Phantom Test Protocol, MR phantom-based assessment can gauge the following features: geometric accuracy, high-contrast resolution, slice thickness accuracy, slice position accuracy, image intensity uniformity, percent signal ghosting, low-contrast object detectability, SNR and central frequency monitoring [43]. Customized phantoms can also be used to assess the size of geometric distortions even at the sub-millimetric level, for instance by calculating the difference between specific points in the acquired MR images against the true physical features of the same points in the phantom [44, 45]. Phantoms are very handy as they provide a reference image to which the quality of acquired MR images can be compared to, hence enabling the application of FR-IQA methods. Interestingly, there are other advanced types of phantoms like anthropomorphic brain phantoms [46, 47] that offer a realistic depiction of tissue classes, which could improve the IQA process for diverse neuroimaging applications.

There are more than 100 FR-IQA metrics, so it is beyond the scope of this brief review to comprehensively appraise their applicability or relevance to MRI. They include for instance some classic metrics such as mean square error (MSE), peak signal-to-noise ratio (PSNR), contrast-to-noise ratio (CNR), and structural similarity index measure (SSIM) [37, 48, 49]. FR-IQA can be classified further into different families depending on their mode of operation [40, 50]. Other recent IQA methods have shown both high accuracy and robustness that would make them suitable to MR images (e.g. [51, 52]).

A comparison between 43 FR-IQA methods [40] on different datasets illustrates the complexity of the IQA process depending on what artifacts are preponderant in the images of interest. For instance, a recent comparison on MR images [41] reported that some FR-IQA methods performed better than other methods, including the visual saliency-based index. Despite FR-IQA metrics being easily quantifiable and interpretable, the availability of a reference image in MRI is

not always possible. For example, some MR contrasts exploit inherent artifacts in the acquired images to distinguish normal versus abnormal tissue in patients, e.g. susceptibility artifacts as a proxy for the detection of cerebral microbleeds when using susceptibility-weighted imaging. This makes the distinction between what is artifact and what is relevant signal extremely hard to replicate with standard phantoms. This issue highlights the importance of alternative NR-IQA methods.

B. No-reference image quality assessment (NR-IQA):

There is a huge interest in developing NR-IQA approaches that do not require a reference image. These methods can deliver competitive performance as compared to subjective assessments by human experts [53, 54]. One classic example is the use of SNR [55] as a quality indicator to flag up MR images of bad quality [56, 57]; low SNR values would indicate poor image quality. The ratio between signal and noise can be calculated in two ways in MRI, assuming a spatially homogeneous distribution of noise over the whole image: (i) by defining two regions of interest (ROI), signal is set to the average intensity in the ROI of interest (e.g. brain tissue) and noise is set to the standard deviation of pixels intensity in a background ROI, or (ii) using two identical (repeated) images of the brain, signal is defined as the average intensity in the ROI in the first image and noise as the standard deviation of the pixels inside the same ROI of the difference between the two images. Some studies have shown that SNR measures using the two-ROI approach for MRI are not valid unless the statistical intensity distribution of the background noise follows a Rayleigh distribution [6].

Similar to SNR, other alternative metrics were developed. For instance, Mortamet et al. suggested a metric based on the analysis of a single ROI in the background air region of a brain MR image [58]. Specifically, after delineating the background ROI, a model-free quality index is assessed and subsequently combined with another index that examines the noise intensity distribution by fitting a noise model [58]. The usefulness of this methods has been tested on a large dataset with structural MR images. Other NR-IQA metrics were directly compared to SNR on MR images, including indicators such as BLIINDS-II [59] and BRISQUE [60], and were shown to be accurate and robust (see empirical evidence in [56]).

NR-IQA methods can be grouped [40] into different families. They were compared on different datasets [40], showing for instance a superiority of the codebook representation for no-reference image quality assessment (CORNIA) method for different types of artifacts. In another systematic comparison of >200 methods, NR-IQA methods were able to reliably discriminate between undistorted MR images versus MR images contaminated with either noise or distortion [28], though their performance varied with the type and level of distortion. Similarly, other alternative methods based on the examination of images statistical features and local texture showed promising results on MR images [61-63].

It is worth noting that the majority of NR-IQA methods were mainly tested on structural MRI images with very few applications to other MRI domains such functional and perfusion imaging. I also note that the performance of NR-IQA methods strongly depends on the type of the artifact and its severity, and there is no single NR-IQA technique that can handle diverse artifacts across many MR contrasts. For instance, some of NR-IQA metrics can be optimal for identifying scanner-related artifacts but might not be optimal for patient-related artifacts. Last but not least, some of NR-IQA methods presumably rely on prior knowledge about image features (e.g. location of a given brain tissue as an ROI versus background), making them more likely to operate as RR-IQA methods. In the light of these challenges, I briefly discuss in the next paragraph what AI can bring to the table.

IV. AI-BASED IQA

AI has opened new opportunities to automate the IQA process, taking advantage of the existence of annotated data [64-70]. Such AI tools, including deep learning methods, can classify images into good versus poor quality images after learning from images with known labels; for review see [71]. Popular AI architectures for IQA include convolutional neural networks (CNN) [65, 72-74], and end-to-end AI-based solutions to IQA already exist (e.g. [75, 76]).

AI-based IQA techniques, evaluated on MR images [77-80], were found to be in good agreement with human expert evaluation (see discussion in [54]). AI-based IQA tools are typically designed for one specific type of artifacts in MR images (e.g. motion artifacts [81, 82]). Despite their huge potential for IQA, they pose many challenges. First, IQA methods based on deep learning require copious amounts of annotated data, which might not be available for all types of artifacts or MR contrasts. Second, the implementation of a variety of neural network architectures entails sophisticated hyperparameters optimization procedures in the search hyperspace. Such optimization of parametrization is fundamentally an empirical question that depends on data, thus making the generalizability of designed architectures to other MR data a difficult question (e.g. data collected with other scanners or with different sequences). Third, many of existing AI-based IQA methods operate on 2D mode, which can hamper their effectiveness for the detection of heterogenous artifacts that extend across multiple slices in the collected 3D volumes. Extending to 3D inputs however increases dramatically the complexity of AI architectures.

Furthermore, the performance of AI methods might be bounded by the image quality criteria used by the human experts, which raises questions on how to deal with uncertainty in the definition of the labels (ground-truth) during the training phase. Likewise, class imbalance is another challenge as some artifacts might occur less frequently than other artifacts. Moreover, artifacts in MRI are underpinned by different interacting sources where the severity of an artifact might also be exacerbated by the presence of another artifact (e.g. the presence of motion can worsen geometric distortions), which might affect the performance of AI tools for multiclass problems. These methodological issues warrant future research before AI-based IQA tools can be considered robust IQA methods for MRI. Perhaps most importantly, MRI developers are already integrating AI capabilities into their image preprocessing and reconstruction platforms, making the quality assessment process an inherent part of the whole acquisition protocol. This will likely accelerate the development of AI-based IQA methods that are compatible with existing AI image reconstruction tools.

V. BRAIN-INSPIRED IQA

To mimic the role of human experts in subjective IQA, it would make sense to develop IQA methods based on how the visual system processes information [83-86]. Several functional properties of the human visual system were incorporated in the design of IQA methods, including human visual sensitivity [87], low-level features [88], receptive fields

representations [89], frequency and spatial features [90, 91], and multi-channel processing modes [74].

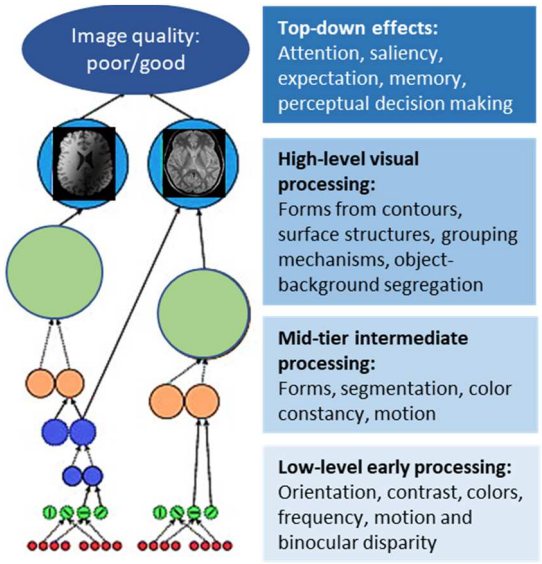


Figure 4: A schematic illustration of how vision works (see for more details [83]): low-level to high-level processes, through multi-resolution representations with increasing receptive fields size, based on segmentation and grouping mechanisms, across functionally specialized areas, via parallel but interacting streams, ensuring invariance to visual transformation, preserving chromatic appearance, and ensuring coherence with prior knowledge during object recognition. Object recognition is the sum of both bottom-up (feedforward) and top-down (feedback) inputs, based on the physical properties of the visual stimulus (e.g. signal intensity and spatial frequency in the MR image) and prior knowledge about MR images (expected contrast and location of brain tissue classes).

There is a huge neuroscience literature about the structure and function of the human visual system, highlighting its hierarchical organization and its subdivision into specialized modules or areas [92]. Below is a selective list of ten characteristics of the human visual system (Figure 4): (i) a rich connectivity from the retina, to the lateral geniculate body, to the primary visual cortex and other higher visual cortices, involving complex feedforward and feedback interactions (ii) retinotopic organization in different visual areas, implying that neurons with receptive fields close together in the visual field are also close together in the cortex, (iii) neurons in visual areas (in particular in low-level areas) are sensitive to orientation, (iv) cortical magnification, with large cortical representations of central (foveal) compared to peripheral regions, (v) size of receptive fields increases from low to high-level visual areas, yielding multiresolution integrated representations of the presented image, (vi) contrast sensitivity varies across visual areas, with the ability of the system to preserve the chromatic appearance of the presented image, (vii) different functional specialization across visual areas, with some areas responding preferably to different physical properties (e.g. area V5/MT for motion, area V4/V8 for color processing), (viii) spatial (and temporal) features of the presented image implicate relatively different parvocellular and magnocellular pathways, (ix) responses in different areas are strongly modulated by visual attention, and (x) image recognition involves complex interplay between top-down and bottom-up interactions.

Not all properties of the visual system can be easily translated into IQA metrics. Thus, rather than considering each functional or structural property of the human visual system separately, it would make sense to take a more holistic view in terms of organizational principles. For example, the visual system can be modeled as an inferential system that aims to give sense to sensory inputs, a system that can learn and update its prior knowledge, while entertaining complex spatiotemporal interactions with other systems such as the semantic memory system, spatial attention, oculomotor system, the salience network, decision making, and other domain-general executive functions (Figure 4). One organizational principle relevant to this framework is the free energy principle [93]. This principle offers a unified theory to explain perception, learning and action. More specifically, it offers a generalization of (i) the Bayesian brain hypothesis with the brain as an inference machine that actively predicts and explains its sensations, (ii) the infomax principle about neuronal activity encoding sensory information in an efficient and parsimonious fashion, and (iii) that perception is an inevitable consequence of active exchange with the environment. The free energy principle assumes that an agent must have an implicit generative model of how causes produce sensory data [93]. For the human visual system, one can assume a system that integrates two sets of inputs: (i) generated prediction errors at low-level areas conveyed with forward driving connections, and (ii) constructed predictions at higher areas conveyed with backward connections. The ultimate goal is to use these predictions to explain away prediction errors in low-level areas (Figure 5).

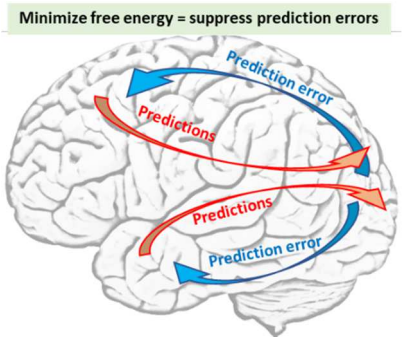


Figure 5: A schematic illustration of the interplay between predictions (red arrows) and prediction errors (blue arrows). Predictions are constructed according to an internal generative model. According to the free energy principle [93], action can reduce free-energy (i.e. prediction errors) by changing sensory input, whereas perception reduces free-energy by changing predictions.

Therefore, and following Friston's discussion of hierarchical models of sensory input [93, 94], it is possible to define IQA as a process that suppresses free energy. Here, the (input) data are the MR images and the (generative) model encodes how a typical artifact-free MR image should look like for the IQA process. Accordingly, an MR image with artifacts would be expected to increase prediction error (i.e. increase surprise). As discussed by Zhai and colleagues [95], one critical step in the design of IQA methods inspired by the free energy principle is the definition of the internal generative model. An optimal generative model should be able to approximate any given visual input with high precision [95]. Previous studies have suggested a couple of free-energy based IQA metrics [96-99], using in particular an autoregressive model as a generative model despite its high computational cost [95]. However, their usefulness and robustness for MRI have yet to be evaluated in a systematic way. Given the diversity in contrasts and artifacts in MRI, it would make sense to design free-energy inspired IQA methods specific to each artifact type. This is because a single generative model

that can *understand* such variety in MR contrasts and artifacts might be computationally impractical. This issue warrants future research.

VI. Implications for MRI at Ultralow and Ultrahigh Magnetic Fields

One emerging trend in neuroimaging with MRI is the possibility to scan patients at ultralow (<0.1 Tesla) or ultrahigh (>7 Tesla) magnetic fields. As severity and occurrence of different artifacts can vary with magnetic field strength, it is likely that current IQA methods, in particular NR-IQA and AI-based IQA methods, developed and tested on MR images collected at traditional field strengths (i.e. 1.5T and 3T), might not be applicable at ultralow or ultrahigh fields. A search in current literature shows that there is no systematic examination of IQA metrics on MRI images acquired at either ultralow or ultrahigh fields. Below, I describe what is so special about MRI at such field strengths.

MRI at ultrahigh fields is gaining in popularity, as it can increase SNR [100, 101]. High SNR offers the possibility to increase spatial resolution, yielding MR images at the submillimeter level; see for example recent development in layer-specific functional MRI. However, the downside is the increase in susceptibility artifacts with field strength, resulting in strong geometric distortions that are notoriously difficult to correct. Such distortions deteriorate with large head motion artifacts in patients and they tend to be spatially heterogenous, e.g. severe distortions in brain areas around air cavities [102]. IQA methods should thus be sensitive to the inherent heterogenous spatial distortions at ultrahigh fields. They could potentially be combined with recent AI tools that were implemented for susceptibility artifacts correction [103-105]. Furthermore, high-resolution MRI at ultrahigh fields will generate large images that are not always easy to handle with current IQA methods. For instance, a high-resolution anatomical image collected at 0.1 mm resolution took almost 2TB of raw k-space data [106], a storage size that needs to be multiplied by many folds when dealing with data from a large number of subjects. Computationally efficient IQA methods are thus needed for MRI at ultrahigh fields. This is an important issue as magnetic field strengths are expected to go even higher in the next decade [107].

Likewise, at the other end of the spectrum, MRI at ultralow fields is gaining in popularity, thanks to the emergence of portable MRI scanners [108, 109]. For instance, images collected with a low-field portable MRI (0.064 Tesla) were shown to be clinically useful for the evaluation of intracerebral hemorrhage [110]. However, as MRI signal decreases with magnetic field strength, MR images collected at ultralow fields have very poor SNR. Other limitations also include poor spatial resolution and reduced contrast between gray and white matter tissue [111]. A poor SNR would be a challenge for many IQA methods, in particular for NR-IQA methods including AI-based methods. More specifically, poor SNR may mask the manifestation of some subtle artifacts, making them very hard to spot with typical NR-IQA metrics. This domain is still in its infancy, and thus new IQA metrics are needed in the near future to assess MR images collected with portable MRI scanners at ultralow magnetic fields.

VII. Conclusion

Poor quality MR images can lessen diagnosis accuracy and increases costs of repeated scans. The majority of IQA methods for MRI were borrowed from existing methods developed for natural images. Consequently, many IQA methods are agnostic to the type of contrast and texture that specifically characterize MR images. In addition to FR-IQA methods that can be part of typical phantom imaging protocols, NR-IQA methods including AI-based methods are highly valuable. There is currently a growing interest in developing AI-based IQA methods that can be integrated or combined with AI image reconstruction and processing tools. This endeavor will be facilitated by recent initiatives to share large annotated datasets; see for example the MR-ART dataset about motion artifacts [112]. However, it is important that researchers adhere to current best practices when developing AI methods and reporting findings in order to improve reproducibility and replicability (e.g. see guidelines in [113, 114]). Similarly, new IQA metrics inspired by how the brain processes information will open new avenues for informed IQA metrics, including for instance methods based on the free energy principle. One aspect not particularly emphasized in this brief review is IQA for multimodal MRI. Image quality varies with MR modality, hence optimal IQA for multimodal MRI can either exploit variability (between-modality differences) or complementarity (image fusion techniques) across modalities [115]. Last but not least, tailored IQA methods for MR images acquired at ultralow or ultrahigh magnetic fields warrant future research.

Acknowledgment

The author would like the thank Prof Azeddine Beghdadi for the invitation to contribute a paper to EUVIP2023 meeting. This work was funded by Khalifa University (grant numbers RC2-2018-022 and FSU-2022-006).

References

[1] S.Y. Huang, R.T. Seethamraju, P. Patel, P.F. Hahn, J.E. Kirsch, A.R. Guimaraes, Body MR Imaging: Artifacts, k-Space, and Solutions, Radiographics, 35 (2015) 1439-1460.

[2] Y. Arefeen, O. Beker, J. Cho, H. Yu, E. Adalsteinsson, B. Bilgic, Scan-specific artifact reduction in k-space (SPARK) neural networks synergize with physics-based reconstruction to accelerate MRI, Magn Reson Med, 87 (2022) 764-780.

[3] K.H. Jin, J.Y. Um, D. Lee, J. Lee, S.H. Park, J.C. Ye, MRI artifact correction using sparse + low-rank decomposition of annihilating filter-based hankel matrix, Magn Reson Med, 78 (2017) 327-340.

[4] H. Gudbjartsson, S. Patz, The Rician distribution of noisy MRI data, Magn Reson Med, 34 (1995) 910-914.

[5] S. Aja-Fernandez, A. Tristan-Vega, C. Alberola-Lopez, Noise estimation in single- and multiple-coil magnetic resonance data based on statistical models, Magn Reson Imaging, 27 (2009) 1397-1409.

[6] O. Dietrich, J.G. Raya, S.B. Reeder, M.F. Reiser, S.O. Schoenberg, Measurement of signal-to-noise ratios in MR images: influence of multichannel coils, parallel imaging, and reconstruction filters, J Magn Reson Imaging, 26 (2007) 375-385.

[7] K. Murphy, J. Bodurka, P.A. Bandettini, How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration, Neuroimage, 34 (2007) 565-574.

[8] S. Gobbi, Y. Lee, I. Homolya, P.N. Tobler, T.A. Hare, Z. Nagy, On the reproducibility of in vivo temporal signal-to-noise ratio and its utility as a predictor of subject-level t-values in a functional magnetic resonance imaging study, Int J Imaging Syst Technol, 31 (2021) 1849-1860.

[9] W.S. Yamanashi, K.K. Wheatley, P.D. Lester, D.W. Anderson, Technical artifacts in magnetic resonance imaging, Physiol Chem Phys Med NMR, 16 (1984) 237-250.

[10] E.M. Bellon, E.M. Haacke, P.E. Coleman, D.C. Sacco, D.A. Steiger, R.E. Gangarosa, MR artifacts: a review, AJR Am J Roentgenol, 147 (1986) 1271-1281.

[11] S.A. Mirowitz, MR imaging artifacts. Challenges and solutions, Magn Reson Imaging Clin N Am, 7 (1999) 717-732.

[12] K. Krupa, M. Bekiesinska-Figatowska, Artifacts in magnetic resonance imaging, Pol J Radiol, 80 (2015) 93-106.

[13] S. Heiland, From A as in Aliasing to Z as in Zipper: Artifacts in MRI, Clinical Neuroradiology, 1 (2008) 25-36.
[14] C. Noda, B. Ambale Venkatesh, J.D. Wagner, Y. Kato, J.M. Ortman, J.A.C. Lima, Primer on Commonly Occurring MRI Artifacts and How to Overcome Them, Radiographics, 42 (2022) E102-E103.
[15] E. Pusey, C. Yoon, M.L. Anselmo, R.B. Lufkin, Aliasing artifacts in MR imaging, Comput Med Imaging Graph, 12 (1988) 219-224.
[16] L.F. Czervionke, J.M. Czervionke, D.L. Daniels, V.M. Haughton, Characteristic features of MR truncation artifacts, AJR Am J Roentgenol, 151 (1988) 1219-1228.
[17] M.H. Buonocore, L. Gao, Ghost artifact reduction for echo planar imaging using image phase correction, Magn Reson Med, 38 (1997) 89-100.
[18] M. Zaitsev, J. Maclaren, M. Herbst, Motion artifacts in MRI: A complex problem with many partial solutions, J Magn Reson Imaging, 42 (2015) 887-901.
[19] J.F. Schenck, The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds, Med Phys, 23 (1996) 815-850.
[20] C.J. Bakker, M.A. Moerland, R. Bhagwandien, R. Beersma, Analysis of machine-dependent and object-induced geometric distortion in 2DFT MR imaging, Magn Reson Imaging, 10 (1992) 597-608.
[21] B. Belaroussi, J. Milles, S. Carme, Y.M. Zhu, H. Benoit-Cattin, Intensity non-uniformity correction in MRI: existing methods and their validation, Med Image Anal, 10 (2006) 234-246.
[22] M.A. Gonzalez Ballester, A.P. Zisserman, M. Brady, Estimation of the partial volume effect in MRI, Med Image Anal, 6 (2002) 389-405.
[23] A. Macovski, Noise in MRI, Magn Reson Med, 36 (1996) 494-497.
[24] P. Coupe, J.V. Manjon, E. Gedamu, D. Arnold, M. Robles, D.L. Collins, Robust Rician noise estimation for MR images, Med Image Anal, 14 (2010) 483-493.
[25] S. Aja-Fernandez, G. Vegas-Sanchez-Ferrero, Statistical Analysis of Noise in MRI: Modeling, Filtering and Estimation, Springer Cham2016.
[26] L. Leveque, M. Outtas, H. Liu, L. Zhang, Comparative study of the methodologies used for subjective medical image quality assessment, Phys Med Biol, 66 (2021).
[27] R. Obuchowicz, M. Oszust, A. Piorkowski, Interobserver variability in quality assessment of magnetic resonance images, BMC Med Imaging, 20 (2020) 109.
[28] J.P. Woodard, M.P. Carley-Spencer, No-reference image quality metrics for structural MRI, Neuroinformatics, 4 (2006) 243-262.
[29] H.H. Barrett, J. Yao, J.P. Rolland, K.J. Myers, Model observers for assessment of image quality, Proc Natl Acad Sci U S A, 90 (1993) 9758-9765.
[30] A. Burgess, Image quality, the ideal observer, and human performance of radiologic decision tasks, Acad Radiol, 2 (1995) 522-526.
[31] X. He, S. Park, Model observers in medical imaging research, Theranostics, 3 (2013) 774-786.
[32] A.E. Burgess, Visual perception studies and observer models in medical imaging, Semin Nucl Med, 41 (2011) 419-436.
[33] M. Anton, A. Khanin, T. Kretz, M. Reginatto, C. Elster, A simple parametric model observer for quality assurance in computer tomography, Phys Med Biol, 63 (2018) 075011.
[34] M.A. Lago, C.K. Abbey, M.P. Eckstein, Medical image quality metrics for foveated model observers, J Med Imaging, 8 (2021) 041209.
[35] A. Viry, C. Aberle, T. Lima, R. Treier, S.T. Schindera, F.R. Verdun, D. Racine, Assessment of task-based image quality for abdominal CT protocols linked with national diagnostic reference levels, Eur Radiol, 32 (2022) 1227-1237.
[36] F.R. Verdun, D. Racine, J.G. Ott, M.J. Tapiovaara, P. Toroi, F.O. Bochud, W.J.H. Veldkamp, A. Schegerer, R.W. Bouwman, I.H. Giron, N.W. Marshall, S. Edyvean, Image quality in CT: From physical measurements to model observers, Phys Med, 31 (2015) 823-843.
[37] Y. Ding, Medical Image Quality Assessment, Visual Quality Assessment for Natural and Medical Image, Springer2018, pp. 215–264.
[38] S. Wang, Y. Ding, H. Dai, D. Qian, X. Yu, M. Zhang, Generalized relative quality assessment scheme for reconstructed medical images, Biomed Mater Eng, 24 (2014) 2865-2873.
[39] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans Image Process, 13 (2004) 600-612.
[40] S. Athar, Z. Wang, A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms, IEEE Access, 7 (2019) 140030-140070.
[41] S. Kastryulin, J. Zakirov, N. Pezzotti, D.V. Dylov, Image Quality Assessment for Magnetic Resonance Imaging, IEEE Access, (2023).
[42] M.J. Firbank, R.M. Harrison, E.D. Williams, A. Coulthard, Quality assurance for MRI: practical experience, Br J Radiol, 73 (2000) 376-383.
[43] C.C. Chen, Y.L. Wan, Y.Y. Wai, H.L. Liu, Quality assurance of clinical MRI scanners using ACR MRI phantom: preliminary results, J Digit Imaging, 17 (2004) 279-284.
[44] T. Torfeh, R. Hammoud, S. Paloor, Y. Arunachalam, S. Aouadi, N. Al-Hammadi, Design and construction of a customizable phantom for the characterization of the three-dimensional magnetic resonance imaging geometric distortion, J Appl Clin Med Phys, 22 (2021) 149-157.
[45] A. Walker, P. Chlap, T. Causer, F. Mahmood, J. Buckley, L. Holloway, Development of a vendor neutral MRI distortion quality assurance workflow, J Appl Clin Med Phys, 23 (2022) e13735.
[46] S. Wood, N. Krishnamurthy, T. Santini, S.B. Raval, N. Farhat, J.A. Holmes, T.S. Ibrahim, Design and fabrication of a realistic anthropomorphic heterogeneous head phantom for MR purposes, PLoS One, 12 (2017) e0183168.
[47] N. Crasto, A. Kirubarajan, D. Sussman, Anthropomorphic brain phantoms for use in MRI systems: a systematic review, MAGMA, 35 (2022) 277-289.
[48] A. Eskicioglu, P. Fisher, S. Chen, Image quality measures and their performance. , IEEE Trans Commun, 43 (1995) 2959-2965.
[49] U. Sara, M. Akter, M. Shorif Uddin, Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study, Journal of Computer and Communications, 7 (2019) 8-18.
[50] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans Image Process, 15 (2006) 3440-3451.
[51] K. Ding, K. Ma, S. Wang, E. Simoncelli, Image Quality Assessment: Unifying Structure and Texture Similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2022) 2567-2581.
[52] R. Reisenhofer, S. Bosse, G. Kutyniok, T. Wiegand, A Haar wavelet-based perceptual similarity index for image quality assessment, Signal Processing: Image Communication, 61 (2018) 33-43.
[53] L.S. Chow, H. Rajagopal, R. Paramesran, I. Alzheimer's Disease Neuroimaging, Correlation between subjective and objective assessment of magnetic resonance (MR) images, Magn Reson Imaging, 34 (2016) 820-831.
[54] I. Stepien, M. Oszust, A Brief Survey on No-Reference Image Quality Assessment Methods for Magnetic Resonance Images, J Imaging, 8 (2022).
[55] L. Kaufman, D.M. Kramer, L.E. Crooks, D.A. Ortendahl, Measuring signal-to-noise ratios in MR imaging, Radiology, 173 (1989) 265-267.
[56] S. Yu, G. Dai, Z. Wang, L. Li, X. Wei, Y. Xie, A consistency evaluation of signal-to-noise ratio in the quality assessment of human brain magnetic resonance images, BMC Med Imaging, 18 (2018) 17.
[57] R. Li, G. Dai, Z. Wang, S. Yu, Y. Xie, Using signal-to-noise ratio to connect the quality assessment of natural and medical images, Tenth International Conference on Digital Image Processing (ICDIP 2018), Shanghai, China, 2018, pp. 08064Q.
[58] B. Mortamet, M.A. Bernstein, C.R. Jack, Jr., J.L. Gunter, C. Ward, P.J. Britson, R. Meuli, J.P. Thiran, G. Krueger, I. Alzheimer's Disease Neuroimaging, Automatic quality assessment in structural brain magnetic resonance imaging, Magn Reson Med, 62 (2009) 365-372.
[59] M.A. Saad, A.C. Bovik, C. Charrier, DCT statistics model-based blind image quality assessment., IEEE ICIP, 1 (2011) 3093-3096.
[60] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans Image Process, 21 (2012) 4695-4708.
[61] J. Jang, K. Bang, H. Jang, D. Hwang, I. Alzheimer's Disease Neuroimaging, Quality evaluation of no-reference MR images using multidirectional filters and image statistics, Magn Reson Med, 80 (2018) 914-924.
[62] L.S. Chow, H. Rajagopal, Modified-BRISQUE as no reference image quality assessment for structural MR images, Magn Reson Imaging, 43 (2017) 74-87.
[63] M. Osadebey, M. Pedersen, D. Arnold, K. Wendel-Mitoraj, No-reference quality measure in brain MRI images using binary operations, texture and set analysis, IET Image Processing, 11 (2017) 672-684.
[64] W. Hou, X. Gao, D. Tao, X. Li, Blind image quality assessment via deep learning, IEEE Trans Neural Netw Learn Syst, 26 (2015) 1275-1286.
[65] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal, Image and Video Processing, 12 (2018) 355-362.
[66] H. Talebi, P. Milanfar, NIMA: Neural Image Assessment, IEEE Trans Image Process, DOI 10.1109/TIP.2018.2831899 (2018).
[67] M. Rehman, I.F. Nizami, M. Majid, DeepRPN-BIQA: Deep architectures with region proposal network for natural-scene and screen-content blind image quality assessment, Displays, 71 (2022) 102101.

[68] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, Y. Yang, MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1191-1200.
[69] J. You, J. Korhonen, Attention integrated hierarchical networks for no-reference image quality assessment, Journal of Visual Communication and Image Representation, 82 (2022) 103399.
[70] M. Oszust, M. Bielecka, A. Bielecki, I. Stepien, R. Obuschowicz, A. Piorkowski, Blind image quality assessment of magnetic resonance images with statistics of local intensity extrema, Information Sciences, 606 (2022) 112-125.
[71] X. Yang, F. Li, H. Liu, A Survey of DNN Methods for Blind Image Quality Assessment, IEEE Access, 7 (2019) 123788-123806.
[72] S. Bosse, D. Maniry, K.R. Muller, T. Wiegand, W. Samek, Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment, IEEE Trans Image Process, 27 (2018) 206-219.
[73] J. Kim, A.D. Nguyen, S. Lee, Deep CNN-Based Blind Image Quality Predictor, IEEE Trans Neural Netw Learn Syst, 30 (2019) 11-24.
[74] M. Liu, J. Huang, D. Zeng, X. Ding, J. Paisley, A Multiscale Approach to Deep Blind Image Quality Assessment, IEEE Trans Image Process, 32 (2023) 1656-1667.
[75] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-End Blind Image Quality Assessment Using Deep Neural Networks, IEEE Trans Image Process, 27 (2018) 1202-1213.
[76] X. Lan, M. Zhou, X. Xu, X. Wei, X. Liao, H. Pu, J. Luo, T. Xiang, B. Fang, Z. Shang, Multilevel Feature Fusion for End-to-End Blind Image Quality Assessment, IEEE Transactions on Broadcasting, (2023) 1-11.
[77] S.J. Sujit, I. Coronado, A. Kamali, P.A. Narayana, R.E. Gabr, Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks, J Magn Reson Imaging, 50 (2019) 1260-1267.
[78] A. Gupta, A.R. Sadri, S.E. Viswanath, P. Tiwari, Quality assessment of brain MRI scans using a dense neural network model and image metrics, Proc. SPIE 11312, Medical Imaging: Physics of Medical Imaging, 2020.
[79] K. Lei, A.B. Syed, X. Zhu, J.M. Pauly, S.S. Vasanawala, Artifact- and content-specific quality assessment for MRI with image rulers, Med Image Anal, 77 (2022) 102344.
[80] I. Stepien, M. Oszust, No-Reference Image Quality Assessment of Magnetic Resonance images with multi-level and multi-model representations based on fusion of deep architectures, Engineering Applications of Artificial Intelligence, 123 (2023) 106283.
[81] I. Fantini, C. Yasuda, M. Bento, L. Rittner, F. Cendes, R. Lotufo, Automatic MR image quality evaluation using a Deep CNN: A reference-free method to rate motion artifacts in neuroimaging, Comput Med Imaging Graph, 90 (2021) 101897.
[82] A. Largent, K. Kapse, S.D. Barnett, J. De Asis-Cruz, M. Whitehead, J. Murnick, L. Zhao, N. Andersen, J. Quistorff, C. Lopez, C. Limperopoulos, Image Quality Assessment of Fetal Brain MRI Using Multi-Instance Deep Learning Methods, J Magn Reson Imaging, 54 (2021) 818-829.
[83] Z. Wang, A.C. Bovik, L. Lu, Why is image quality assessment so difficult?, IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 2002.
[84] G. Ginesu, F. Massidda, D.D. Giusto, A multi-factors approach for image quality assessment based on a human visual system model, Signal Processing: Image Communication, 21 (2006) 316-333.
[85] X. Gao, W. Lu, D. Tao, W. Liu, Image quality assessment and human visual system, Proceedings of SPIE - The International Society for Optical Engineering, 2010.
[86] J. Yang, C. Hou, R. Xu, J. Lei, New metric for stereo image quality assessment based on HVS, Int J Imaging Syst Technol, 20 (2010) 301-307.
[87] N. Sinha, A.G. Ramakrishnan, Quality assessment in magnetic resonance images, Crit Rev Biomed Eng, 38 (2010) 127-141.
[88] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans Image Process, 20 (2011) 2378-2386.
[89] H.W. Chang, Q.W. Zhang, Q.G. Wu, Y. Gan, Perceptual image quality assessment by independent feature detector, Neurocomputing, 151 (2015) 1142-1152.
[90] B. Wang, Z. WAng, Y. Liao, X. Lin, HVS-based structural similarity for image quality assessment, 9th International Conference on Signal Processing, Beijing, China, 2008.
[91] J. You, Z. Zhang, Visual Mechanisms Inspired Efficient Transformers for Image and Video Quality Assessment, in: K. Arai (Ed.) Advances in Information and Communication. FICC 2023. Lecture Notes in Networks and Systems, Springer, Cham2023, pp. 455-473.
[92] D.J. Kravitz, K.S. Saleem, C.I. Baker, L.G. Ungerleider, M. Mishkin, The ventral visual pathway: an expanded neural framework for the processing of object quality, Trends Cogn Sci, 17 (2013) 26-49.
[93] K.J. Friston, The free-energy principle: a unified brain theory?, Nat Rev Neurosci, 11 (2010) 127-138.
[94] K. Friston, The free-energy principle: a rough guide to the brain?, Trends Cogn Sci, 13 (2009) 293-301.
[95] G. Zhai, X. Min, N. Liu, Free-energy principle inspired visual quality assessment: An overview, Digital Signal Processing, 91 (2019) 11-20.
[96] G. Zhai, X. Wu, X. Yang, W. Lin, W. Zhang, A psychovisual quality metric in free-energy principle, IEEE Trans Image Process, 21 (2012) 41-52.
[97] K. Gu, G. Zhai, X. Yang, W. Zhang, Using Free Energy Principle For Blind Image Quality Assessment, IEEE Transactions on Multimedia, 17 (2015) 50-63.
[98] N. Liu, G. Zhai, Free energy adjusted peak signal to noise ratio (fea-psnr) for image quality assessment, Sensing and Imaging, 18 (2017) 11.
[99] W. Zhu, G. Zhai, X. Min, M.-C. Hu, J. Liu, G. Guo, X. Yang, Multi-channel de-composition in tandem with free-energy principle for reduced-reference image quality assessment, IEEE Transactions on Multimedia, DOI (2019).
[100] R. Pohmann, O. Speck, K. Scheffler, Signal-to-noise ratio and MR tissue parameters in human brain imaging at 3, 7, and 9.4 tesla using current receive coil arrays, Magn Reson Med, 75 (2016) 801-809.
[101] M. Cosottini, L. Roccatagliata, Neuroimaging at 7 T: are we ready for clinical transition?, Eur Radiol Exp, 5 (2021) 37.
[102] T. Yamamoto, M. Fukunaga, S.K. Sugawara, Y.H. Hamano, N. Sadato, Quantitative Evaluations of Geometrical Distortion Corrections in Cortical Surface-Based Analysis of High-Resolution Functional MRI Data at 7T, J Magn Reson Imaging, 53 (2021) 1220-1234.
[103] S.T.M. Duong, S.L. Phung, A. Bouzerdoum, S.P. Ang, M.M. Schira, Correcting Susceptibility Artifacts of MRI Sensors in Brain Scanning: A 3D Anatomy-Guided Deep Learning Approach, Sensors, 21 (2021).
[104] Y. Qiao, Y. Shi, Unsupervised Deep Learning for FOD-Based Susceptibility Distortion Correction in Diffusion MRI, IEEE Trans Med Imaging, 41 (2022) 1165-1175.
[105] B. Zahneisen, K. Baeumler, G. Zaharchuk, D. Fleischmann, M. Zeineh, Deep flow-net for EPI distortion estimation, Neuroimage, 217 (2020) 116886.
[106] B.L. Edlow, A. Mareyam, A. Horn, J.R. Polimeni, T. Witzel, M.D. Tisdall, J.C. Augustinack, J.P. Stockmann, B.R. Diamond, A. Stevens, L.S. Tirrell, R.D. Folkerth, L.L. Wald, B. Fischl, A. van der Kouwe, 7 Tesla MRI of the ex vivo human brain at 100 micron resolution, Sci Data, 6 (2019) 244.
[107] A. Nowogrodzki, The world's strongest MRI machines are pushing human imaging to new limits, Nature, 563 (2018) 24-26.
[108] C.Z. Cooley, P.C. McDaniel, J.P. Stockmann, S.A. Srinivas, S.F. Cauley, M. Sliwiak, C.R. Sappo, C.F. Vaughn, B. Guerin, M.S. Rosen, M.H. Lev, L.L. Wald, A portable scanner for magnetic resonance imaging of the brain, Nat Biomed Eng, 5 (2021) 229-239.
[109] Y. Liu, A.T.L. Leong, Y. Zhao, L. Xiao, H.K.F. Mak, A.C.O. Tsang, G.K.K. Lau, G.K.K. Leung, E.X. Wu, A low-cost and shielding-free ultra-low-field brain MRI scanner, Nat Commun, 12 (2021) 7238.
[110] M.H. Mazurek, B.A. Cahn, M.M. Yuen, A.M. Prabhat, I.R. Chavva, J.T. Shah, A.L. Crawford, E.B. Welch, J. Rothberg, L. Sacolick, M. Poole, C. Wira, C.C. Matouk, A. Ward, N. Timario, A. Leasure, R. Beekman, T.J. Peng, J. Witsch, J.P. Antonios, G.J. Falcone, K.T. Gobeske, N. Petersen, J. Schindler, L. Sansing, E.J. Gilmore, D.Y. Hwang, J.A. Kim, A. Malhotra, G. Sze, M.S. Rosen, W.T. Kimberly, K.N. Sheth, Portable, bedside, low-field magnetic resonance imaging for evaluation of intracerebral hemorrhage, Nat Commun, 12 (2021) 5119.
[111] T.C. Arnold, C.W. Freeman, B. Litt, J.M. Stein, Low-field MRI: Clinical promise and challenges, J Magn Reson Imaging, 57 (2023) 25-44.
[112] A. Narai, P. Hermann, T. Auer, P. Kemenczky, J. Szalma, I. Homolya, E. Somogyi, P. Vakli, B. Weiss, Z. Vidnyanszky, Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans, Sci Data, 9 (2022) 630.
[113] M.L. Seghier, Ten simple rules for reporting machine learning methods' implementation and evaluation on biomedical data, Int J Imaging Syst Technol, 32 (2022) 5-11.
[114] L.M. Stevens, B.J. Mortazavi, R.C. Deo, L. Curtis, D.P. Kao, Recommendations for Reporting Machine Learning Analyses in Clinical Research, Circulation: Cardiovascular Quality and Outcomes, 13 (2020) 782-793.
[115] L. Tang, Y. Hui, H. Yang, Y. Zhao, C. Tian, Medical image fusion quality assessment based on conditional generative adversarial network, Front Neurosci, 16 (2022) 986153.

Deep Learning Models for Low Dose CT Simulation

Lumi XIA
IETR - UMR 6164 F-35000
Univ. Rennes, INSA Rennes
Rennes, France
Lumi.Xia@insa-rennes.fr

Meriem OUTTAS
IETR - UMR 6164 F-35000
Univ. Rennes, INSA Rennes
Rennes, France
Meriem.Outtas@insa-rennes.fr

Lu ZHANG
IETR - UMR 6164 F-35000
Univ. Rennes, INSA Rennes
Rennes, France
Lu.Ge@insa-rennes.fr

Eric FRAMPAS
Central Department of Radiology and Medical Imaging
Universitary Hospital CHU Nantes
Nantes, France
eric.frampas@chu-nantes.fr

Olivier DEFORGES
IETR - UMR 6164 F-35000
Univ. Rennes, INSA Rennes
Rennes, France
Olivier.Deforges@insa-rennes.fr

***Abstract*—Finding the optimal balance between reducing radiation risk for patients and preserving image quality in CT imaging remains a challenging task. To explore the impact of lower radiation doses on CT image quality, the simulation of low dose CT images is highly demanded. Traditional methods often rely on generating Gaussian/Poisson noise on raw data using mathematical models for specific scanners, which requires access to many medical resources. Inspired from image denoising models, we propose two deep learning models (ResNet and U-Net) for low dose CT images simulation, among which the U-Net model demonstrated better performance. Moreover, the trained U-Net model proves its applicability to untrained CT images from various scanners, showcasing its potential as a versatile and generic low dose CT simulation tool. By bypassing the need for complex phantom experiments and mathematical modeling, deep learning for low dose CT simulation emerges as a promising and powerful approach, akin to solving the inverse problem of well-established image denoising techniques.**

***Index Terms*—Deep Learning, Image Processing, Computed Tomography**

I. Introduction

X-ray Computed Tomography (CT) is an advanced medical imaging technology that offers detailed 3D visualization of the human body, providing rich diagnostic information. With its principle of detecting the attenuation of the 360 degree rotating X-Ray when passing tissues with different densities during a continuous time, the patients receive about 100 to 500 times more radiation dose than in convention X-ray [1]. Thus, strict protocols exist for all CT exams, which is a difficult compromise between the quality and the radiation risk. Is it possible to reduce radiation dose while maintaining diagnostic quality? What is the lowest dose level at which CT scans can reliably characterize specific pathologies? These questions remain unanswered. Consequently, low dose CT (LDCT) is undoubtedly indispensable in the pursuit of balancing radiation exposure and diagnostic quality.

Regarding the difficulty and restriction of conducting multiple dose CT scans experiments in clinics, developing convenient and reliable LDCT simulation tools becomes a crucial need. Traditionally, the addition of Gaussian or Poisson noise to raw data has been employed to generate degraded sinograms, specifically tailored to the characteristics of a particular CT scanner. The resulting reconstructed noisy image is then considered as the simulated LDCT [2]–[4]. Recent advancements have brought forth simulation methods in the image domain that no longer require access to raw CT data. These methods can be categorized into two approaches, based on how they generate noise associated with the degradation of LDCT images compared to normal dose images. The first approach involves direct noise estimation in the image domain [5], [6], and the second one entails reconstructing noise from simulated noise sinograms [7]. Both approaches rely on intricate mathematical models, necessitating a series of blank scans or even phantom experiments, whose complexity and specificity actually limit their generic application.

When those direct methods for LDCT simulation are unrealizable, can we simulate these LDCT in another way? Inspired from abundant studies of LDCT denoising / reconstruction [8]–[12] where the LDCT is considered as the noisy version of high dose ones (HDCT), we had the idea to treat this simulation problem as the inverse problem of denoising problem. Denoising methods can be broadly divided into two groups: model-based optimization methods and convolutional neural network (CNN)-based methods [13]. Among these, we are particularly interested in CNN-based methods considering our limited resources. Firstly introduced by He in 2016 [14], residual learning has been successfully applied to denoise both natural and medical images [15], [16]. While U-Net is predominantly known for image segmentation [17], it has recently gained increasing attention in the field of image denoising as well [18], [19].

In 2021, Niu *et al.* proposed a Noise-Entangled Generative Adversarial Network (NE-GAN) to simulate LDCT from HDCT, where the noise image generated from HDCT are

979-8-3503-4218-5/23/$31.00©2023 IEEE

scaled and added to the latter ones to form the simulated LDCT [20]. This is the first study which applied this inverse-denoising principle for LDCT simulation with satisfactory results. However, this study has been conducted on a simple-source dataset, where all CT images were collected at the same time by the same scanner [21], the reproducibility of this method is then not sure; besides, the results were presented only visually, have not been evaluated in a more objective way.

In this paper, we employed two types of CNN models, Deep Residual Network (ResNet) and U-Net, which give most interesting results among many denoising models we have experimented, for our LDCT simulation; then we have successfully developed a generic LDCT simulation tool based on the best-performance trained model. The code and the example of dataset are available on GitHub [1]. Our contributions are demonstrated in two aspects: First, validated the possibility to treat the LDCT simulation problem as inverse problem of image denoising, by transforming two commonly-used deep learning (DL) denoising models as LDCT simulation models. Second, proposed a generic DL-based LDCT simulation tool, which is applicable for any CT scans, without knowing the parameters of scanners. To our knowledge, this is one of the earliest works on this specific subject.

II. Methods and Materials

Fig.1 illustrates the workflow of our study, consisting of two main stages: I. Model Training: two DL networks inspired from denoising models are experimented to realize the low dose simulation task on CT images of one healthy subject (Subject 0); II. Model Application: the best-performance trained model is employed to simulate desired low dose image for any patient X (different than the subject 0 in model training stage).

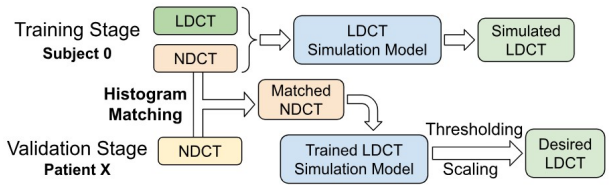


Fig. 1: Workflow of proposed low dose simulation model

A. Dataset

The dataset is retrospectively collected from Nantes university hospital, composed of two parts: Part I, CT scan of subject 0, a healthy subject, containing both the normal and low dose images; Part II, multi-sources CT scans from 77 patients with adrenal lesion, containing only the normal dose images.

The healthy subject underwent examinations with two different tensions: 120 KVp and 80 KVp, referring to normal dose CT (NDCTs) (computed tomography dose index, CTDI=10.8), and LDCTs (CTDI=7.7). This experiment has been conducted by an experimented radiologist and in accordance with ethical standards, thereby without any ethical concerns. As a result, Part I dataset consists of 227 axial CT images (dimension 512*512 pixels, 1mm thickness, using non-contrast enhanced sequence, 80 multidetector CT, Aquilion PRIME, Canon Medical Systems Corporation, Otawara, Japan). We obtained 109 pairs of NDCT and LDCT images by matching their slice locations as recorded in the metadata. Among these pairs, 87 were used for training data, and 22 were used for testing. Before feeding them into CNN models, all images were divided into 64x64 patches.

The CT scans in Part II were collected from anonymized adrenal lesion patients across different years (2005-2022) and were conducted by various scanners, as shown in the Table.I.

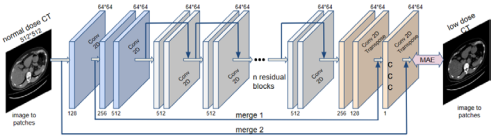
TABLE I: CT scanners used in Dataset part II

Manufacturer	Model Name	Number of patients
SIEMENS	Sensention 16	15
	SOMATOM Definition AS	3
GE	LightSpeed VCT	8
MEDICAL SYSTEMS	BrightSpeed	8
	BrightSpeed QX/i	3
	Optima CT660	3
TOSHIBA	Acquilion PRIME	28
PHILIPS	Ingenuity Flex	7

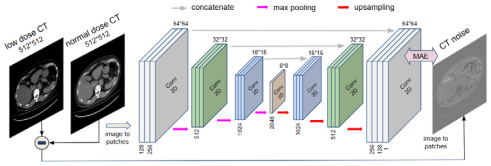
B. Deep Learning Models

1) ResNet denoising model: Inspiring from a ResNet denoising model for natural image denoising [22], we applied a network with n residual blocks on our dataset, with NDCTs as input and corresponding LDCTs as output, as shown in Fig.2 (a). Here we set $n = 8$, the same as the default setting in the original model.

2) U-Net denoising model: Adopted from a U-Net model for reduced dose MRI restoration [23], we applied the model shown in Fig.2 (b), with NDCTs as input and corresponding noise image as output, which refers to the image difference between ND/LDCT image pair.



(a) ResNet model



(b) U-Net model

Fig. 2: Deep learning models for LDCT simulation

Both models are trained (1500 epochs) on the dataset mentioned in II-A, using Mean Absolute Error (MAE) as the loss function, Adam as optimizer. The GPU GP102 (GeForce

[1]https://github.com/LumiereSummer/LowDoseSimulation

GTX 1080 Ti) and TU102 (GeForce RTX 2080 Ti Rev. A) have been utilised in our training process.

C. Evaluation Metrics

1) Intra-image evaluation metrics: To compare the characteristics of simulated LDCT images with those of real ND/LDCT image pairs, the mean and standard deviation (SD) of CT numbers are calculated, as well as image entropy. Mean CT numbers reflects luminance levels, SD tells the variations of CT numbers. LDCT are generally darker than NDCT when displayed in the same window, with "rougher" texture. Entropy of images is considered as a measure of both image noise and resolution [24], which could also differentiate LDCT from NDCT.

2) Inter-image evaluation metrics: Apart the intra-image evaluation, the inter-image metrics are also calculated. Beside the most commonly used quantitative measurements for evaluating image denoising methods [25] as Root mean squared errors (RMSE), peak signal-to-noise ratio (PSNR) and structure similarity index measure (SSIM), here we also borrowed two metrics often used for image registration, histogram correlation (HC) [26] and mutual information (MI) [27], [28] as similarity metrics. A higher HC or MI value indicates a greater similarity between the two images. Moreover, the power spectra of images are also compared to reflect the noise level.

D. Model Application

1) Histogram Matching: As mentioned in section II-A, the 77 CT scans in the dataset Part II are generated by different scanners, giving them different imaging parameters and characteristics. To make these CT images compatible with the trained U-Net model, a standardization step as preprocessing is necessary.

Histogram matching (*i.e.* histogram equalization) is employed here to fulfill this goal, by matching the histogram of different images to that of original input images (NDCT of Subject 0, *i.e.* reference images). Then the trained model could work on these standardized image (*i.e.* matched image) properly. Since the histogram of an image is related to its content, here we need to first pair the original image with the reference image, by searching the smallest structure distance of the former one and all reference images. Therefore, these two paired images share similar anatomic structure and their histogram difference majorly comes from the different imaging characteristics, which will be resolved by the following histogram matching step, as shown in Fig.3.

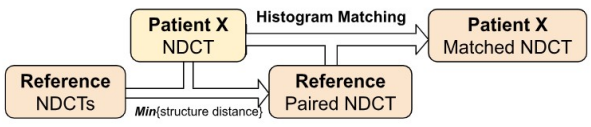
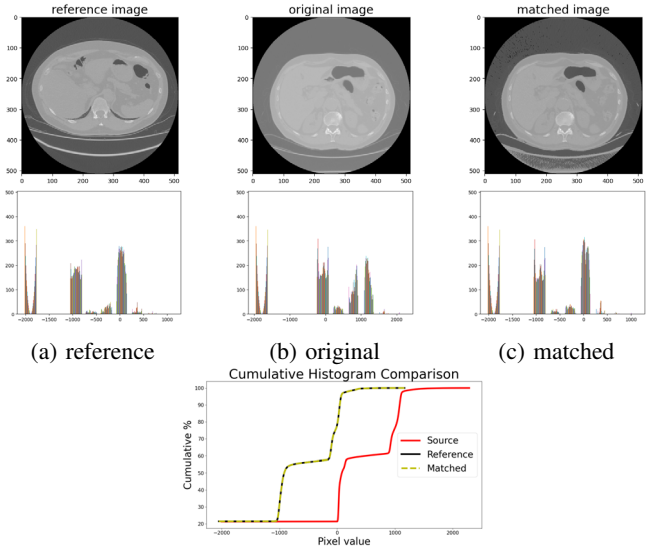


Fig. 3: Diagram of Histogram Matching

Fig.4 presents an example of histogram matching. From (a) to (c), the first row shows the CT images, and the second row illustrates the corresponding histogram; Cumulative histogram presented in (d) demonstrated the principle of histogram matching: mapping each pixel value in original image to the corresponding pixel value with the same probability in the reference image. As we can see, the matched image has a grayer air background, and the contrast of main structures has been slightly enhanced.



(a) reference (b) original (c) matched

(d) Cumulative Histogram

Fig. 4: Example of Histogram matching (on CT scan of Patient 26, generated by BrightSpeed model from GE MEDICAL SYSTEMS manufacturer, no window applied).

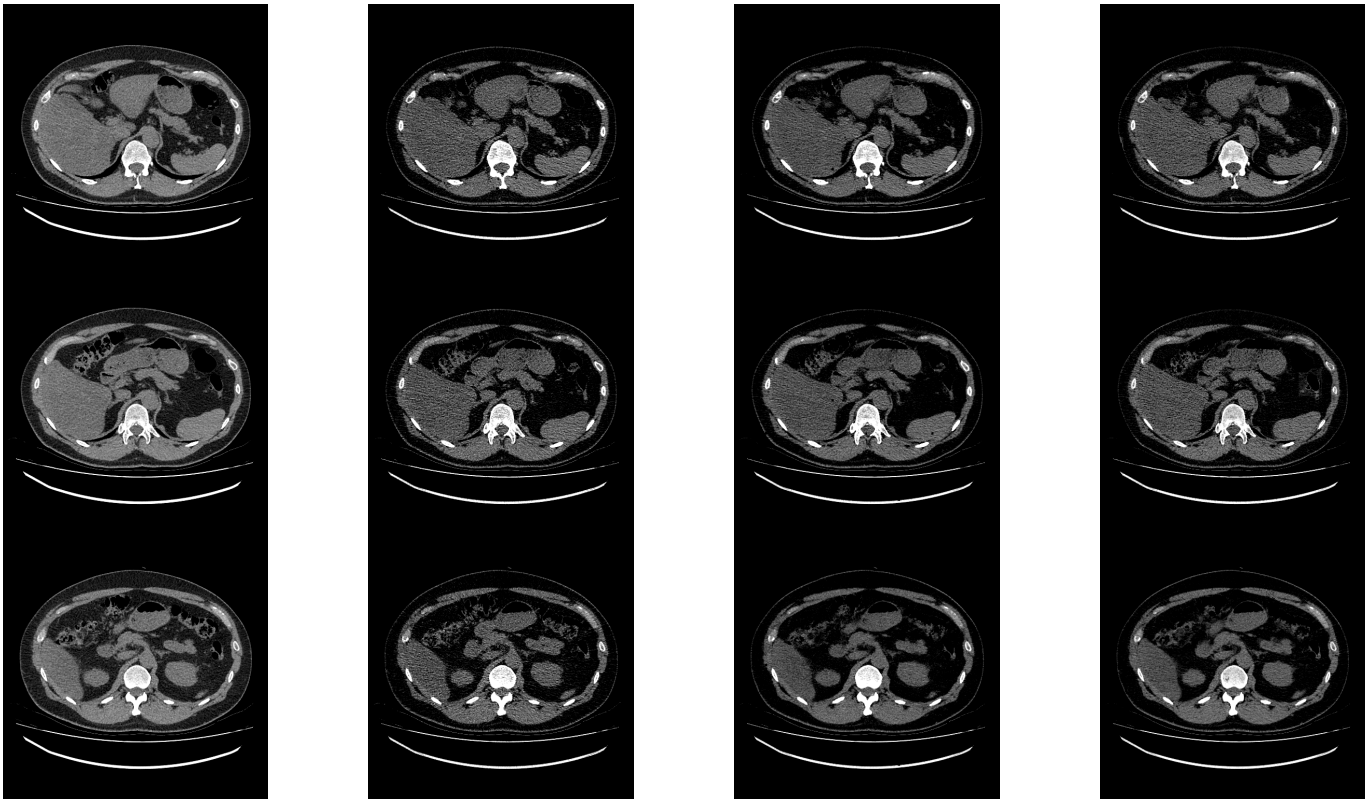
2) Thresholding and Scaling: The output of trained U-Net model also needs post-processing to better serve a specific low dose simulation task.

As introduced in section II-B2, the direct output of U-Net model is the noise image of an input NDCT, which might contain some pixel anomalies (caused by patches borders etc.), Thresholding can solve this problem by assigning those anomalies to the threshold value, as in Eq.1.

$$\begin{aligned} I_{simLD,\lambda} &= I_{NDCT} - \lambda * thres[I_{noise}] \\ I_{simLD} &= \max_{\lambda} HC(I_{LDCT}, I_{simLD,\lambda}) \end{aligned} \tag{1}$$

The noise image after thresholding will be subtracted from the original NDCT images of patient X, to form the corresponding simulated LDCT. To imitate different doses CT image, we introduced a scaling factor λ, which is determined by finding the maximal value of HC of the desired LDCT image I_{LDCT} and simulated image I_{simLD}, as illustrated in Fig.6.

In our case, the desired LDCT images is the reference LDCTs, *i.e.*, the CT scans of 80KVP (CTDI=7.7), about 70% dose of the NDCT ones (120KVP, CTDI=10.8). For avoiding the influence of image content on the value of HC, here we also need to return to the same image pair found in section II-D1 in the research of the best λ, so that the histogram correlation we calculate depends mainly on the texture of image (*i.e.* noise level), not on the different structures.



(a) NDCT (b) LDCT (c) ResLD (d) UnetLD

Fig. 5: Simulated LDCTs from both models after 1500 epochs, with comparison of the real ND/LDCT pair (displayed in window [-150,250]).

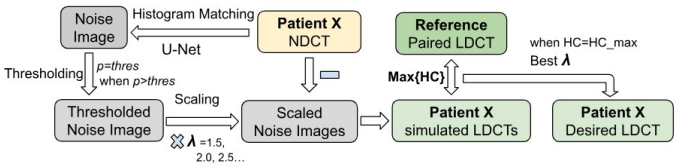


Fig. 6: Diagram of Thresholding and Scaling

III. Results and Discussion

The results are presented as two parts: the quantitative and qualitative evaluation of proposed models' performance on Subject 0, where the ground truth (real LDCT) is available; the application of the best-performance trained model on any Patient X, where the desired dose level could be customized.

A. Model Evaluation on Subject 0

1) Quantitative Evaluation: Using the evaluation metrics in section.II-C, the statistical analysis of these simulated images has been done for both models, as in Table II. In our intra-image evaluation, both ResLD and UnetLD are quite similar to LDCT (difference$<6\%$; in terms of luminance (mean), variance (SD) of CT numbers, the performance of ResNet model is slightly better than U-Net, while UnetLD's entropy difference with LDCT is smaller that of ResLD. Correspondingly, in inter-image evaluation, U-Net model shows better performance than U-Net, either in terms of traditional metrics like RMSE, PSNR and SSIM, or in terms of the nouvel metrics HC amd MI, where the performance difference of two models is more obvious.

TABLE II: Evaluation of simulated LDCT generated from ResNet model and U-Net model, comparing with real ND/LDCT pairs

(a) Intra-image evaluation

Image	Mean	SD	Entropy
NDCT	-92.240	89.589	3.000
LDCT	-109.926	79.069	2.331
ResLD	-112.253	76.278	2.198
UnetLD	-112.567	75.293	2.230
ResLD-LD(%)	+2.117	-3.530	-5.706
UnetLD-LD(%)	+2.403	-4.776	-4.333

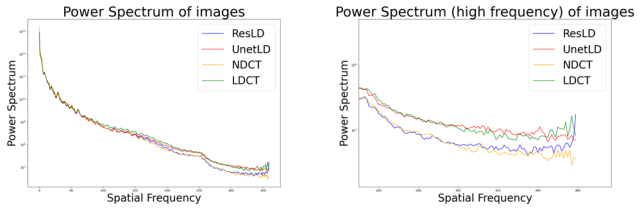
(b) Inter-image evaluation

Image	RMSE	SSIM	PSNR	HC ⇑	MI ⇑
LD vs ND	43.857	0.597	15.311	0.909	0.720
ResLD vs ND	40.199	0.606	16.062	0.744	0.780
UnetLD vs ND	39.661	0.623	16.183	0.773	0.907
ResLD vs LD	25.492	0.707	20.021	0.953	0.734
UnetLD vs LD	24.143	0.748	20.508	0.964	0.769

In other words, the indices calculated the global absolute difference of CT values could hardly differential the performances of these two models, while the indices focus more on the total distribution are more in favor of the U-Net model.

2) Qualitative evaluation: Three example images of simulated LDCT from both models are presented in Fig.5, as well as the corresponding ND/LDCT pairs. Apparently both simulated images are more noisy than NDCT image. From the point of view of one radiologist expert, the U-Net simulation results are also more similar to the real LDCTs than the ResNet ones, when assessing only by visualization.

Fig.7 shows the power spectra of the first example images in Fig.5, we can see that the power spectrum of UnetLD is much more close to the one of LDCT than to NDCT, on the contrary of the spectrum of ResLD, specially in high frequencies, which can reflect the noise level of images.



(a) all frequencies (b) high frequencies

Fig. 7: power spectra comparison of simulated LDCTs and real ND/LDCT pair

Considering all the information above, we can conclude that the U-Net model outperforms the ResNet model in LDCT simulation, as it can better imitate the texture of noise. This phenomenon might be attributed to two factors: I. the method of learning from image difference of ND/LDCT pairs (noise image) is simply more effective than the direct image translation from NDCT to LDCT; II. the incorporation of max-pooling and up-sampling steps in U-Net model emphasized the significant features of the noise image, thereby improved the simulation performance. Nevertheless, these hypothesis still need to be further investigated, and the results of such studies will be reported in our future work.

B. Model Application on Patient X

Fig.8 presents the simulated low dose images of trained U-Net model with different scaling factors, referring to different doses. From left to right, there are original CT images (NDCT), and simulated images at three different dose levels ($\lambda = 1.5$, $\lambda = 2.0$, $\lambda = 2.5$, among which the last column images showed the greatest histogram correlation ($HC > 0.999$) with the LDCT of Subject 0. From top to bottom, they are CT images from Patient 26, Patient 48 and Patient 75, from three different scanners (BrightSpeed model from GE MEDICAL SYSTEMS, Acquilion PRIME model from TOSHIBA, and Ingenuity Flex model from PHILIPS, respectively). We can easily observe that our model has successfully introduced noticeable quality degradation on all these CT images, despite their different sources.

Theoretically, with the same procedure presented in section II-D, this DL-based simulation tool could be employed to realize any other desired low dose CT images without the knowledge of scanners, as long as we have at least one example of this desired dose CT scan (serving as the simulation target I_{LDCT} in Eq.1). Further analysis will be conducted to assess the stability of this model in the presence of artifacts such as streaking, rings etc. [29].

IV. Conclusion

In this paper we proposed a solution to simulate low-dose CT images without requiring raw data or scanner parameters, as the inverse problem of image denoising, and proved its feasibility by transforming two DL denoising models (ResNet and U-Net model) into LDCT simulation models. The best-performance model (U-Net) has been further proved as a generic tool for LCDT simulation, applicable for any patient and desired dose (as long as one desired dose CT scan is available). Furthemore, we share the trained generic LDCT simulation model as well as the reference LDCT-NDCT pairs for further studies.

More subjective assessments for specific pathology (adrenal lesions) are expected to be done in the near future. Further study on the influence of radiation dose on CT image quality in terms of its diagnostic performance is now ongoing, based on the LDCT dataset generated by this deep learning model.

References

[1] R. Smith-Bindman, "Is computed tomography safe?" *The New England Journal of Medicine*, vol. 363, no. 1, pp. 1–4, 2010.

[2] W. J. Veldkamp, L. J. Kroft, J. P. A. van Delft, and J. Geleijns, "A technique for simulating the effect of dose reduction on image quality in digital chest radiography," *Journal of Digital Imaging*, vol. 22, no. 2, pp. 114–125, 2009.

[3] S. Žabić, Q. Wang, T. Morton, and K. M. Brown, "A low dose simulation tool for ct systems with energy integrating detectors," *Medical physics*, vol. 40, no. 3, p. 031102, 2013.

[4] D. Zeng, J. Huang, Z. Bian, S. Niu, H. Zhang, Q. Feng, Z. Liang, and J. Ma, "A simple low-dose x-ray ct simulation from high-dose scan," *IEEE transactions on nuclear science*, vol. 62, no. 5, pp. 2226–2233, 2015.

[5] S. E. Divel and N. J. Pelc, "Accurate image domain noise insertion in ct images," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1906–1916, 2019.

[6] M. Elhamiasl and J. Nuyts, "Low-dose x-ray ct simulation from an available higher-dose scan," *Physics in Medicine & Biology*, vol. 65, no. 13, p. 135010, 2020.

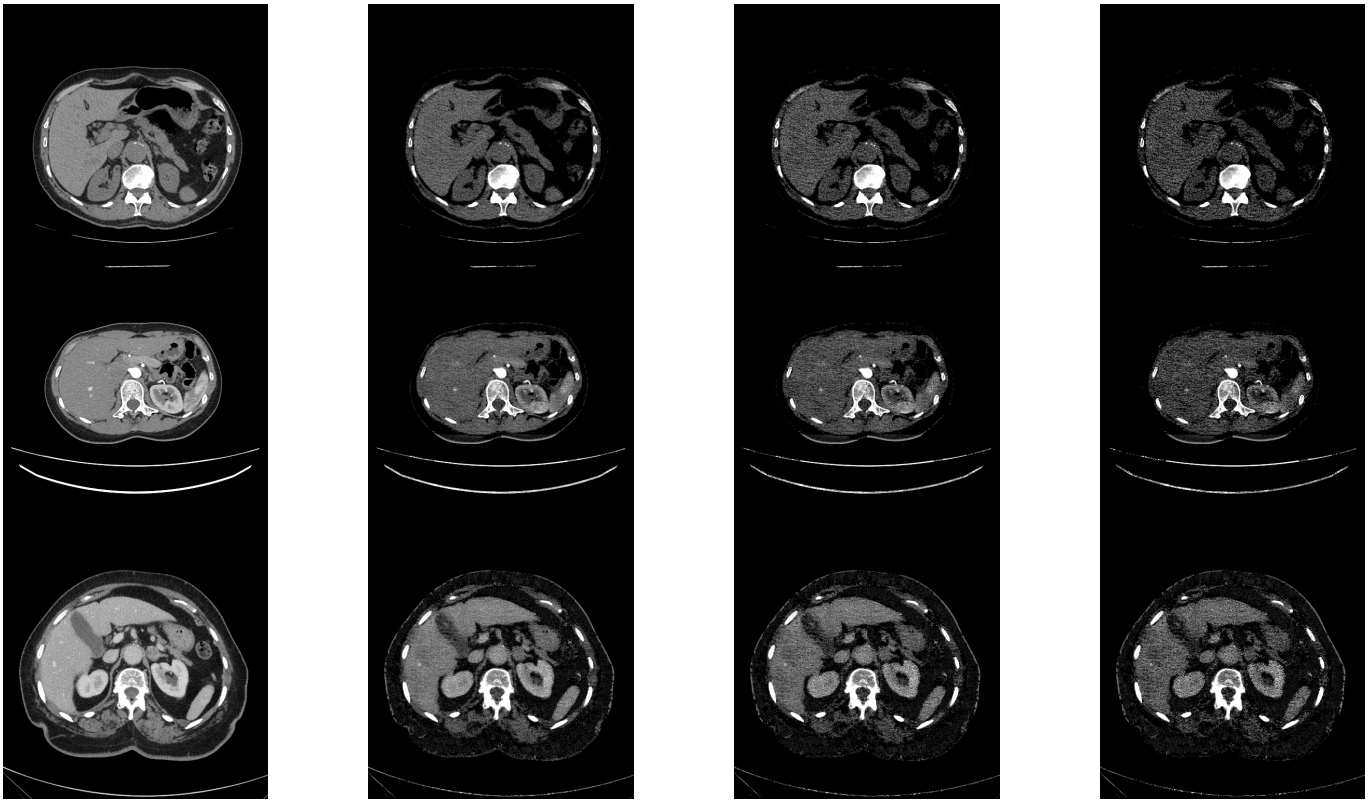
[7] C. Won Kim and J. H. Kim, "Realistic simulation of reduced-dose ct with noise modeling and sinogram synthesis using dicom ct images," *Medical physics*, vol. 41, no. 1, p. 011901, 2014.

[8] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose ct denoising with convolutional neural network," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 143–146.

[9] Z. Huang, Z. Chen, Q. Zhang, G. Quan, M. Ji, C. Zhang, Y. Yang, X. Liu, D. Liang, H. Zheng, and Z. Hu, "Cagan: A cycle-consistent generative adversarial network with attention for low-dose ct imaging," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1203–1218, 2020.

[10] X. Yi and P. Babyn, "Sharpness-aware low-dose ct denoising using conditional generative adversarial network," *Journal of digital imaging*, vol. 31, pp. 655–669, 2018.

[11] C. You, Q. Yang, H. Shan, L. Gjesteby, G. Li, S. Ju, Z. Zhang, Z. Zhao, Y. Zhang, W. Cong *et al.*, "Structurally-sensitive multi-scale deep neural network for low-dose ct denoising," *IEEE access*, vol. 6, pp. 41 839–41 855, 2018.



(a) NDCT (b) UnetLD1 (c) UnetLD2 (d) UnetLD3

Fig. 8: Simulated LDCTs on external dataset (displayed in window [-150,250]).

[12] I. Shiri, A. Akhavanallaf, A. Sanaat, Y. Salimi, D. Askari, Z. Mansouri, S. P. Shayesteh, M. Hasanian, K. Rezaei-Kalantari, A. Salahshour *et al.*, "Ultra-low-dose chest ct imaging of covid-19 patients using a deep residual neural network," *European radiology*, vol. 31, pp. 1420–1431, 2021.

[13] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] M. Gholizadeh-Ansari, J. Alirezaie, and P. Babyn, "Low-dose ct denoising with dilated residual network," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5117–5120.

[16] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu, "Medical image denoising using convolutional neural network: a residual learning approach," *The Journal of Supercomputing*, vol. 75, no. 2, pp. 704–718, 2019.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[18] Y. Han and J. C. Ye, "Framing u-net via deep convolutional framelets: Application to sparse-view ct," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1418–1429, 2018.

[19] E. Gong, J. M. Pauly, M. Wintermark, and G. Zaharchuk, "Deep learning enables reduced gadolinium dose for contrast-enhanced brain mri," *Journal of magnetic resonance imaging*, vol. 48, no. 2, pp. 330–340, 2018.

[20] C. Niu, G. Wang, P. Yan, J. Hahn, Y. Lai, X. Jia, A. Krishna, K. Mueller, A. Badal, K. Myers *et al.*, "Noise entangled gan for low-dose ct simulation," *arXiv preprint arXiv:2102.09615*, 2021.

[21] Q. Yang, M. K. Kalra, A. Padole, J. Li, E. Hilliard, R. Lai, and G. Wang, "Big data from ct scanning," *JSM Biomed. Imag*, vol. 2, no. 1, pp. 1003–1, 2015.

[22] Z. Wang, L. Wang, S. Duan, and Y. Li, "An image denoising method based on deep residual gan," in *Journal of Physics: Conference Series*, vol. 1550, no. 3. IOP Publishing, 2020, p. 032127.

[23] S. Pasumarthi, J. I. Tamir, S. Christensen, G. Zaharchuk, T. Zhang, and E. Gong, "A generic deep learning model for reduced gadolinium dose in contrast-enhanced brain mri," *Magnetic Resonance in Medicine*, vol. 86, no. 3, pp. 1687–1700, 2021.

[24] D.-Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," *Journal of digital imaging*, vol. 21, no. 3, pp. 338–347, 2008.

[25] L. Fan, F. Zhang, H. Fan, and C. Zhang, "Brief review of image denoising techniques," *Visual Computing for Industry, Biomedicine, and Art*, vol. 2, no. 1, pp. 1–12, 2019.

[26] N. I. Radwan, N. M. Salem, and M. I. El Adawy, "Histogram correlation for video scene change detection," in *Advances in computer science, engineering & applications*. Springer, 2012, pp. 765–773.

[27] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[28] D. B. Russakoff, C. Tomasi, T. Rohlfing, and C. R. Maurer, "Image similarity using mutual information of regions," in *European conference on computer vision*. Springer, 2004, pp. 596–607.

[29] J. F. Barrett and N. Keat, "Artifacts in ct: recognition and avoidance," *Radiographics*, vol. 24, no. 6, pp. 1679–1691, 2004.

Medical Point Clouds Enhancement at the Network Edge

Paolo Giannitrapani, Tiziana Cattai, Stefania Colonnese
Dept. of Information Engineering, Electronics and Telecommunications
Sapienza University of Rome
Rome, Italy
(paolo.giannitrapani, tiziana.cattai, stefania.colonnese)@uniroma1.it

***Abstract*—The paper proposes a method for the enhancement of medical point clouds suitable for implementation at the edge of a next generation network. This method exploits the 2D point clouds projection employed in compression algorithms and enhances the point cloud by applying a diffusion sampling model in the flattened domain. The proposed approach, to which we refer to as Projection Sampling based Point Cloud Enhancement, perfectly fits the e-health service architecture in next generation network since it can be implemented at the network edge or within the network, at an intermediate transcoding stage. The experimental findings with medical point clouds demonstrate the method's efficacy in mitigating noise and preserving texture information, making it a valuable tool for incorporating point cloud enhancement into an Extended Reality transmission system.**

***Index Terms*—Point cloud, Extended reality, Image quality assessment, Diffusion models**

I. Introduction

The paper presents the Projection Sampling based Point Cloud Enhancement (PS-PCE) method for the enhancement of medical Point Clouds (PCs) suitable for implementation at the edge of a next generation network. [1]

Extended Reality (XR) e-health services are expected to play a significant role in next generation networks, and PCs are excellent candidates for volumetric data representation. Due to PC acquisition or generation errors, the vertex locations may be estimated in presence of an additive error, or a few points cloud vertices may lack of attribute information, such as color. Thereby, enhancement methods such as geometry or texture PC denoising as well as PC inpainting could improve the PC quality and definitely the XR service feasibility. PCs are defined into a 3D, non-Euclidean domain and several enhancement methods in the literature, while focusing on novel processing tools, overlook the feasibility of performing enhancement within an XR e-health service.

The PS-PCE method stems on standard architectures for medical XR data encoding and transmission over next generation networks. In particular, PS-PCE acts on multiple flattened representations of the PC, which is the same adopted in MPEG V-PCC coding standard [1]. The enhancement applies a well established diffusion network model to the flattened representations, and then recomputes the corresponding 3D data. The PS-PCE method approach perfectly fits the e-health service architecture in next generation network and it has a few advantages: i) it can be implemented at the network edge, or at user's premises, thereby exploiting edge computing capabilities; ii) it can be realized within the network, at an intermediate transcoding stage. Experimental results of PS-PCE method on medical PC datasets prove that the method fast and effectively counteracts noise and texture information loss, providing an effective tool to encompass PC enhancement in a XR transmission system.

[1]This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART".

II. Background review

PCs are sets of data characterized by locations and color/texture attributes, acquired by purely passive systems with an RGB camera, or both active and passive systems, including a RGB plus depth camera or a Light Detection and Ranging (LIDAR).

Visual quality of PCs is a matter of strong interest in different scientific communities, from signal processing [2], [3] to deep learning [4], [5]. Several methods that estimate the visual quality of the PCs stem from metrics specifically developed for images or meshes [6], [7]. In general, tools belonging to $2D$ processing have been adapted to elaborate PCs for different purposes, such as the video-based PC compression [1] that is based on projections of the PC on the $2D$ space in order to use state-of-the-art methods for video coding [8]. Another class of problems is the PC completion processing, where several methods apply state-of-the-art methods for 2D completion to 3D PCs by means voxelization and 3D convolution [9]–[11].

PCs are expected to enable XR services such as telemedicine or remote surgery in 5G and beyond networks. Still, PC compression is challenging due to the non-uniform distribution of points in space and the typically large number of points even within a single object. The ISO Moving Picture Expert Group (MPEG) is addressing standardization of both LIDAR and surface PC compression and of dynamic multimedia PCs; for these latter, the Video PC Coding (V-PCC) standard has been designed. The V-PCC algorithn for dynamic content leverages an intermediate PC representation, by suitable projections of the PC in 2D space.

We leverage the PC encoding architecture to perform the PC enhancement in the flattened representation computed by

979-8-3503-4218-5/23/31.00 © 2023$IEEE$

the V-PCC code: the enhancement applies without training in the flattened domain, leveraging either edge computing or network-assisted processing in a next generation network.

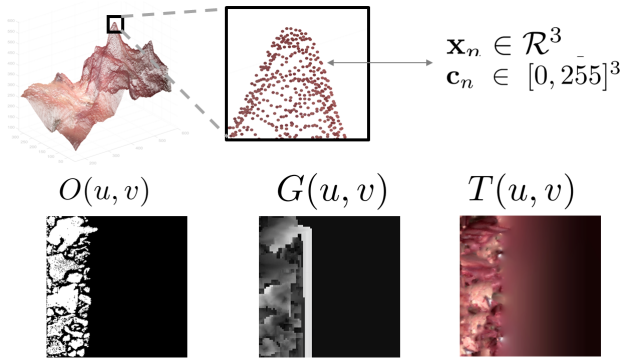


Fig. 1. From 3D PC vertex locations $\mathbf{x}_n \in \mathcal{R}^3$, $n = 0,...N-1$, and texture attributes $\mathbf{c}_n \in [0,255]^3$ to 2D maps $O(u,v)$, $G(u,v)$, $T(u,v)$, $u,v = 0,\cdots M-1$.

III. Projection Sampling Point Cloud Enhancement

A PC is a collection of N points (vertices) in the 3D space, where each point is associated with position and attribute information (e.g., luminance and color). Let $\mathbf{x}_n \in \mathcal{R}^3$ denote the PC vertex locations $n = 0,...N-1$, and $\mathbf{c}_n \in [0,255]^3$ the associated texture attributes. For medical PCs, such as those extracted from laparoscopic measurements [12], the texture attributes typically consists in photometric information. Due to acquisition limits, the PC may present noisy or missing attributes at some vertices. The purpose of the proposed enhancement is to recover these kinds of errors, in the same 2D domain where compression is carried out.

A. Projection generation

The V-PCC encoding represents the N vertices PC by three $M \times M$ bidimensional maps, representing the PC Occupancy, Geometry and Texture, respectively. Let us denote the flattening operators as Ψ_O, Ψ_G, and Ψ_T, respectively. In a nutshell, each operator aggregates different lateral 2D projections of the PCs into a 2D map, avoiding self-occlusion or hidden surfaces [1], and possibly applying an edge smoothing operator; finally, the operators yield the following bi-dimensional data:

$$\begin{cases} O(u,v) = \Psi_O\left(\mathbf{x}_0,\cdots \mathbf{x}_{N-1}\right) \\ G(u,v) = \Psi_G\left(\mathbf{x}_0,\cdots \mathbf{x}_{N-1}\right) \\ T(u,v) = \Psi_T\left(\mathbf{x}_0,\cdots \mathbf{x}_{N-1}; \mathbf{c}_0,\cdots \mathbf{c}_{N-1}\right) \end{cases} \quad (1)$$

for $u,v = 0,\cdots M-1$. By this representation, the n-th vertex of the PC is associated to a pixel $(u_n, v_n) = \Phi(\mathbf{x}_n)$. Let us remark that the operators Ψ_O, Ψ_G, and Ψ_T are not perfectly reversible, since they rely on a quantization of the spatial coordinates of the PC vertices to map them on the bidimensional (u,v) grid.

An example of the maps $O(u,v)$, $G(u,v)$, $T(u,v)$, for $u,v = 0,\cdots M-1$ is shown in Fig. 1.

The occupancy map $O(u,v)$ is a binary indicator map of pixels corresponding to PC points; the geometry map $G(u,v)$ provides the depth information of pixels in space before projection (e.g., the distance between the pixel location in 3D space and the projection plane); the attribute map $T(u,v)$ preserves the texture (luminance and chrominances) information of the initial PC.

B. Projection sampling

Herein, we address the enhancement of the maps $O(u,v)$, $G(u,v)$, $T(u,v)$ through diffusion models. According to diffusion models, the observed degraded image is obtained from the original one by a Markovian, step-wise degradation. By retracing the chain of states in reverse, the model learns to reconstruct the original image starting from the degraded one.

The literature presents several supervised and unsupervised diffusion models tackling inverse problems, both [13]–[16]. The models are slow in training due to the high dimensionality of the latent variable, which allows to generalize well.

The applicability of classical diffusion models to the enhancement of $O(u,v)$, $G(u,v)$, $T(u,v)$ is somewhat limited. The maps $O(u,v)$, $G(u,v)$, $T(u,v)$, despite being bidimensional, do not fit statistical properties of natural images [17], such as Markovianity (smoothness). The abrupt edges between different 2D projections of the original PC data lead to highly varying, fragmented bidimensional sequences.

We resort here to the class of Denoising Diffusion Restoration Models (DDRM) in [18], which immediately generalize without requiring specific training. This feature perfectly suits the enhancement of the $O(u,v)$, $G(u,v)$, $T(u,v)$ maps. DDRM iteratively solves a recovery problem exploiting any pre-trained diffusion model at each iteration. DDRM assumes that the observed image $\mathbf{y}$ is linearly related to the unknown original image $\mathbf{x}$, namely $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, being $\mathbf{H}$ the known linear degradation matrix and $\mathbf{w}$ an additive Gaussian noise of known variance σ_n^2. The model iteratively solves a generalized linear inverse problem: the idea is to tighten the noise present in the measurement with the noise of the diffusion process $\mathbf{x}_{1:T}$, ensuring that the state $\mathbf{x}_0$ is faithful to the measurement $\mathbf{y}$ (state without any added noise). The diffusion process leverages the Singular Value Decomposition (SVD) of $\mathbf{H}$ and it acts as an iterative orthogonalization of the signal and of the noise. At the k-th iteration, at state i, the distribution of the unknown image $\mathbf{x}_i$ is computed as $p^{(k)}_{\mathbf{X}_i}(\mathbf{x}_i) = \mathcal{N}\left(\hat{\mathbf{x}_i}^{(k)}, \mathbf{\Sigma}^{(k)}\right)$, where $\hat{\mathbf{x}_i}^{(k)}$ is a weighted combination of the measurements $\mathbf{y}$ and of the nonlinear estimate $\tilde{\mathbf{x}_i}^{(k)} = \eta\left(\hat{\mathbf{x}_i}^{(k+1)}\right)$, computed using any pre-trained diffusion model such as those in [13], [14]. We exploit the DDRM generalization property to perform projection sampling starting from the observed maps $O(u,v)$, $G(u,v)$, $T(u,v)$; the recovered maps are then back projected in the 3D space leading to inpainting as well as geometry or texture denoising on the observed PC.

Projection sampling requires knowledge of the degradation matrix $\mathbf{H}$, as well as of an estimate of the variance of the additive texture or geometry noise.

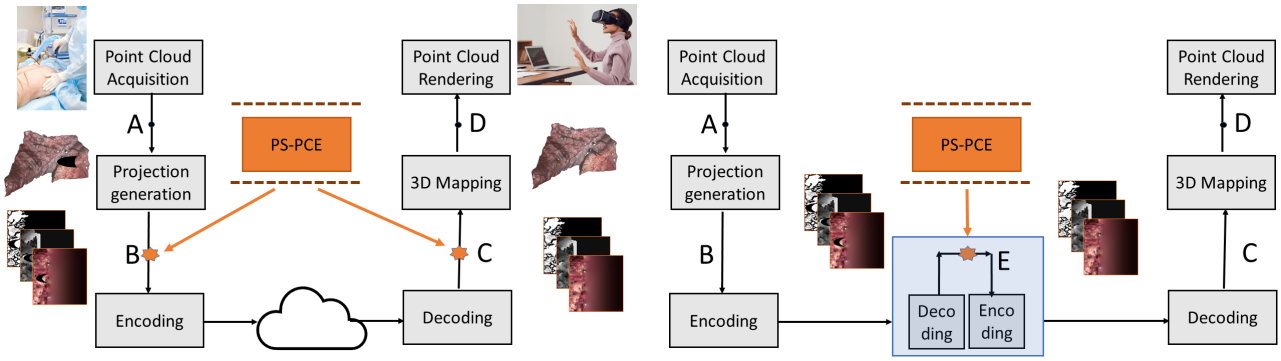


Fig. 2. PS-PCE implementation within a Next generation networks XR service: edge computing (left) and network assisted rendering (right).

In inpainting, a subset $\mathcal{I}$ of the PC vertices lacks of the attribute information, i.e., $\mathbf{c}_n$ is missing for $n \in \mathcal{I}$, due to a PC acquisition or generation error. For any $n \in \mathcal{I}$, we label the mirror point (u_n, v_n) as missing in all the three maps. In denoising, we set the additive error variance to a known value.

IV. PS-PCE at Network Edge

Next generation networks encompass delivery of XR [19] and Immersive Video (IV) data [20]. Medical XR data, either directly acquired by LIDAR or Time-of-Flight (TOF) cameras, or reconstructed by Red Green Blue Depth (RGBD) single or multi camera rigs, are taking the lion's share in e-health services, yet they require effective encoding algorithms and edge-assisted rendering services for clinical [21] or educational purposes [22], [23]. Artificial Intelligence based network resources management is expected not only to provide high transmission rate at low latency, but also network-assisted processing and edge computing facilities [24].

The PS-PCE approach is natively designed to work on intermediate variables of the encoding process, and it can naturally fit the XR service architecture. This is illustrated in Fig. 2 (left side). The PS-PCE can be realized both at the source and destination side, on the maps produced for the purpose of PC encoding. The PS-PCE can be realized in edge computing, i.e., exploiting computation resources at the network edge and at the user premises. Besides, we observe that the computation architecture may encompass network assisted processing, as illustrated in Fig. 2 (right side). In this scenario, the encoded PC data is only partially decoded at an in-network element (transcoder); then, PS-PCE is applied, and the enhanced data are re-encoded and transmitted to the final destination. Different computation architectures face different problems. Application of PS-PCE at point A counteracts acquisition errors and missing attributes, whereas PS-PCE at point B can effectively recover distortions due to lossy compression or provide transmission error concealment. Finally, the in-network processing provided by applying PS-PCE at point E can ease the computational effort at the destination side, and this can extend the feasibility of the approach to mobile end user devices.

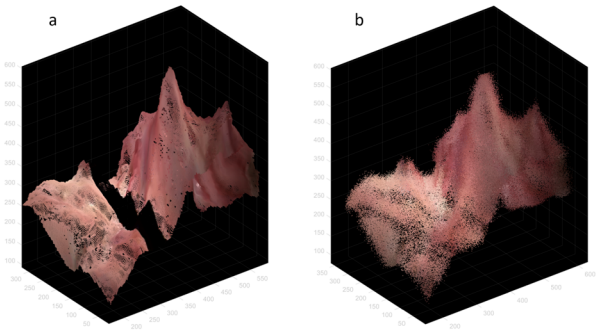


Fig. 3. Endoscopic PC: a) case of vertices with missing attributes and b) case of additive Gaussian noise.

V. Numerical results

To prove the potential of PS-PCE, we consider two different kinds of medical data, belonging to endoscopic PC [12] and aneurysm PC [25].

We firstly analyze the endoscopic dataset [12]. The PC has $N = 123200$ points, and RGB attributes at each vertex. The vertex locations are scaled and shifted so as to fit the interval $[0, 640]$ per axis (the axis values are empirically chosen to contain all frame projections in a single 1280×1280 image).

In the case of inpainting, a set $\mathcal{I}$ of vertex attributes of initial PC has been removed. After application of the Ψ_o these points correspond to non zero values in the occupancy and geometry maps, whereas they are represented in black the attribute maps.

Fig. 3 shows the endoscopic PC (a) in the case of inpainting, i.e., vertices with missing attributes, and (b) in presence of additive Gaussian noise. Fig. 4 shows the different maps produced by the V-PCC in presence of inpainting, namely the occupancy $O(u, v)$, geometry $G(u, v)$, and attribute $T(u, v)$ maps corresponding to the PC in Fig. 3 a). The vertices without attributes are clearly visible as non zero areas in the occupancy and geometry maps, and as black areas in the attribute map. In Fig. 4, mask is directly identified on the projections in the attribute map. In Fig. 5 we present the $O(u, v)$, $G(u, v)$, and $T(u, v)$ maps in the case of PC affected in additive Gaussian noise for the PC in Fig. 3.

The map selected for processing is tiled into blocks of size

Fig. 4. Endoscopic PC inpainting: occupancy $O(u, v)$, geometry $G(u, v)$, and attribute $T(u, v)$ maps with missing pixels related to the PC vertices in $\mathcal{I}$.

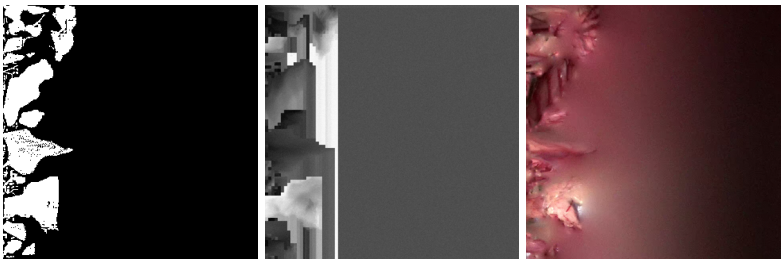


Fig. 5. Endoscopic PC denoising: occupancy $O(u, v)$, geometry $G(u, v)$, and attribute $T(u, v)$ maps with additive Gaussian noise with $\sigma_n = 0.05$.

Fig. 6. Endoscopic PC inpainting: occupancy $O(u, v)$, geometry $G(u, v)$, and attribute $T(u, v)$ maps after inpainting.

Fig. 7. Endoscopic PC denoising: occupancy $O(u, v)$, geometry $G(u, v)$, and attribute $T(u, v)$ maps after denoising.

256×256 to adapt it to the DDRM library [18] and composed back after the processing.

The DDRM model assumes that the matrix $\mathbf{H}$ is known. In the inpainting case, the matrix $\mathbf{H}$ is represented by the mask provided along with the attribute map and identifies the areas of missing pixels to be reconstructed. We set $\mathbf{H} = \text{diag}(b_0, \cdots b_{N-1})$. In the denoising case, the degradation consists in the additive Gaussian noise that can affect both the PC vertices coordinates and the attribute at each vertex. This translates into an additive noise on both the attribute map and the geometry map. The DDRM assumes the acquisition noise variance to be known, so as to properly weight the measurements and the current estimate throughout the iterative recovery algorithm. At each iteration, DDRM exploits a pretrained model as nonlinear image estimator; in the simulations, we resort to the pretrained ImageNet diffusion model as nonlinear estimator [15] because it demonstrates better performance in solving inverse problems on out-of-distribution images with general content.

Figs. 6 and 7 display the attribute maps after processing with the DDRM model in the inference phase, whereas Figs. 8 a) and b) display the final PCs after reconstruction.

We take into consideration another PC in order to apply our method to other possible scenarios [25]. We represent

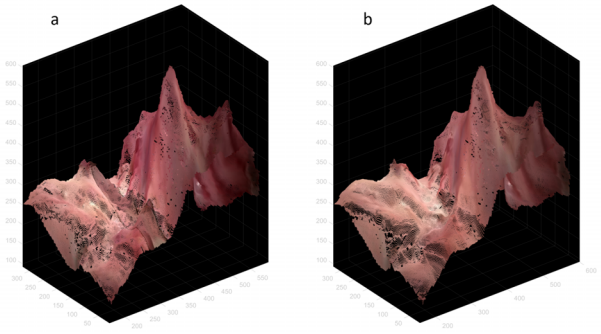


Fig. 8. PS-PCE results on endoscopic PC: a) inpainted and b) denoised PC.

in Fig. 9 a PC modeling an aneurysm, where we added an additive Gaussian noise of known variance. This PC contains only geometry information, without attribute. We represent the occupancy $O(u, v)$ and the geometry $G(u, v)$ in this configuration in Fig. 10. Then we apply all the steps of our proposed denoising method, and we can see results in terms of occupancy and geometry maps in Fig. 11 and in terms of denoised PC in Fig. 12. We can remark that after the denoising the geometry of the aneurysm PC is more defined and more details are visible.

Tabs. 1, 2 and 3 show the Image Quality Assessment (IQA) metrics, computed given the reference image. These Full Reference objective quality metrics estimate either the loss of visual information [17], [26] or the similarity between the reference image and the corrupted/restored image [27]–[29] by modeling the Human Visual System.

Tabs. 1 and 2 present the IQA metrics for the endoscopic test, for the inpainting and denoising ($\sigma_n = 0.05$) cases respectively, before and after DDRM processing. All the metrics improvement, apart for VIF in the attribute map case. VIF is sensitive to the artifacts that do not fit natural images models [30]. Future work will extend the diffusion model to account for joint processing of the geometric, occupancy and texture information.

In Tab. 3, the values for the aneurysm test are shown, specifically for the denoising case (with $\sigma_n = 0.05$ and $\sigma_n = 0.1$). The structure of the aneurysm PC after denoising exhibits a high number of details compared to the noisy case. As can be seen from the quality indices, the DDRM model is able to recover the final PC close to the original, pristine PC, even in the case of higher noise with $\sigma_n = 0.1$.

It is important to highlight that even a moderate amount of noise added to the maps generates high levels of noise in the final PCs.

VI. Conclusion and Future Work

This paper has presented the Projection Sampling - Point Cloud Enhancement method suitable for implementation at the edge of a next generation network. The method exploits the 2D PCs projection employed in compression algorithms and enhances the PC by applying a diffusion sampling model in the

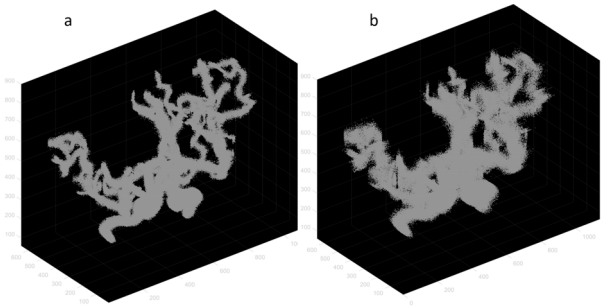


Fig. 9. Aneurysm PCs in presence of additive Gaussian noise.

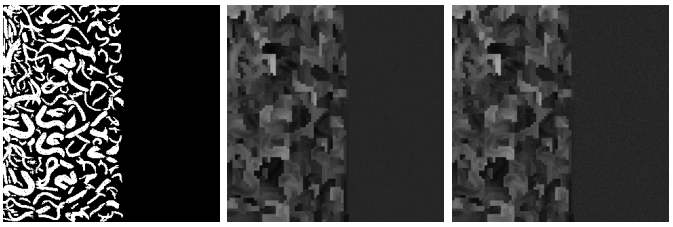


Fig. 10. Aneurysm PC denoising: occupancy $O(u, v)$, geometry $G(u, v)$ maps with additive Gaussian noise with $\sigma_n = 0.05$ (center) and $\sigma_n = 0.1$ (right).

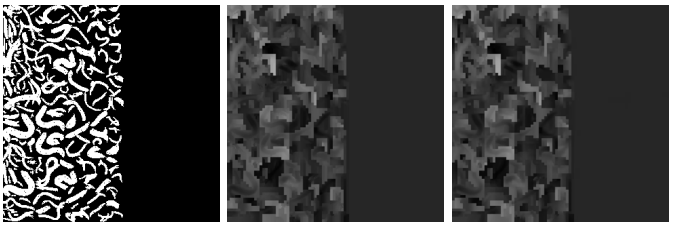


Fig. 11. Aneurysm PC denoising: occupancy $O(u, v)$, geometry $G(u, v)$ after denoising.

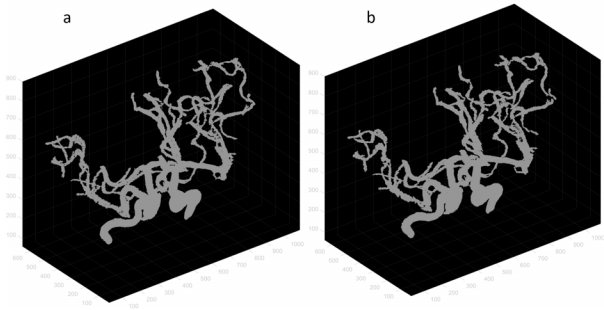


Fig. 12. Effect of PS-PCE denoising on Aneurysm PCs of Fig. 9.

flattened domain. We present different architectures to encompass PS-PCE in e-health service architecture in next generation network. PS-PCE can be implemented at the network edge or within the network, at an intermediate transcoding stage.

Experimental results on medical PCs show that PS-PCE effectively counteracts noise and texture information loss on the PC by applying a diffusion model on the PC projection. This paves the way to extend several results achieved by

TABLE I
INPAINTING EXPERIMENT (ENDOSCOPIC TEST).

IQA metrics	before DDRM attribute map	after DDRM attribute map
MS-SSIM	0.9670	0.9715
FSIM	0.9552	0.9635
GMSD	0.0614	0.0482
VIF	0.7369	0.7192
MAD	124.27	128.80
PSNR	21.98	25.43

TABLE II
DENOISING EXPERIMENT (ENDOSCOPIC TEST).

IQA metrics	before DDRM attribute map $\sigma_n = 0.05$	after DDRM attribute map $\sigma_n = 0.05$	before DDRM geometry map $\sigma_n = 0.05$	after DDRM geometry map $\sigma_n = 0.05$
MS-SSIM	0.9175	0.9920	0.8720	0.9980
FSIM	0.9420	0.9773	0.8400	0.9988
GMSD	0.0736	0.0208	0.1041	0.0117
VIF	0.4184	0.3241	0.5501	0.6288
MAD	97.10	44.57	98.75	12.37
PSNR	31.89	42.35	28.91	49.10

TABLE III
DENOISING EXPERIMENT (ANEURYSM TEST).

IQA metrics	before DDRM geometry map $\sigma_n = 0.05$	after DDRM geometry map $\sigma_n = 0.05$	before DDRM geometry map $\sigma_n = 0.1$	after DDRM geometry map $\sigma_n = 0.1$
MS-SSIM	0.8860	0.9927	0.7234	0.9884
FSIM	0.9890	0.9940	0.9629	0.9877
GMSD	0.0999	0.0196	0.1982	0.0307
VIF	0.5065	0.6084	0.3147	0.5400
MAD	86.71	23.67	116.77	35.54
PSNR	28.88	44.14	23.03	42.36

diffusion network to PCs. Future work will address the joint modeling of the different kinds of bidimensional information associated to the PC (shape, geometry, texture) into an unified diffusion model.

REFERENCES

[1] D. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and A. Tabatabai, "An overview of ongoing point cloud compression standardization activities: Video-based (v-pcc) and geometry-based (g-pcc)," *APSIPA Trans. on Signal and Information Processing*, vol. 9, p. e13, 2020.

[2] I. Viola and P. Cesar, "A reduced reference metric for visual quality evaluation of point cloud contents," *IEEE Signal Processing Letters*, vol. 27, pp. 1660–1664, 2020.

[3] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards subjective quality assessment of point cloud imaging in augmented reality," in *2017 IEEE 19th Int. Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, 2017.

[4] A. Chetouani, M. Quach, G. Valenzise, and F. Dufaux, "Deep learning-based quality assessment of 3d point clouds without reference," in *2021 IEEE Int. Conf. on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2021.

[5] L. Zhou, G. Sun, Y. Li, W. Li, and Z. Su, "Point cloud denoising review: from classical to deep learning-based approaches," *Graphical Models*, vol. 121, p. 101140, 2022.

[6] G. Lavoué and R. Mantiuk, "Quality assessment in computer graphics," in *Visual Signal Quality Assessment: Quality of Experience (QoE)*, pp. 243–286, Springer, 2014.

[7] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Trans. on Multimedia*, vol. 23, pp. 3877–3891, 2020.

[8] S. Schwarz and M. Pesonen, "Real-time decoding and ar playback of the emerging mpeg video-based point cloud compression standard," *Nokia Technologies; IBC: Helsinki, Finland*, 2019.

[9] A. Dai, C. Ruizhongtai Qi, and M. Nießner, "Shape completion using 3d-encoder-predictor cnns and shape synthesis," in *Proc. of the IEEE Conf. on computer vision and pattern recognition*, pp. 5868–5877, 2017.

[10] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu, "High-resolution shape completion using deep neural networks for global structure and local geometry inference," in *Proc. of the IEEE Int. Conf. on computer vision*, pp. 85–93, 2017.

[11] Y. Liu, B. Fan, S. Xiang, and C. Pan, "Relation-shape convolutional neural network for point cloud analysis," in *Proc. of the IEEE/CVF Conf. on computer vision and pattern recognition*, pp. 8895–8904, 2019.

[12] L. Xi, Y. Zhao, L. Chen, Q. H. Gao, W. Tang, T. R. Wan, and T. Xue, "Recovering dense 3d point clouds from single endoscopic image," *Computer Methods and Programs in Biomedicine*, vol. 205, p. 106077, 2021.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. on Learning Representations*, 2021.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," *arXiv e-prints*, p. arXiv:2011.13456, Nov. 2020.

[17] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, pp. 430–444, Feb. 2006.

[18] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[19] 3GPP, *Extended Reality (XR) in 5G, 3GPP TR 26.928*, 12 2020.

[20] 3GPP, *Immersive Teleconferencing and Telepresence for Remote Terminals (ITT4RT) Operation and Usage Guidelines, 3GPP TR 26.962*, 11 2021.

[21] M. Sugimoto, "Cloud xr (extended reality: Virtual reality, augmented reality, mixed reality) and 5g mobile communication system for medical image-guided holographic surgery and telemedicine," *Multidisciplinary Computational Anatomy: Toward Integration of Artificial Intelligence with MCA-based Medicine*, pp. 381–387, 2022.

[22] T. Bieg, R. Schatz, S. Egger-Lampl, B. Roszipal, and K. Kinzer, "Better experience, better performance? results of a study on vr training effectiveness in healthcare," in *2022 14th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, pp. 1–4, IEEE, 2022.

[23] P. S. Mathew and A. S. Pillai, "Role of immersive (xr) technologies in improving healthcare competencies: a review," *Virtual and Augmented Reality in Education, Art, and Museums*, pp. 23–46, 2020.

[24] 3GPP, *Study on 5G Media Streaming Extensions for Edge Processing, 3GPP TR 26.803*, 8 2020.

[25] X. Yang, D. Xia, T. Kin, and T. Igarashi, "Intra: 3d intracranial aneurysm dataset for deep learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 2656–2666, 2020.

[26] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," tech. rep., Video Quality Experts Group, http://www.vqeg.org, Mar. 2000.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[28] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Trans. on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[29] W. Xue, L. Zhang, X. Mou, and A. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. on Image Processing*, vol. 23, pp. 684–695, Feb. 2014.

[30] E. D. Di Claudio, P. Giannitrapani, and G. Jacovitti, "Predicting blur visual discomfort for natural scenes by the loss of positional information," *Vision Research*, vol. 189, pp. 33–45, 2021.

ENHANCED RESIDUE PREDICTION FOR LOSSLESS CODING OF MULTIMODAL IMAGE PAIRS BASED ON IMAGE-TO-IMAGE TRANSLATION

Daniel S. Nicolau[⋆‡], *Joao O. Parracho*[⋆†], *Lucas A. Thomaz*[⋆‡],
Luis M. N. Tavora[⋆‡], *and Sergio M. M. Faria*[⋆‡]

[‡]ESTG, Polytechnic of Leiria, Leiria, Portugal
[⋆] Instituto de Telecomunicações, Leiria, Portugal
[†]PEE, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

ABSTRACT

Multimodal medical imaging combine data obtained from multiple techniques simultaneously, yielding more detailed information about the content, which is a clear advantage over independent acquisition techniques. As these images are acquired using different imaging modalities and sometimes even in different dimensions, they commonly require a geometrical registration process. However, when they are encoded using standard image codecs the prediction methods do not exploit the redundancies related to the multimodal acquisition. In this paper, a novel lossless multimodal prediction module is introduced. The proposed method employs a deep learning-based approach with Image-to-Image translation for the purpose of joint coding of Positron Emission Tomography (PET) and Computed Tomography (CT) image pairs. Prior to the coding stage, a Generative Adversarial Network (GAN) is used for multimodal image translation. Then, a weighted estimated image is utilised as the I-frame, while the weighted sum of the original and synthesised image from the same modality serves as the P-frame for inter prediction. By employing weighted frames, the predictive frame approximates the reference frame more accurately, enhancing the overall performance of the prediction process. The experimental results, on a publicly available PET-CT dataset, demonstrate that the proposed prediction scheme outperformed the previously proposed method, and attains coding gains up to 13.20% when compared with the single modality intra coding of the Versatile Video Coding (VVC) lossless standard.

Index Terms— Lossless image coding, Multimodal image coding, Learning based prediction, Generative predictive coding, Versatile Video Coding

Authors' e-mails: {daniel.nicolau, jparracho, lucas.thomaz, luis.tavora, sergio.faria}@co.it.pt.This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the projects UIDB/50008/2020, LA/P/0109/2020 under project CIBME, and 2022.09914.PTDC under project CoMBINNe.

1. INTRODUCTION

Medical imaging has emerged as a critical resource to assist healthcare professionals [1]. It is used in diverse applications such as for example diagnosis, monitoring, and treatment planning. The continuous development of medical imaging technologies [2] have pushed the boundaries of image resolution and bit-depth across multiple modalities. As a consequence of the increased image quality, the requirements for storage and transmission of such images have changed, creating new challenges for hospitals, clinics, and research institutions. The use of multimodal images [3], which represent objects captured through different technologies simultaneously, helps to overcome the limitations of independent acquisition techniques, facilitating the extraction of complementary information such as structural and functional scans.

Deep learning has gained significant attention in the research domain, offering solutions to a wide range of problems. This has led to the development of various algorithms, including Deep Convolutional Neural networks (CNN), Autoencoders (AE), GANs, and Vision Transformers. These techniques have been increasingly applied to leverage the potential of multimodal medical images, particularly in tasks such as medical image segmentation, synthesis, and coding. The use of multimodal information can lead to the development of more sophisticated and efficient methods. The current state-of-the-art work using multimodal medical images has been predominately focused on tasks such as segmentation and synthesis. Regarding the segmentation task, in [4], a novel whole-body segmentation framework for accurately identifying heterogeneous Metastatic Melanoma (MM) lesions in 3D Fluorodeoxyglucose (FDG)-PET / CT images is described. *Ziqi Yu et al.* [5] proposed MouseGAN++, a novel framework for segmenting mouse brain fine structures with limited multimodal Magnetic Resonance Imaging (MRI) data. It employs a disentangled and contrastive GAN-based approach to synthesise missing modalities and enhance segmentation performance. Experimental results demonstrate superior performance compared to state-of-the-art methods,

showcasing its effectiveness in fusing cross-modality information and achieving robust segmentation outcomes.

A generative adversarial approach for medical image synthesis (ResViT) is described in [6]. It combines vision transformers, convolution operators, and adversarial learning to achieve high-quality results. The generator utilises aggregated residual transformer (ART) blocks, promoting representation diversity and task-relevant information. Experimental evaluations demonstrate its superiority in synthesising missing sequences in multi-contrast MRI and CT images from MRI, when compared to competing methods.

The continuous demand for the improvement of image quality led to the development of more sophisticated algorithms in medical imaging. However, this progress has also resulted in a significant increase in data volume especially when multimodal imaging is employed, requiring efficient coding methods for storage and transmission. To address these concerns, lossless coding techniques are commonly employed in medical compression applications to ensure the information is not modified. The emphasis on preserving the details plays an important role in maintaining diagnostic integrity and achieving high accurate analysis in medical imaging. Regarding lossless multimodal medical image coding, in the literature, few works address this problem using a deep learning approach. In [7, 8] a pipeline based on an Attention-Guided Generative Adversarial Network (AGGAN) that generates an estimated PET to be jointly coded with the CT was proposed. For the sake of simplification, the AGGAN will be referred to as GAN throughout the remainder of the paper. In [7], the authors suggest to losslessly intra-code the original CT and the residue between the original and the estimated PET. A different prediction method is proposed in [8], where the original PET is encoded as a P-frame using the estimated one as a reference in the inter prediction. The resulting P-frame residue is computed using the VVC standard motion estimation module.

In this paper, a deep learning approach based on an Image-to-Image translation (I2I) [9] to jointly code PET-CT image pairs is adopted. A new prediction module integrated with the lossless VVC standard is proposed. The main contributions presented in this work can be summarised as follows: (i) a new lossless cross-modality prediction, that approximates the reference to the P-frame used for the inter-coding, (ii) ablation studies that demonstrate the efficiency of the proposed method, and (iii) experimental results on a PET-CT paired dataset that demonstrates the proposed prediction method can improve the coding performance when compared to the standard lossless VVC with intra modality prediction and other state-of-the-art learning-based predictions schemes.

The remainder of the paper is organised as follows: Section 2 describes the proposed prediction method, Section 3 presents the experimental results, analysing the compression performance and comparing it with other coding strategies. Section 4 presents the concluding remarks.

2. PROPOSED METHOD

In this work, a cross-modality encoder based on I2I is applied for PET-CT pairs within the context of lossless coding. The proposed lossless prediction scheme, shown in Figure 1, employs a module based on the VVC encoder to extract the coding residue and generate the compressed bitstream. The encoding process involves bypassing the transform (T) and quantisation (Q) functions of standard hybrid codecs, where the original CT image (CT_o) is encoded using intra coding (according with the Common Test Conditions configuration file *encoder_intra_vtm.cfg*), and PET_o as inter prediction.

Initially, CT_o is intra coded, then a GAN produces an estimated PET frame(PET_e) having CT_o as input. Then, the quality of PET_e is enhanced using the optimisation method proposed in [8]. A Nelder–Mead simplex algorithm [10] was used to minimise the mean absolute error (MAE) between PET_o and PET_e, by rescaling, adjusting the brightness and spatial alignment of the estimated one. Two different prediction methods are used to inter encode PET_o, selecting the one that results in the lower bitstream: In the first one, PET_o is encoded as a P-frame with PET_e as the reference. The VVC standard's motion estimation/compensation module is utilised to compute the P-frame residue, compensating for spatial shifts between PET_o and PET_e and minimising the energy of the residue. In the second prediction method, PET_e and PET_o are combined by means of a weighted sum of the frames. The resulting combined frame (PET_p), to be encoded as a P-frame, is expressed as:

$$PET_p = \alpha_1 \times PET_e \; + \; \alpha_2 \times PET_o, \tag{1}$$

where α_1, and α_2 are the PET_e and PET_o weights, respectively. The reference frame (I-frame) is obtained by applying a weighted sum of α_1 and α_2 to PET_e, which can be expressed as:

$$PET_i = (\alpha_1 + \alpha_2) \times PET_e \tag{2}$$

This procedure enables the approximation of the P-Frame to the I-Frame, which is expected to minimise the residual energy during the coding process.

A consequence of the summation of two images with a given bit depth naturally results in an image with a higher bit depth. α_1 and α_2 are calculated in such manner that the maximum bit depth of the resulting frames does not surpass 16, which corresponds to VVC maximum coding bit depth. Optimisation of brightness and contrast is performed on the P-Frame to minimise the MAE between the reference (I) and predictive frames (P). This optimisation process is applied to both prediction methods. The resulting prediction residue is then subjected to entropy coding and integrated into the compressed stream. The computed optimisation parameters, including scaling, contrast, brightness, and spatial alignment, are also incorporated as side information within the compressed stream.

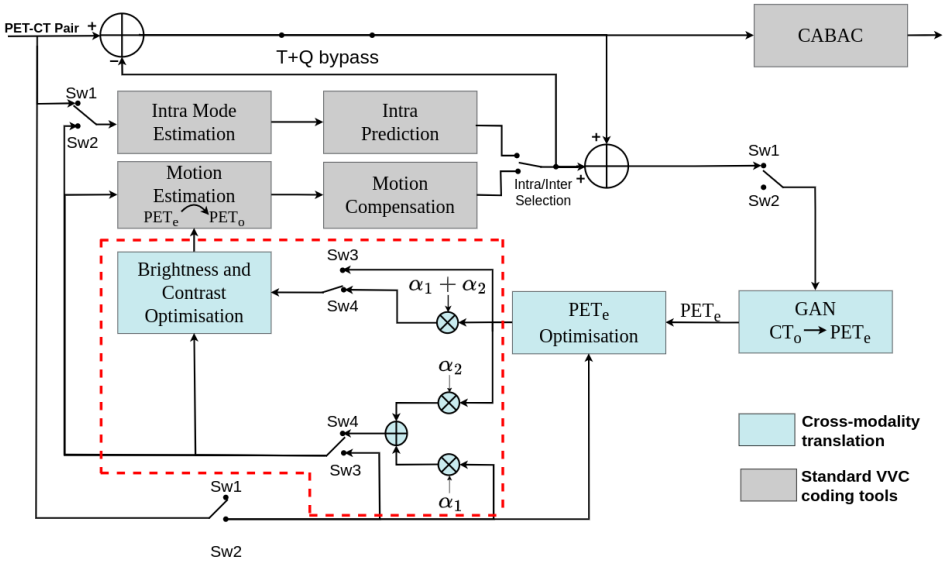


Fig. 1. Proposed lossless cross-modality prediction scheme (main contributions are shown within the dashed red box).

At the decoding phase, the initial step involves decoding CT_o, which is subsequently used to generate PET_e, with the same GAN employed in the encoder. Moreover, the bit depth information is extracted from the bitstream to identify the employed prediction method. Finally, the PET_o is decoded, ensuring the reconstruction of the original PET-CT pair without any loss of information. The selection of input images between CT_o and PET_o is facilitated by Sw1 and Sw2. Sw1 is activated when CT_o is the input, and Sw2 when it is PET_o. In addition, the proposed framework incorporates Sw3 and Sw4, enabling the selection of the most efficient prediction scheme for each PET-CT pair. Sw4 is responsible for choosing the prediction method that utilises the weighted sum of the original and estimated PET as the P-Frame.

2.1. Cross modality image translation network

In order to perform the CT to PET image translation the GAN architecture proposed in [11], with a reformulation of the generator and discriminator loss functions, is used. This architecture is comprised of two primary components: a generator (G) and a discriminator (D) network. The generator network is responsible for producing synthetic data, while the discriminator network is designed to differentiate between real and generated data. By training these two networks in an adversarial manner, the GAN aims to generate realistic and high-quality data that closely resembles the real data distribution. An essential characteristic of the GAN is its embedded attention mechanism, which enables the detection of the most distinctive semantic regions within images across different domains

For the adversarial loss, the least square approach described in [12] was used. This loss function imposes a penalty on the generated images proportional to their deviation from the assigned label (1 for real and 0 for generated data), providing informative feedback to the generator regarding the proximity of the generated images to be classified as real. This, in turn, leads to higher gradients and enhances the effectiveness of the generator. It is worth noting that the absence of such distance information, as highlighted in [13], can give rise to the well-known issue of vanishing gradients. Therefore, as discussed in [12], the adoption of the least square loss function improves training stability, and it can be expressed as:

$$\begin{aligned}\mathcal{L}_{\mathrm{LSGAN}}(D) = &\frac{1}{2}\mathbb{E}_{y\sim p_{\mathrm{data(y)}}}[(D(y)-1)^2] \\ &+ \frac{1}{2}\mathbb{E}_{x\sim p_{\mathrm{data(x)}}}[(D(G(x))^2],\end{aligned} \tag{3}$$

$$\mathcal{L}_{\mathrm{LSGAN}}(G) = \mathbb{E}_{x\sim p_{\mathrm{data(x)}}}[(D(G(x)-1)^2], \tag{4}$$

where x and y correspond to the original CT and PET respectively, and $\hat{y} = G(x)$ to the estimated PET. $\mathbb{E}_{x\sim p_{\mathrm{data(x)}}}$ and $\mathbb{E}_{y\sim p_{\mathrm{data(y)}}}$ represents the expected value of CT and PET data distribution, respectively. To enhance the quality of the estimated images by the GAN, four non-adversarial losses were implemented (pixel, perceptual, style, and entropy). The pixel reconstruction loss (an L1 loss) adopted, was proposed by *Isola et al.* in [14], and it serves to reduce the structural differences between the original and estimated images. This loss can be defined as:

$$\mathcal{L}_{\mathrm{Pixel}}(G) = \mathbb{E}_{x\sim p_{\mathrm{data(x)}}}[\|G(x)-y\|_1]. \tag{5}$$

The pixel reconstruction loss relies on the pixel distance measurement between a target image y and an estimated image $\hat{y}$, which has been found to produce blurry results, as

discussed in [15, 16]. To effectively assess the learned feature representations of the GAN network, the perceptual loss used in [17] was adopted. This loss minimises the discrepancy between high-level perceptual features from the target y and estimated $\hat{y}$ images, and consequently forces both feature representations to be similar. The feature maps can be either extracted from pre-trained feature extractors, or directly from the discriminator D layers. The adapted perceptual loss uses feature maps extracted from the discriminator D and can be expressed as:

$$\mathcal{L}_{Percep}(G) = \frac{1}{h_i w_i d_i} \sum_{i=0}^{M} \lambda_{cP,i} (\|D_i(y) - D_i(\hat{y})\|_1), \quad (6)$$

where D_i represents the intermediate feature map extracted from the i^{th} layer of the D. The discriminator consists of multiple hidden layers, denoted by M. Moreover, $\lambda_{cP,i}$ denotes the weighting factor assigned to the i^{th} layer, indicating its relative importance in the overall loss calculation.

To improve the generator's capability to capture complex patterns, Gatys *et al.* [18, 19] proposed a style reconstruction loss. This loss utilises the feature maps extracted from the D and computes their correlation along the depth dimension using the squared Frobenius norm of the difference between the Gram matrix of the original ($G_{ri}(y)$) and estimated ($G_{ri}(\hat{y})$) image. The Gram matrix and the style reconstruction loss can be defined as:

$$G_{ri}(y)_{m,n} = \frac{1}{h_i w_i d_i} \sum_{h=0}^{hi} \sum_{w=0}^{wi} D_i(y)_{h,w,m} D_i(y)_{h,w,n}, \quad (7)$$

$$\mathcal{L}_{Style} = \sum_{i=0}^{M} \lambda_{cS,i} \frac{1}{4 d_i^2} (\|Gr_i(y) - Gr_i(\hat{y})\|_F^2), \quad (8)$$

where $D_i(y)_{h,w,m}$ denotes the feature map extracted from the i^{th} layer of the discriminator D, with h_i, w_i, and d_i representing the height, width, and depth of the extracted feature space, respectively. The $\lambda_{cS,i}$ determines the influence of the Gram matrix from the i^{th} layer in the style reconstruction loss.

Efficient lossless image compression involves a crucial step: entropy coding. For the proposed prediction method, PET estimates closest to the original PET are desirable for coding purposes, as they result in lower residue entropy. Lower entropy residues are known to enhance the coding efficiency. To reinforce this, a loss term, denoted as $L_{Entropy}(G)$, is used. This loss aims to minimise the energy of the estimated residue, thus promoting higher efficiency in coding. The loss term is formulated as a function of the estimated entropy of the residue.

$$\mathcal{L}_{Entropy}(G) = -\sum_{i=1}^{M} p_i \log_2 p_i, \quad (9)$$

where p_i denotes the probability of occurrence for each of the M individual pixel values in the residue (y - $\hat{y}$).

The final loss function used for the proposed GAN framework for CT-to-PET image translation is defined as follows:

$$\begin{aligned} \mathcal{L}(D,G) = \mathcal{L}_{\mathrm{LSGAN}}(D) + \lambda_{\mathrm{LSGAN}} \times \mathcal{L}_{\mathrm{LSGAN}}(G) \\ + \lambda_{\mathrm{p}} \times \mathcal{L}_{\mathrm{Pixel}}(G) + \mathcal{L}_{\mathrm{Percep}}(G) + \mathcal{L}_{\mathrm{Style}}(G) \\ + \lambda_{\mathrm{E}} \times \mathcal{L}_{Entropy}(G), \end{aligned} \quad (10)$$

where λ_{E}, λ_{LSGAN}, and λ_{p} are the weights of the entropy, least-square, and pixel reconstruction loss respectively. The discriminator used in this work adopted the 70×70 PatchGAN architecture proposed by Isola *et al.* [14], featuring five discriminative layers, while the generator employed the ResNet architecture [20] with 9 residual blocks.

3. EXPERIMENTAL ASSESSEMENT

3.1. Dataset

In order to evaluate the performance of the proposed predicted scheme, "The Cancer Imaging Archive (TCIA)" [21] was used. This dataset comprises a total of 2111 PET-CT pairs, from where 1391 were allocated for training and 721 for testing. The PET images have a resolution of 128×128, while the corresponding CT images have a resolution of 512×512. Both PET and CT images are represented in grayscale with a bit depth of 8. To ensure consistency and remove irrelevant regions, the images underwent a pre-processing step, by first defining a head segmentation mask, cropping the background. Subsequently, zero-padding was employed in order to equalise the variable resolution, resulting in 100×100 for PET images and 320×320 for CT.

3.2. Experimental Setup

The GAN was trained for 1500 epochs, employing a batch size of 4, and batch normalisation. Different hyperparameters were tested using a grid search method. The following parameters were set as follows: $\lambda_E = 10$, $\lambda_{Pixel} = 10$, $\lambda_{LSGAN} = 1$, $\lambda_{cP,0} = 5$, $\lambda_{cP,1} = 5$, $\lambda_{cP,2} = 2.5$, $\lambda_{cP,3} = 1.5$, $\lambda_{cP,4} = 1$, $\lambda_{cS,0} = 5$, $\lambda_{cS,1} = 5$, $\lambda_{cS,2} = 2.5$, $\lambda_{cS,3} = 1.5$, and $\lambda_{cS,4} = 1$. The Adam optimizer [22] was used with momentum terms $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 2×10^{-4}. A linear learning rate policy was employed, where the learning rate decreases linearly to 0 after 250 epochs. During the training process, an additional convolutional layer was incorporated as the initial layer of the generator to perform the downsampling of CT images in order to align them with the PET resolution.

The coding gain (CG) was adopted as the performance metric to evaluate the coding efficiency of the proposed approach, which is defined as follows:

$$CG = \frac{\text{Size(ref)} - \text{Size(x)}}{\text{Size(ref)}} \times 100 \quad (11)$$

where Size(ref) and Size(x) are the compressed data size obtained using the VVC lossless in intra mode, and with the proposed prediction method, respectively.

The optimal values for α_1 and α_2 were determined by means of a grid search algorithm. The coding performance of various configurations of α_1 and α_2 were evaluated, with α values restricted to positive integers to guarantee lossless coding compatibility with the VVC. Moreover, considering that images within the dataset have a bit depth of 8, and the VVC standard supports a maximum bit depth of 16, it is important that the sum of α_1 and α_2 does not exceed 256.

3.3. Results

To assess the effectiveness of the proposed framework and find optimal values of α_1 and α_2, an initial evaluation was conducted without the use of Sw3. The results obtained with the proposed prediction method are presented in Table 1 for different combinations of α_1 and α_2. The table presents the average CG (%) for PET images across, with negative values indicating a decrease in coding efficiency, compared to the VVC intra-coding.

Table 1. Average CG (%) for PET images using different α values

α_2 \ α_1	**1**	**2**	**3**	**4**	**5**
1	**7.14**	-4.46	-5.75	-6.97	-7.81
2	-14.86	-17.27	-18.86	-18.40	-19.47
3	-25.38	-28.1	-30.24	-32.09	-33.24
4	-31.47	-28.2	-34.70	-35.80	-36.40
5	-38.06	-39.8	-41.54	-43.05	-43.92

The results in Table 1 also make it clear that only when $\alpha_1 = \alpha_2 = 1$ the proposed prediction method yields positive coding gains (7.14%, in the case). Based on these results, $\alpha_1 = \alpha_2 = 1$ is chosen as the optimal configuration for the alpha values. The percentage of PET images where coding gains were obtained (compared to the alternative $\alpha_1 = 0$ and $\alpha_2 = 1$) was also assessed, and the results are presented in Table 2. These results cover a range of α values tested in the grid search that resulted in at least one PET image with superior coding efficiency.

Table 2. Percentage (%) of the number of PET images with higher coding efficiency when compared to [8]

α_2 \ α_1	**1**	**2**	**3**	**4**	**5**
1	**53.6**	0.41	0.14	0.14	0.14

The results shown in Table 2 (improvements only on 53.6% of the PET images compared to [8]) motivated to the development of a more adaptive and versatile solution. It was observed that the configuration with $\alpha_1 = 0$ and $\alpha_2 = 1$ (Sw3 in Fig. 1) approximates the results of [8]. Accordingly, aiming to maximise the overall coding gains, and based on real-time coding efficiency assessment, an embedded decider that enables the dynamic selection of the best coding method for each frame was implemented (thus stating the selection of Sw3 and Sw4). The relative performance for the three different prediction methods, with and without the brightness and contrast optimisation procedure, is shown in Table 3.

Table 3. PET average CG(%)

Brightness and Contrast Optimisation	**[8]**	**Proposed**	**Proposed with embedded decider**
-	7.17	7.14	13.04
✓	7.19	7.46	**13.20**

The results in Table 3, show that without the optimisation procedure, the prediction method described in [8] achieved a CG of 7.17%. Also, that the proposed method without the embedded decider, achieved a similar CG of 7.14%. However, with the incorporation of the embedded decider, an improvement in coding efficiency was achieved, resulting in a CG of 13.04%. This set of results clearly demonstrates that the optimisation leads to CG improvements for all prediction methods. Additionally, by using the brightness and contrast optimisation the proposed method surpassed the results of [8] even without the use of the embedded decider. The highest coding gain of 13.20% was obtained when using the proposed method combined with the embedded decider and optimisation. The obtained results reinforce the efficiency of the proposed prediction approach in enhancing the coding efficiency of PET images, namely when the embedded decider is used.

Table 4. PET-CT pair average CG(%).

Brightness and Contrast Optimisation	**[8]**	**Proposed**	**Proposed with embedded decider**
-	0.70	0.70	1.30
✓	0.70	0.76	**1.32**

The coding efficiency of the combined PET-CT image pairs was also evaluated, and the results are shown in Table 4. It is worth noting that the CG values are lower than those of the PET modality alone discussed earlier. This happens since CT images have a much larger size and higher coding efficiency compared to PET images. However, as discussed about Table 3, the usage of the optimisation procedure and the inclusion of the embedded decider further improves the CG value in PET-CT pairs, where the value of 1.32% was achieved when using both proposed configurations.

4. CONCLUSIONS

This paper introduces a novel approach to enhance the lossless coding efficiency of paired images using the VVC stan-

dard by effectively leveraging multimodal medical image information. The proposed approach includes an embedded decider that dynamically selects the optimal prediction method for each frame based on its coding efficiency, thereby improving the overall performance. The multimodal prediction method outperforms previously proposed cross-modality coding schemes, achieving coding gains of up to 13.20% compared to the reference VVC intra-coding of the corresponding PET image. The experimental results demonstrate effectiveness of the proposed method in improving the efficiency of lossless coding for multimodal medical images.

5. REFERENCES

[1] J. C. Waterton and Martin Braddock, *Chapter 1 Medical Imaging: Overview and the Importance of Contrast*, pp. 1–20, Royal Society of Chemistry, 2012.

[2] S. Coughlin and D. Roberts et al., "Looking to tomorrow's healthcare today: a participatory health perspective," *Internal Medicine J.*, vol. 48, no. 1, pp. 92–96.

[3] J. Xingyu, M. Jiayi, and et al. X. Guobao, "A review of multimodal image matching: Methods and applications," *Information Fusion*, vol. 73, pp. 22–71, 2021.

[4] T. Vagenas, T. Economopoulos, and C. Sachpekidiset al., "A decision support system for the identification of metastases of metastatic melanoma using whole-body fdg PET/CT images," *IEEE J. of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1397–1408, 2023.

[5] Z. Yu, X. Han, and S. Zhang et al., "Mousegan++: Unsupervised disentanglement and contrastive representation for multiple MRI modalities synthesis and structural segmentation of mouse brain," *IEEE Tran. on Medical Imaging*, vol. 42, no. 4, pp. 1197–1209, 2023.

[6] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multimodal medical image synthesis," *IEEE Tran. on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[7] J. Parracho, L. A. Thomaz, L. M. N. Távora, S. M. M. Faria, and P. A. Assunção, "Cross-modality lossless compression of PET-CT images," *Proc. Conf. on Telecommunications - ConfTele*, Feb. 2021.

[8] J. Parracho, L. A. Thomaz, L. M. N. Távora, and S. M. M. Faria et al., "Lossless coding of multimodal image pairs based on image-to-image translation," *European Workshop on Visual Information Processing*, Jun. 2022.

[9] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen, "Image-to-image translation: Methods and applications," *IEEE Tran. on Multimedia*, pp. 1–1, 2021.

[10] J. Nelder and R. Mead, "A simplex method for function minimization," *The Computer J.*, vol. 7, no. 4, pp. 308–313, Jan. 1965.

[11] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," *CoRR*, Mar. 2019.

[12] X. Mao, Q. Li, and H. Xie et al., "Least squares generative adversarial networks," in *IEEE Int. Conf. on Computer Vision*, Venice, Italy, Oct. 2017, pp. 2813–2821.

[13] I. Goodfellow, J. Pouget-Abadie, and M. Mirza et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, vol. 27, pp. 2672–2680.

[14] P. Isola, J. Zhu, and T. Zhou et al., "Image-to-image translation with conditional adversarial networks," in *IEEE Conf. on Comp. Vision and Patt. Recognition*, Honolulu, HI, United States, Jul. 2017, pp. 5967–5976.

[15] D. Pathak, P. Krähenbühl, and J. Donahue et al., "Context encoders: Feature learning by inpainting," in *Conf. on Comp. Vision and Pattern Recognition*, Las Vegas, NV, USA, Jul. 2016, pp. 2536–2544.

[16] R. Zhang, P. Isola, and A. Efros, "Colorful image colorization," *CoRR*, vol. abs/1603.08511, 3 2016.

[17] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, Mar. 2016.

[18] S. Schreiber, J. Geldenhuys, and H. de Villiers, "Texture synthesis using convolutional neural networks with long-range consistency and spectral constraints," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics Int. Conf.*, Stellenbosch, South Africa, Nov. 2016, pp. 1–6.

[19] L. Gatys, A. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 8 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Dec. 2016, pp. 770–778.

[21] K. Clark, B. Vendt, and K. Smith et al., "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *J. of digital imaging*, vol. 26, pp. 1045–1057, 7 2013.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Representations*, San Diego, CA, USA, 12 2014, pp. 1–13.

Automatic lung nodule classification in CT images using Two-stage CNNs and Soft-voting of Multi-scale Classifiers

Lipeng Xie
School of Cyber Science and Engineering, Zhengzhou University
Zhengzhou, China, 450001
lipengxie@zzu.edu.cn

Yubing Tong
Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania, United States, 19104
yubing@pennmedicine.upenn.edu

Yuan Wan
Department of Biomedical Engineering, Binghamton University, Binghamton, NY, United States 13902
ywan@binghamton.edu

***Abstract*—Lung nodule classification in Computed Tomography is an essential procedure for the diagnosis of lung cancer. Though some automatic methods have been proposed with high accuracy, the performance of these methods was heavily dependent on the amount and quality of data annotation and sensitive to the distribution of tissue density in CT. In this study, we propose a novel lung nodule classification system based on convolutional neural networks, achieving nodule detection and classification with good accuracy even using coarse annotated and low-quality data. Firstly, we constructed a nodule center-point detection method to predict the coarse coordinate of the nodule and the distance offset between the corner-points and the center-point of the nodule. Then, we extracted the multi-scale region of interest s and shape features of the nodule within an region of interest and then input them into the nodule classification network for predicting the nodule grade and type label. The proposed method was tested on 822 cases yielding a precision of 0.962 and a recall of 0.934 for nodule detection and an accuracy of 0.759 for nodule grade classification.**

Keywords—convolutional neural network, classification, lung nodule, Computed Tomography

I. INTRODUCTION

Automatic lung nodule classification in CT (Computed Tomography) plays a crucial role in the early detection and diagnosis of lung cancer[1]. Lung nodules in CT images are small, round or oval-shaped abnormalities found in the lungs, and they can be an early sign of lung cancer or other lung diseases. However, manual detection and classification of the nodules is time-consuming, labor-intensive, and suffers from cognitive variability. Therefore, there is an urgent need for the development of automatic lung nodule classification systems, aiming to assist radiologists in detecting and characterizing these nodules in CT scans more efficiently and accurately. In general, the computer-aided diagnosis system of lung cancer involves analyzing the large volumes of image data generated by CT scans and identifying potential nodules for further evaluation. The main challenges for this task are the complex shape variability, low contrast, and poor signal-to-noise ratio of CT.

To overcome the problems, several traditional machine learning based lung classification methods had been proposed. For example, Lee et al. [2] proposed an ensemble classification aided by clustering method to automatically predict the nodule grade of 2D CT images with a high sensitivity of 0.98 and specificity of 0.97. In [3], the authors utilized 3D active contour method to segment the nodule and linear discriminant analysis classifier to classify the grade of nodule. However, these traditional machine learning models were sensitive to the image noise. Recently, encouraged by the development of deep learning (DL) technology, some DL based automatic lung detection and classification models were presented. Xie et al. [4] utilized the Faster R-CNN (convolutional neural networks) to construct a lung nodule detection for 2D CT image, obtaining a good accuracy on the public dataset. To reduce the false-positive rate of nodule detection, El-Regaily et al. [5] proposed a multi-view 3D CNN to output the multi-view detection results and classify the detected nodules. Similarly, the authors in [6] utilized vision transformer model to locate and classify lung nodule in CT. However, the performance of these models heavily depends on the amount and quality of annotation data.

In this study, we propose a novel nodule classification system based on convolutional neural networks for CT images, achieving a good agreement with the manual detection and classification results. The main innovations of our study include: i) A fully automatic and accurate system for nodule grade and type classification in CT; ii) A novel design of network architecture to detect the region of interest (ROI) of nodule allowing a rough bounding box of the nodule to train the model; iii) An effective verification that the handcrafted shape feature is helpful to nodule classification, iv) A new design of multi-scale classification networks fusion

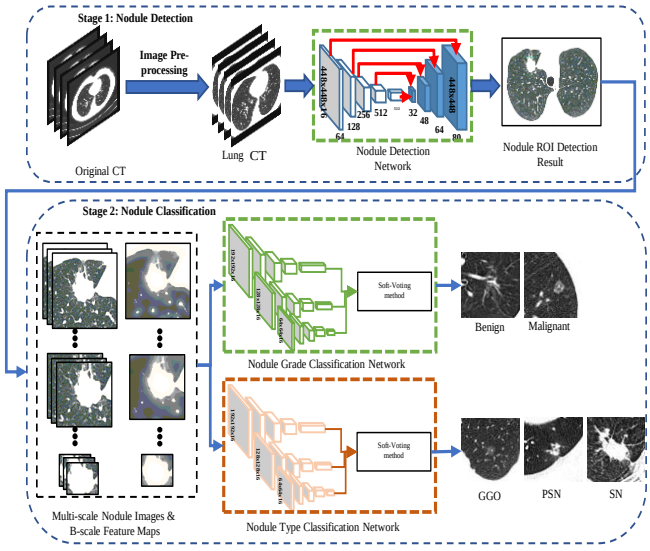


Figure 1. The structure of the proposed lung nodule classification system

method to improve the nodule classification performance.

II. MATERIELS AND METHOD

Figure 1 is a schematic depiction of the main stages in our method. In the first stage, we proposed a nodule detection network based on the corner-point concept and U-Net [7] to locate the nodule in the 3D images. To reduce the negative effect of the surrounding tissue and normalize the 3D size of CT images, a simple pre-process was utilized by first thresholding out lung tissue (lung CT) and then do morphological processing before nodule detection. Based on the output from first stage, multi-scale ROIs are implemented and B-scale filter were utilized [8] to compute the B-scale values of each pixel in the ROIs, representing the largest ball of homogeneous intensity. Then, we input the multi-scale ROIs with B-scale feature maps into the nodule grade and type classification networks with the soft-voting method for classifying the grade and type label of the lung nodule (2nd stage in Figure 1).

A. Data Acquisition, Annotation, and Augmentation

Table 1. Summary of the Lung CT data set

Dataset	Benign			Malignant		
	GGO	PSN	SN	GGO	PSN	SN
Training set (subjects)	61	63	240	470	517	293
Validation set (subjects)	10	11	40	78	86	49
Testing set (subjects)	30	31	120	235	259	147
Total (subjects)	101	105	400	783	862	489
Image Size (voxel)	695x695x5~695x695x46					

Data acquisition and annotation are essential prerequisites for supervised learning. We collected 2,740 axial CT scans with data allocation for training, validating and testing in Table 1. Each nodule was annotated with a grade label (Benign vs. Malignant) and a type of label (GGO (ground-glass opacity), PSN (part-solid), and SN (solid)), as shown in Figure 2.

In this study, we divided the CT images of subjects into training, validation, and testing set by the ratio of 6:1:3. To increase the diversity of samples, we utilized four data augmentation methods to enhance the training set, including annotation position drift, contrast adjustment, image rotation, and image resize. The total training data is 26,304 CT images.

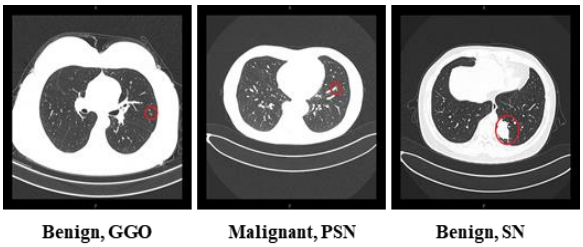


Figure 2. Examples of lung images annotation

B. Pre-processing for Lung CT Images

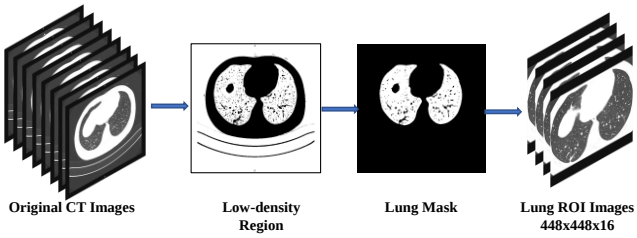


Figure 3. The pre-processing operations for lung CT images

The tissue surrounding the lungs produces abundant unnecessary computation. We designed pre-processing operations including: i) segmenting the low-density regions using the threshold for lung tissue, ii) searching the connected regions and remaining the two smallest regions as the lung mask, and iii) calculating rectangle boundaries of lung mask and extracting the lung ROI images. To unify the image size of training, we resized the lung ROI images to 448x448x16 using linear interpolation method. In our experiments, the threshold value was set to 120.

C. Nodule Detection Network

In this study, we only have the rough location from a clinician, who just utilized a red elliptic (Figure 2) of suitable size to represent the coarse location of nodules in the CT images in order to save time. We designed a center-point (of a nodule) based detection network and calculated the top-left and bottom-right points of elliptic annotation as the ground truth label. Based on the previous work [9], we designed a novel nodule detection network CNN based on U-Net [7] to segment the center-point of nodule ROI and predict the distance offset between the top-left and bottom-right points and center-point of nodule ROI, as shown in Figure 4.

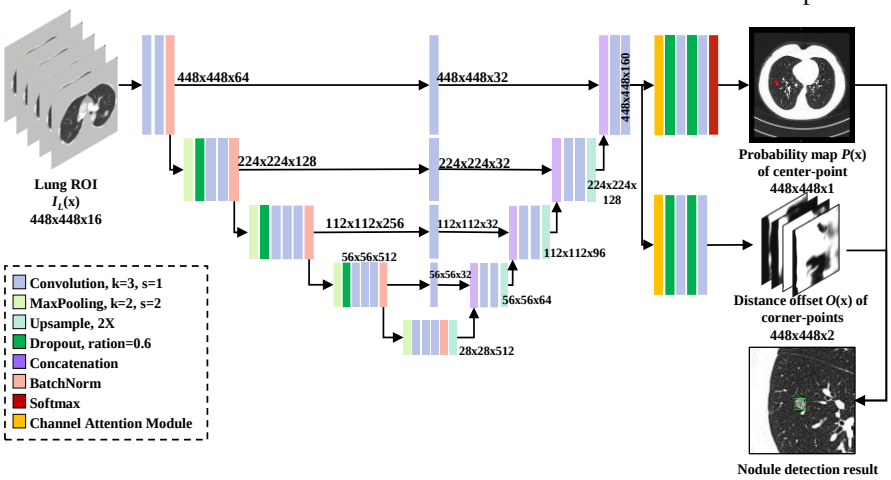


Figure 4. The structure of the proposed lung nodule detection network

The proposed network includes four modules: i) the backbone network (modified VGG19 [10]) for learning the multi-scale and level feature from the input image, including 16 convolutional layers with kernel size of 3x3 and stride of 2, 4 max-pooling layers with kernel size of 2x2 and stride of 2, 5 batch-normalization layers, and 3 dropout layers with the dropout ratio of 0.6, ii) the feature fusion network for integrating the multi-scale feature maps with size 448x448x32, 224x224x32, 112x128x32, 56x56x32, and 28x28x32 together as a feature pyramid, including 4 up-sampling layer based on linear interpolation method with the scale factor of 2, 8 convolutional layers with kernel size of 3x3 and stride of 1, 4 convolutional layers with kernel size of 1x1, stride of 2 and output channel number of 32, and 4 concatenation layers based on channel stacking operation, and iii) the pixel-wise classifier for predicting the probability of each pixel attributed to nodule center-point category, including 1 channel attention module [11], 2 convolutional layers with kernel size of 3x3 and stride of 1, 2 dropout layers with ratio of 0.6, and 1 soft-max function, and iv) the pixel-wise regressor for predicting the distance between each pixel and top-left and bottom-right points, with the same structure of pixel-wise classifier but without soft-max function.

The nodule center-point segmentation result can be used to calculate the coarse coordinate of the nodule. To improve the nodule detection performance, we proposed a post-processing method by fusing the center-point and distance offset information. Given the nodule center-point segmentation probability map $P(x) \in \mathbb{R}^{448\times448\times1}$ and the distance offset map $O(x) \in \mathbb{R}^{448\times448\times4}$, our model would search the maximum point (x_m, y_m) in $P(x)$ as the center-point of nodule and compute the coordinate of top-left and bottom-right points $(x_{TL}, y_{TL}, x_{BR}, y_{BR})$ of nodule ROI by adding the coordinate of center-point with the offset value as following:

$$x_{TL} = x_m + \alpha O(x_m, y_m, 0), \quad (1)$$
$$y_{TL} = y_m + \alpha O(x_m, y_m, 1), \quad (2)$$
$$x_{BR} = x_m + \alpha O(x_m, y_m, 2), \quad (3)$$
$$y_{BR} = y_m + \alpha O(x_m, y_m, 3), \quad (4)$$

where α represents the distance scale factor with a constant value. Then, we employed the coordinate of top-left and bottom-right points to extract the nodule ROI from the lung ROI images.

To train this model, we defined the loss function by combing the cross-entropy with L2-norm loss functions as follows:

$$L(S_G, O_G; W_D) = -\frac{1}{N}\sum_{x\in\Omega}\log\big(P(S_G(x)|x)\big) + \frac{\lambda_1}{N}\sum_{x\in\Omega}\|O_G(x) - O(x)\|_2 + \lambda_2\|W_D\|_1, \quad (5)$$

where $S_G(x)$ and $O_G(x)$ represent the ground truth of the center-point label and distance offset value at pixel x, $P(S_G(x)|x)$ represents the probability value of pixel x classified as ground truth $S_G(x)$, $O(x)$ denotes the predicted distance offset at pixel x, W_D and Ω represent the parameters of the network and image domains, N is the total number of pixels, $\|\cdot\|_1$ and $\|\cdot\|_2$ are the L1-norm and L2-norm, and λ_1 and λ_2 serve as trade-off parameters among the three terms.

D. Multi-scale Nodule Classification Network with Soft-voting

As shown in Figure 5, we proposed a nodule classification network using VGG-19 [10] network and B-scale feature map, including three steps: i) extracting the nodule ROI with image size 128x128x16 from the original 3D CT image based on the nodule detection result and computing the B-scale feature of the ROI, ii) learning the fusion feature of nodule from the nodule ROI and B-scale feature map, and iii) predicting the grade or type label of nodule ROI by the soft-max classifier. The structure of nodule classification network is similar with the VGG-19 network, adding the normalization layer and channel attention module into the network to improve the feature learning ability. In addition, we utilized the fully connection layer and soft-max function for predicting the possibility of the inputted nodule ROI belongs to different nodule grades and types. We utilized the focal loss to define the loss function of nodule classification as follows:

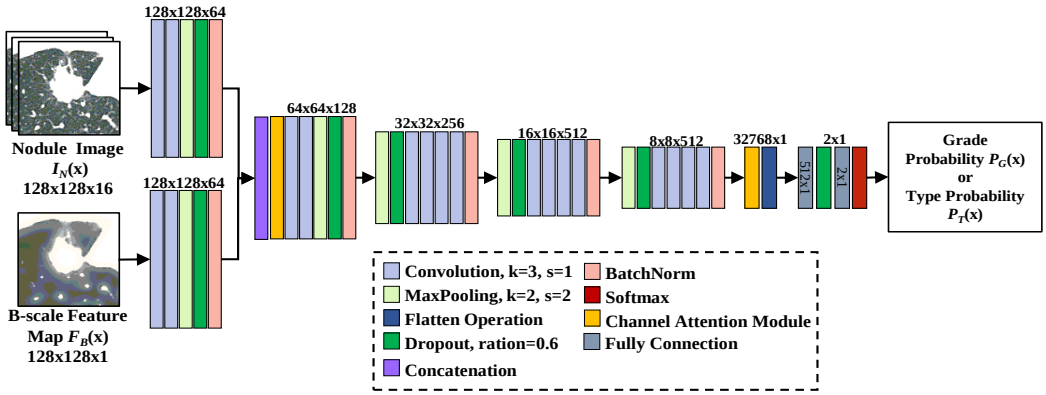


Figure 5. The structure of the lung nodule classification network

$$L(L_i, I_i; W_C) = -\frac{1}{M}\sum_{i=1}^{M}\alpha_i(1 - P_C(L_i|I_i))^{\beta}\log\big(P(L_i|I_i)\big) + \lambda_3\|W_C\|_1, \quad (6)$$

where L_i represents the ground truth label of image I_i, $P_C(L_i|I_i)$ denotes the possibility value of image I_i as ground truth label L_i, α_i and β are the category weight value and modulating factor, M is the total number of samples, and W_C represents the parameters of nodule classification network.

By analyzing the experimental results, we found that the nodule ROI size has an obvious effect on nodule

classification accuracy. In general, the predicted classification possibility value of nodule would be varied with ROI size. To improve the nodule classification performance, we trained 5 nodule classification networks using 5 different ROI sizes, including 192x192x16, 160x160x16, 128x128x16, 96x96x16, and 64x64x16, and fusing the predicted classification possibility by soft-voting method as follows:

$$P_m(I_j) = \sum_{i=1}^{5} w_i P_i(I_j), \quad (7)$$

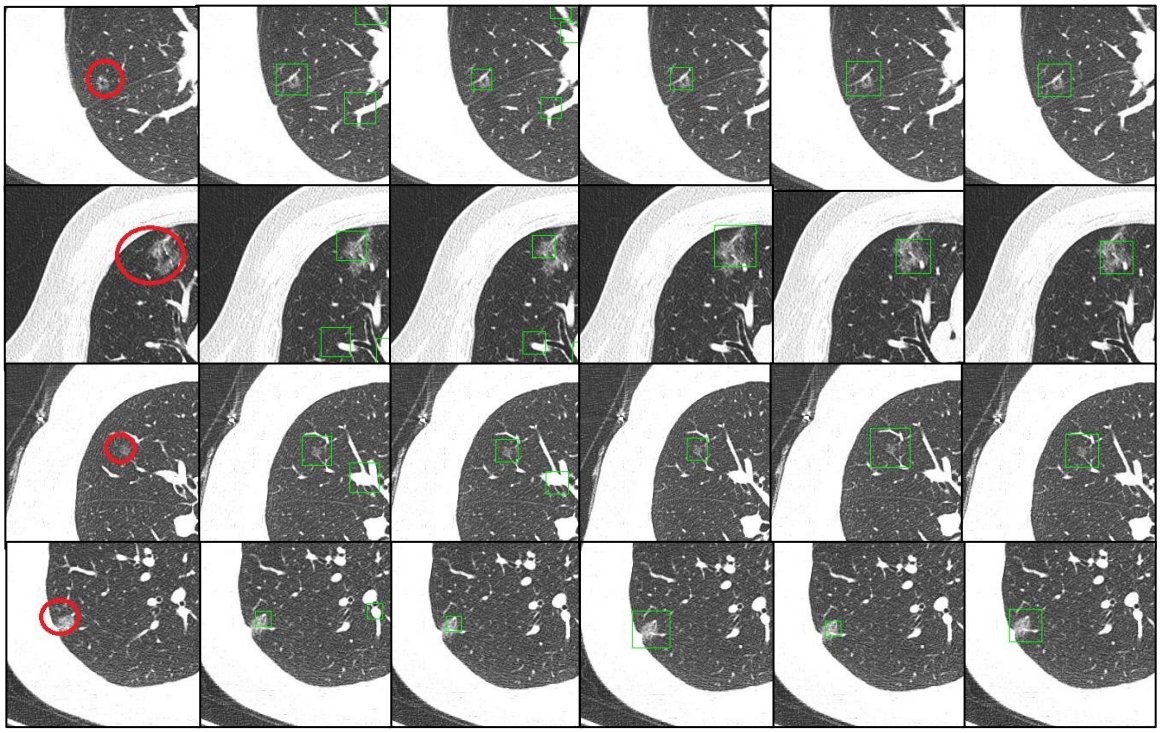
where $P_i(I_j)$ and $P_{\mathrm{m}}(I_j)$ represent the output classification probability of *i*-th classifier and multi-scale classifier for the image I_j, and w_i denotes the weighted value of the *i*-th classifier. The weight value w_i can be calculated as follows:

$$w_i = \frac{\mathrm{auc}_i}{\sum_{j=1}^{5} \mathrm{auc}_j}, \quad (8)$$

where auc_i represents the AUC (Area Under Curve) value of the *i*-th classifier on the validation data set.

III. Experiments and Results

A. Experimental details



(a) (b) (c) (d) (e) (f)

Figure 6. Examples of nodule detection (in green boxes). The green rectangle and red elliptic represent the automatic and manual nodule detection results. (a) Ground truth, (b) Nodulenet, (c) SANet, (d) proposed nodule detection network, (e) proposed nodule detection network with pre-processing method, and (e) proposed nodule detection network with pre- and post-processing methods.

In this study, we utilized the open source library TensorFlow to implement the proposed nodule detection and classification models and the training data in Table 1 to train models. All experiments were conducted on a PC with an Intel i7-7700K CPU and two NVIDIA 1080 Ti GPUs. The hyper-parameters for optimizing the models were as follows: learning rate (0.00001), batch size of training data (50), and number of iterations (500). In addition, we employed the same conditions to train several classic detection and classification networks, including Nodulenet [12], SANet [13], AlexNet [14], VGG-19 [10], ResNet-18 [15], and ViT [16].

B. Metrics

We employed our model and comparison algorithms to detect and classify the nodule on the testing set for evaluating the performance of all models. We applied several metrics to quantitatively analyze the nodule detection performance as following:

$$Precision = \frac{TP}{FP + TP}, \quad (9)$$

$$Recall = \frac{TP}{FN + TP}, \quad (10)$$

where TP, FP, and FN represent the true positive, false positive, and false negative results, respectively. In our experiments, when the center-point of the detected nodule locates in the red circle, it would be viewed as a true positive one otherwise a false positive one. Similarly, when a manual annotated nodule is not detected by the automatic method, it would be viewed as one false negative one.

To quantitatively analyze the nodule classification performance, we computed the classification accuracy (Acc) as below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (11)$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative results, respectively. Besides, we changed the classification thresholding value from 0.05 to 0.95 and computed the true positive rate (TPR) and false positive rate (FPR) as following:

$$TPR = \frac{TP}{TP + FN}, \quad (12)$$

$$FPR = \frac{FP}{TN + FP}. \quad (13)$$

Then, we plotted the ROC curve of nodule classification, in which the horizontal and vertical axis are the TPR and FPR values.

C. Nodule Detection Results

As shown in Figure 6, we show four nodule detection results in CT images. We find that our method obtains good agreement with the manual ground truth and that the pre- and post-processing methods can improve the accuracy of nodule detection. In addition, the false-positive problem of Nodulenet and SANet is very obvious. The main reason is that the nodule detection performance of the two networks is highly dependent on the quality of annotation data. Table 2 summarizes the quantitative results of nodule detection for all models. Our method achieves a high precision vale (0.962) and a high recall value (0.934), meaning that our method has a low false-positive and false-negative rates. The comparison among the performance value in the last three rows shows that the pre-processing operations can slightly increase the precision and recall of nodule detection and the post-processing method can remove most false-positive detection results with decreasing a little recall value.

Table 2. Nodule detection performance on the testing dataset

Method	Metrics	
	Precision	Recall
Nodulenet [12]	0.552	0.785
SANet [13]	0.636	0.864
Proposed model	0.682	0.923
Proposed model + Pre-processing method	0.786	0.963
Proposed model + Pre-and Post-processing method	**0.962**	**0.934**

D. Nodule Grade (benign vs. malignant) Classification Results

Table 3. Nodule grade classification performance on the testing set

Method	Metrics	
	Acc	AUC
AlexNet	0.646	0.684
VGG-19	0.683	0.716
ResNet-18	0.723	0.723
ViT (patch size=8, depth=6)	0.708	0.704
Proposed model (Nodule ROI size = 128)	0.716	0.72
Proposed model (Nodule ROI size = 128) + B-scale feature	0.732	0.753
Proposed model (Nodule ROI size = [64, 96, 128. 160, 192]) + B-scale feature + Soft-voting method	**0.759**	**0.808**

As illustrated in Table 3, we compare the quantitative grade nodule classification results of our method with other classification models. By analyzing the experimental data, we can see that our method obtains the highest accuracy and AUC values. In addition, we show the ROC curve of all models in the Figure 7, which can intuitively demonstrate the nodule grade classification performance differences among the models. We can find that our model outperforms other comparative methods and the B-scale feature and soft voting of multi-scale classifiers can improve the nodule classification accuracy. The lung nodule grade classification performance of our method seems to be lower than some existing nodule classification systems [17, 18]. But keep in mind that we tried to build our system more robust and practical following the

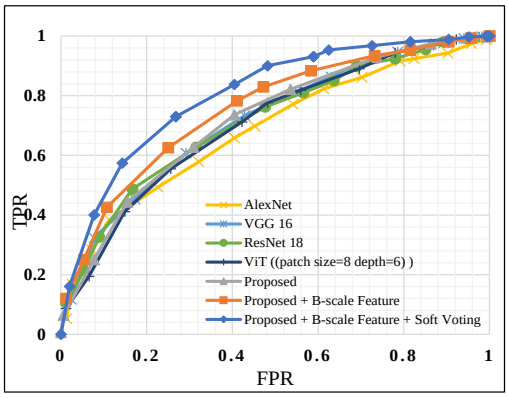


Figure 7. ROC curve for nodule grade classification

real clinician operations. In this study, clinicians have much less burden to label the data we used. In general, the annotation requirements and regularity of dataset in this study are much lower in our system than others [17,18].

E. Extending: Nodule Type (GGO, PSN and SN) Classification

In this study, we trained another proposed nodule

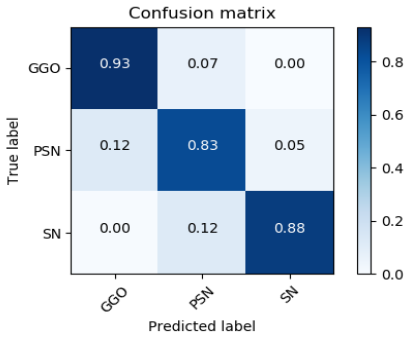


Figure 8. Confusion matrix for nodule type classification

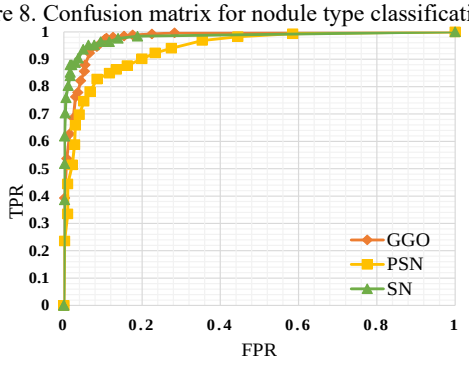


Figure 9. ROC curve for nodule type classification

classification to predict the type label of nodules, achieving an accuracy of 0.878. Figure 8 illustrates the confusion matrix for nodule type classification, demonstrating that our method obtains a high true-positive rate and low false-positive rate both for GGO, PSN, and SN nodule classification. Besides, the ROC curve in Figure 9 illustrates that the nodule type classification performance of our method with high accuracy.

Compared Figure 7 with Figure 9, we can find that there is an obvious difference between grade and type classification performance. That is interesting and has never been reported according to our knowledge. The reason might be that the nodule grade classification is more sensitive to the tissue density value of the nodule. While the nodule type classification is more dependent on one the edge information of the nodule.

IV. Conclusions

In this paper, we utilized CNN and coarse annotated dataset to construct an automatic lung nodule classification system for CT images for improving the accuracy and efficiency of lung cancer diagnosis. Our approach contains two steps: i) detecting the location of the nodule center-point and estimating the distance offset between the corner points and center-point of the nodule for generating the ROI of the nodule; 2) classifying the nodule grade and type label using the multi-scale classification networks with the handcrafted shape features and soft-voting method. Experimental results demonstrate that our approach achieves a good agreement with manual detection and classification of lung nodules in CT. Considering the difference in shape and texture among nodule types, we are analyzing the nodule grade classification performance in different types of nodules. The proposed approach adopted U-Net but can be easily propagated to other architectures, and is much potential for multiple applications of object localization and classification.

References

[1] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, and B. Geurts, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," Medical image analysis, vol. 42, pp. 1-13, 2017.

[2] S. L. A. Lee, A. Z. Kouzani, and E. J. Hu, "Random forest based lung nodule classification aided by clustering," Computerized medical imaging and graphics, vol. 34, no. 7, pp. 535-542, 2010.

[3] T. W. Way, L. M. Hadjiiski, B. Sahiner, H. P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours," Medical physics, vol. 33, no. 7Part1, pp. 2323-2337, 2006.

[4] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," Pattern Recognition, vol. 85, pp. 109-119, 2019.

[5] S. A. El-Regaily, M. A. M. Salem, M. H. A. Aziz, and M. I. Roushdy, "Multi-view Convolutional Neural Network for lung nodule false positive reduction," Expert systems with applications, vol. 162, pp. 113017, 2020.

[6] H. Mkindu, L. Wu, and Y. Zhao, "Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization," Biomedical Signal Processing and Control, vol. 85, pp. 104866, 2023.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation." pp. 234-241.

[8] U. Bagci, J. K. Udupa, and X. Chen, "Ball-scale based hierarchical multi-object recognition in 3D medical images." pp. 1267-1278.

[9] L. Xie, J. K. Udupa, Y. Tong, J. M. McDonough, C. Wu, C. Lott, J. B. Anari, P. J. Cahill, and D. A. Torigian, "Automatic lung segmentation in dynamic thoracic MRI using two-stage deep convolutional neural networks." pp. 948-954.

[10] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks." pp. 7132-7141.

[12] H. Tang, C. Zhang, and X. Xie, "Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation." pp. 266-274.

[13] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, and H. Zhang, "SANet: A slice-aware network for pulmonary nodule detection," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 8, pp. 4374-4387, 2021.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770-778.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[17] C. Tong, B. Liang, Q. Su, M. Yu, J. Hu, A. K. Bashir, and Z. Zheng, "Pulmonary nodule classification based on heterogeneous features learning," IEEE Journal on Selected Areas in Communications, vol. 39, no. 2, pp. 574-581, 2020.

[18] A. Nibali, Z. He, and D. Wollersheim, "Pulmonary nodule classification with deep residual networks," International journal of computer assisted radiology and surgery, vol. 12, pp. 1799-1808, 2017.

Bayesian Multispectral Videos Super Resolution

Abdelhamid N. FSIAN*, Jean-Baptiste THOMAS*, Jon Y. HARDEBERG†, Pierre GOUTON*
* IMVIA, Université de Bourgogne
†Colourlab, department of computer science, NTNU
{firstname.lastname}@u-bourgogne.fr* or @ntnu.no†

***Abstract*—Due to hardware limitations, multispectral videos often exhibit significantly lower resolution compared to standard color videos. These videos capture images in multiple bands of the electromagnetic spectrum, providing valuable additional information that is not available in traditional RGB images. This paper proposes a Bayesian approach to estimate super resolved images from low-resolution spectral videos. We consider adjacent frames from a video sequence to provide a super-resolution image at a time. We include in our proposal the motion between adjacent frames and unlikely to the literature, we estimate the blur and noise while reconstructing the higher resolution image. Experimental results on spectral videos demonstrate the effectiveness of our approach in producing high-quality super resolved images.**

I. Introduction

Multispectral imaging has become a major asset in various fields such as remote sensing, medical imaging, and surveillance. However, the acquired images are often of low resolution, which limits their usefulness in applications that require high-quality images. Super resolution (SR) techniques aim to overcome this limitation by reconstructing high-resolution images from low-resolution ones. In their work, Nasrollahi and Moeslund [1] emphasized the importance of conducting a comprehensive literature review in this particular domain.

While significant progress has been made in the field of spectral image super resolution [2], achieving high-quality super resolution for multispectral video sequences remains a challenging task. Theses challenges arise from various factors, including arbitrary motion of objects and cameras, unknown noise levels, and the presence of motion blur and point spread functions that introduce unknown blur kernels. Prior work often relied on oversimplified assumptions, assuming simple parametric motion forms and known blur kernels and noise levels. However, these assumptions do not hold in practical scenarios, making the super resolution problem more intricate and demanding a more comprehensive approach. Therefore, to develop a practical super resolution system, it is necessary to simultaneously estimate the optical flow [3], noise level [4], and blur kernel [5] ,in addition to reconstructing the high-resolution frames. Since each of these sub-problems has been thoroughly investigated in the field of computer vision, it is natural to integrate them into a unified framework without making oversimplified assumptions.

In this study, we introduce a Bayesian framework using Maximum A Posteriori knowledge (MAP) [6] for reconstructing super resolved multispectral images from low-resolution multispectral videos. Our method takes advantage of the temporal, spatial and spectral correlation between adjacent frames to enhance the super resolution process. Using a sparsity prior for the high-resolution image, flow fields, and blur kernel. The MAP inference iterates between optical flow, noise estimation, blur estimation, and image reconstruction to estimate the optimal values for these parameters. This approach enables us to handle different types of blur kernels and noise models, making our method adaptable and robust in a range of scenarios. Despite different noise levels and blur kernels, our method successfully reconstructs both large-scale structures and small texture features in difficult real-world sequences. The remainder of this paper is organized as follows. In Section 2, we provide a review of related works in the field of spectral super resolution, highlighting the limitations of existing approaches. This review serves as a foundation for understanding the motivation behind our proposed Bayesian framework. Section 3 presents the core of our method, focusing on the image reconstruction process using the Bayesian MAP approach. We describe the key components and their interplay in achieving high-quality super resolution for multispectral video sequences. In Section 4, we present the results obtained from applying our Bayesian framework to real-world multispectral video sequences. We discuss the performance of our method and provide in-depth analysis and discussions on the reconstructed super-resolved images before to conclude.

II. Related Work

A variety of methods have been developed for multispectral super resolution, based on different mathematical models, deep learning architectures, or physical priors. In this section, we provide a review of some of the most representative methods in this field, organized by their underlying principles and approaches.

A. Single Frame Super Resolution

Super resolution from a single frame has been an active area of research in computer vision. Chang et al. [7] made significant contributions to the field, and their work serves as an important foundation. Early research in super resolution addressed the ill-posed problem of reconstructing high-resolution images from low-resolution frames [8]. Schultz et al. [9] utilized spatial priors to overcome the absence of constraints. Bascle et al. [10] considered motion blur using an affine motion model, while Hardie et al. [11] jointly estimated translational motion and the high-resolution image. However, these motion models have limitations in

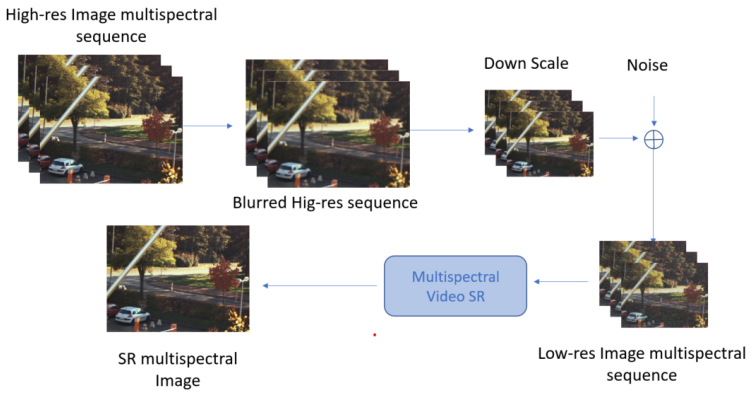


Fig. 1: **super resolution diagram**. To create the observed low-resolution multispectral sequence, the high-resolution sequence is downsampled, each frame is smoothed with a blur kernel, and corrupted with an uncorrelated noise.

capturing the complexity of real-world sequences. Baker and Kanade [12] proposed an approach using optical flow and a parametric motion model to handle motion in super resolution. Fransens et al. [13] introduced a probabilistic framework based on Expectation-Maximization algorithm, but their assumptions about blur kernels and Gaussian priors may affect edge preservation. Recent advancements in optical flow, such as those presented by Brox et al. [14], have provided more reliable techniques based on sparsity priors. On a different scoop, deep learning-based approaches have shown promising results in super resolution of images and videos. However, these methods require a large amount of training data and are computationally expensive [15]. Moreover, Multispectral database are still limited, compared to the RGB database [16].

B. Multi Frame Super Resolution

Super resolution techniques that utilize video sequences or multiple frames have also been explored. Irani et al. [8] proposed a method for enhancing the physical spatial resolution of multispectral images through logical reallocation of spectra. Takeda et al. [17] employed 3D kernel regression inspired by the non-local means technique for video denoising, exploiting spatiotemporal neighboring relationships for video up-sampling. While their technique still requires motion estimation in locations with significant motion, it offers a different approach to super resolution from video sequences. Liu and Freeman [18] developed a video denoising method with accurate motion estimation despite heavy noise. We aim to leverage these advancements in optical flow for more precise super resolution in our work.

In the domain of multispectral image super resolution, Vega et al. [19] proposed a Bayesian approach for super-resolution reconstruction of multispectral images using pansharpening. Pansharpening is the process of fusing high-resolution panchromatic and low-resolution multispectral images to create a single high-resolution color image. The proposed method incorporates prior knowledge on the expected characteristics of multispectral images, including smoothness within each band and correlation between bands. Zhi-Wei et al. [20] introduced an algorithm called SRIF (Multispectral Image Super-Resolution via RGB Image Fusion and Radiometric Calibration) that fuses low-resolution multispectral images with high-resolution RGB images to reconstruct high-resolution multispectral images. However, the linear relationship assumption between multispectral and RGB images may not always hold true, and the requirement of both low-resolution multispectral and high-resolution RGB images can be limiting. Lanaras et al. [21] developed a convex optimization method for improving the spatial resolution of lower-resolution bands in multispectral images. Their adaptive regularizer preserves edges and learns discontinuities, assuming these discontinuities are located in the same positions across all bands, which may not always hold true in practice.

III. Multispectral Super Resolution: Imaging Pipeline and Parameter Estimation

Our objective is to recover the high resolution sequence $\{I\}$ from the low resolution sequence $\{J\}$. In order to take advantage of multispectral video, we attempt to estimate the super resolved frames I_i using the neighboring low resolution frames J_{i-1}, J_i, J_{i+1}. We consider that the low resolution frame J_i is the result of down-sampling of I_i, smoothed with a blur kernel and corrupted with noise. Furthermore, we assume that the noise and kernel blur are consistent across all spectral bands. It simplifies the estimation process and reduces computational complexity. Thus the model of obtaining J_i is illustrated in Figure 1.

In order to estimate the high-resolution sequence and reverse the decay of resolution mentioned above, we need to estimate the noise level, the blur kernel and the motion. Among the unknown parameters in the generative models, the smoothing kernel K, which is equivalent to point spread functions in the imaging process or the smoothing filter when video is down scaled, parameter θ_i which controls the noise, and w_i which represent the motion information between consecutive frames. For this we will use a Bayesian method called MAP as defined in [22] and in the equation (1).

$$\{I', K', \{\theta_i\}', \{w_i\}'\} = \underset{I,K,\{\theta_i\},\{w_i\}}{\arg\max} p\left(I, K, \{\theta_i\}, \{w_i\} | \{J_i\}\right) \quad (1)$$

The model estimates the unknown parameters, such as the smoothing kernel and noise level, using adjacent frames and applies Bayesian MAP inference to find the optimal solution. The goal is to maximize the posterior probability, which is the product of prior and likelihood according to [23], and developed in equation (2).

$$p(I, K, \{\theta_i\}, \{w_i\} | \{J_i\}) \propto p(I)p(K)\prod_i p(w_i)\prod_i p(\theta_i) \quad (2)$$

$$p(J_0|I, K, \theta_0)\prod_{i=-1, i\neq 0}^{1} p(J_i|I, K, \theta_i, w_i)$$

Where i is the mutlispectral frame index. We assume an exponential distribution for the likelihood in order to handle outliers [24] (equation (3)).

$$p\left(J_i|I, K, \theta_i\right) = \frac{1}{Z\left(\theta_i\right)} exp\left\{-\theta_i ||J_i - DBF_{w_i} I||\right\}, \quad (3)$$

where $Z\left(\theta_i\right) = \left(2\theta_i\right)^{-dim(I)}$ and especially the parameter θ_i represents the noise level of frame i. The matrices D and B stand for respectively, down sampling and filtering with kernel blur K. Moreover, F_{w_i} is the warping matrix that correspond to the flow w_i.

To model the priors of image I, optical flow field w_i and blur kernel K, we used sparsity on derivative filter responses. Sparsity on derivative filter responses is a technique used to model the priors of image, optical flow field, and blur kernel in the Bayesian model for super resolution. The sparsity constraint encourages the filter responses to be mostly zero, except for a few significant values, which helps to reduce noise and improve the accuracy of the estimation. This technique is commonly used in signal processing and computer vision applications [25] to promote efficient and robust representations of signals and images (See equations (4) , (5) , (6)). For more mathematical explanation, we refer the reader to this paper [26].

$$p\left(I\right) = \frac{1}{Z_I\left(\eta\right)} exp\left\{-\eta ||\nabla I||\right\} \quad (4)$$

$$p\left(w_i\right) = \frac{1}{Z_w\left(\lambda\right)} exp\left\{-\lambda\left(||\nabla u_i|| + ||\nabla v_i||\right)\right\} \quad (5)$$

$$p\left(K\right) = \frac{1}{Z_K\left(\gamma\right)} exp\left\{-\gamma ||\nabla K||\right\} \quad (6)$$

Where $Z_I(\eta)$ (equation (4)), $Z_w(\lambda)$ (equation (5)), and $Z_K(\gamma)$ (equation (6)) are normalization constants that depends only on η, λ and γ. Moreover, ∇ (equation (7)) is defined as the gradient, by extension

$$\begin{aligned} ||\nabla I|| &= \sum ||\nabla I(n)|| \\ &= \sum \left(|I_x(n)| + |I_y(n)|\right) \end{aligned} \quad (7)$$

where $I_x = \frac{\partial}{\partial x} I$, $I_y = \frac{\partial}{\partial y} I$ and n is the pixel index. The flow field's horizontal and vertical components, u_i and v_i, uses the same notation. As proposed by Liu et al. [4]. We assumed that the conjugate prior for θ_i is a Gamma Distribution (equation (8)) :

$$p\left(\theta_i; \alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \theta_i^{\alpha-1} exp\left\{-\theta_i \beta\right\}. \quad (8)$$

Now that we have the prior and likelihood probability distributions, we can use coordinate descend to do the Bayesian MAP inference. Please note that the model have five free parameters which are $\eta,\ \lambda,\ \gamma,\ \alpha,\ \beta$.

A. Reconstructing High-Resolution Images

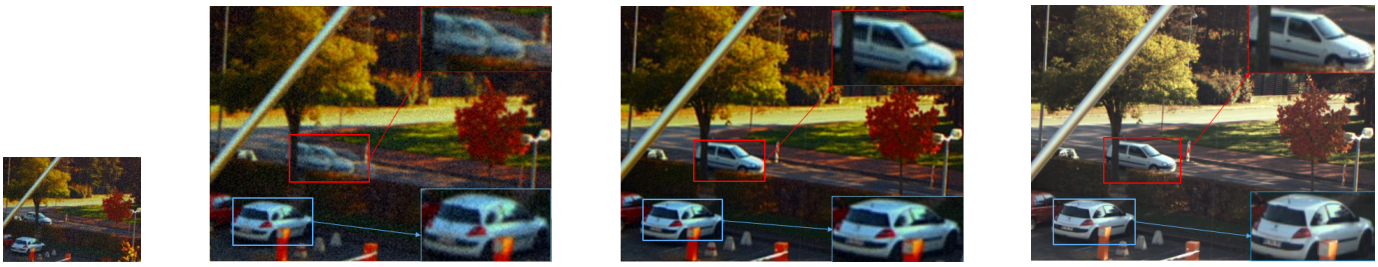
We estimate the high resolution multispectral image by calculating the equation (9), using the most recent estimations for the flow field w_i, the blur kernel K, and the noise level θ_i

$$\begin{aligned} I' = \arg\min_I \theta_0 ||DBI - J_0|| \\ + \eta ||\nabla I|| + \sum_{i=-N, i\neq 0}^{N} \theta_i ||DBF_{w_i} I - J_i|| \end{aligned} \quad (9)$$

The term $\theta_0 ||DBI - J_0||$ measures the data fidelity or the discrepancy between the down sampled and blurred low-resolution image DBI and the observed low-resolution multispectral image J_0. Moreover, the term $\eta ||\nabla I||$, enforces sparsity on the gradient of the estimated high-resolution image I. This term promotes smoothness in the image while preserving edges and details. To use gradient-based methods, we replace the $L1$ norm with a differentiable approximation, which is $\xi\left(x^2\right) = \sqrt{x^2 + \epsilon^2}$ with $\epsilon = 0.001$. Moreover, the parameter η controls the strength of the sparsity regularization. A sum over adjacent frames with i indicating current frame index, where DBF_{w_i} denotes convolution operation using blur kernel estimate at ith frame index; this part encourages consistency between consecutive frames. The overall objective function aims to minimize differences between low resolution input and high resolution output while also incorporating constraints related to motion estimation. To solve this objective function, the iteratively reweighted least squares (IRLS) technique is employed. The IRLS [27] algorithm iteratively estimates the high-resolution image I by updating its estimate in each iteration.

B. Noise and Motion Estimation

We jointly estimate the flow field and the noise level on a Gaussian image pyramid knowing the high resolution image and the blur kernel. Therefore, the optical flow and noise level are iteratively evaluated for each pyramidal level. In the context of super-resolution, the Gaussian image pyramid is created by successively applying Gaussian blurring and down sampling operations to the original high-resolution image. Each level of the pyramid represents a different scale or resolution of the image. The coarsest level is the lowest resolution image, and as we move up the pyramid, the



(a) LR image (b) SR Bicubic (c) MAP (d) HR Image

Fig. 2: Our SR method is able to recover image details with a $\times 2$ upsacale.

resolution increases [28]. The following equation is the closed-form solution for the Bayesian MAP estimate for the noise parameter $\theta_i' = \frac{\alpha+N_q-1}{\beta+N_q\bar{x}}$.

Where $\bar{x} = \frac{1}{N_q}\sum_{q=1}^{N_q}|(J_i - BDF_{w_i}I)(q)|$ is defined as a sufficient statistic used to estimate the noise level in the Bayesian MAP approach (following the convention in [28]). Once the noise is known, the flow field w_i is computed using MAP and IRLS technique as depicted in equation (10).

$$w_i' = \arg\min_{w_i} \theta_i ||BDF_{w_i}I - J_i|| + \lambda||\nabla u_i|| + \lambda||\nabla v_i|| \quad (10)$$

The objective function is a weighted sum of data fidelity and regularization terms, where the regularization term $\lambda(||\nabla u_i|| + ||\nabla v_i||)$ enforces smoothness of the motion field and the high-resolution image. The optimization problem is solved iteratively using the IRLS method, which alternates between solving a linear system and updating the weights based on the current estimate. Here again we approximate the norm $|x|$ by $\xi\left(x^2\right)$

C. Blur Kernel Estimation

Following the notation from [23], we only demonstrate how to estimate the x-component kernel K_x given I and J_0 without losing generality and assuming that the kernel K is $x-$ and $y-$separable : $K = K_x \otimes K_y$, where K_y probability distribution is the same as K_x. Therefore, we define each row of the matrix A as the concatenation of pixels that correspond to the filter K. Moreover, we define $M_y : M_y K_x = K_x \otimes K_y = K$. The estimation of K_x is depicted in equation (11).

$$K_x' = \arg\min_{K_x} \theta_0 ||AM_yK_x - J|| + \gamma||\nabla K_x|| \quad (11)$$

The method involves solving an optimization problem that minimizes the difference between the low-resolution observation and the convolution of the high-resolution image with the estimated kernel, subject to a regularization term that encourages spatial smoothness of the kernel. The optimization problem is solved using the IRLS method.

IV. Results and Discussions

In our study on multispectral SR, we utilized a publicly available database from Benezeth et al. [29], which contains a collection of five VNIR (Visible and Near-InfraRed) multispectral videos containing between 250 and 2300 frames

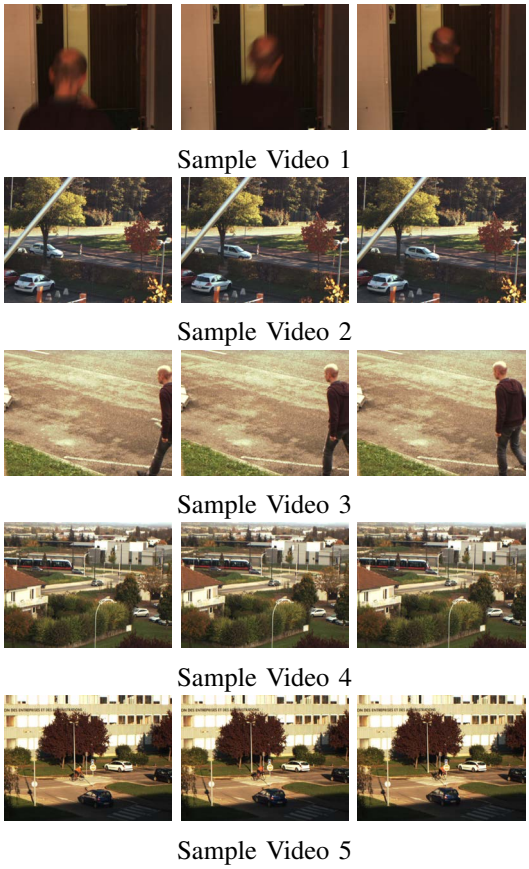


Fig. 3: Sample of three consecutive images from different videos. The video goes from left to right.

of spatial resolution 658×491. Each video in the database consists of a sequence of frames. Each frame contains seven spectral bands of which six are in the visible and one in the near infra-red (NIR), illustrations of the data-set are presented in Figure 3.

The inclusion of the NIR band allows for enhanced perception and analysis of various materials and phenomena that may exhibit distinct spectral characteristics in this region. The dataset provided a valuable resource for evaluating and benchmarking our proposed multispectral SR algorithm, enabling us to assess its performance across different spectral

TABLE I: Comparison of ESRGAN and MAP SR using average PSNR, SSIM, and RMSE from the Videezy4K dataset

Method	PSNR	SSIM	RMSE
ESRGAN $\times 2$	32.09	0.8793	0.00634
Ours $\times 2$	30.54	0.7467	0.076

bands and video sequences.

In our experimental analysis, we conducted a serie of evaluations to assess the performance and effectiveness of our proposed MAP for SR. To create a realistic simulation of real-world imaging conditions, we initially applied a blur operation followed by downscaling to the multispectral sequence. [1] This step aimed to mimic the inherent limitations and degradation commonly encountered in practical imaging scenarios. Furthermore, to account for the presence of noise in real data, we introduced a Gaussian noise into the blurred and downscaled images. Figure 2 shows respectively, the low resolution image, Bicubic SR image, MAP SR image (with a $\times 2$ upscale) and HR image.

In order to demonstrate the effectiveness of our Bayesian approach, we compared its results against those obtained using conventional interpolation methods, including Bicubic, nearest neighbor, and bilinear interpolation. By showcasing the comparative results, we highlight the distinct advantages and improvements achieved by our proposed Bayesian method in terms of both quantitative metrics and visual quality. We selected a set of 7 consecutive frames from each VNIR video containing scenes with moving objects. We considered factors such as object motion, scene complexity, and spectral diversity during the selection process. Specifically, we aimed to include scenes with diverse spectral content and varying degrees of motion to evaluate the performance of our SR MAP algorithm comprehensively. Furthermore, inspired by the versatility of our multispectral SR algorithm, we explore its applicability to RGB images. Unlike traditional SR algorithms that handle all channels simultaneously, our method treats each channel independently. To substantiate our findings, we conducted a comparison with a well-established state-of-the-art SR algorithm: ESRGAN [30] . We employed the RGB dataset from Videezy4K, a benchmark dataset renowned for its diverse and challenging image content. A dataset that contains 11 RGB videos and each videos contains 19 4K RGB images.

Comprehensive metrics including SSIM, PSNR and RMSE are provided in Table II for VNIR videos. The results from our experiments on the RGB dataset sourced from the Videezy 4K dataset are summarized in Table I.

The experimental results clearly demonstrate a significant drop in performance when motion estimation is omitted from the super-resolution process. Without motion estimation, the algorithm fails to capture the temporal coherence between frames, resulting in reduced image quality and an inability to effectively compensate for motion-related artifacts. On the other hand, incorporating motion estimation enables the algorithm to align frames and accurately estimate motion, leading to improved reconstruction quality and better preservation of fine details. These findings highlight the crucial role of motion estimation in achieving enhanced multispectral super-resolution by leveraging temporal information and mitigating artifacts.

V. Conclusion

In this study, we conducted a comprehensive analysis of multispectral super resolution utilizing a publicly available VNIR video database. Our research focused on developing a novel Bayesian method for spectral image SR and compared its performance against conventional interpolation techniques, namely Bicubic, nearest neighbor, and bilinear interpolation. Additionally, we extended our investigation to include a comparison with a state-of-the-art method, ESRGAN, which employs RGB image sequences. The results of this comparative analysis showcased the promising performance of our proposed algorithm. By leveraging neighboring frames to enhance the reference frame without the need for any training. This approach allowed us to achieve improved super-resolution results and narrowed the performance gap between our method and ESRGAN, highlighting the potential of our approach in the field of multispectral super resolution.

In our experiments, incorporating motion estimation, demonstrated superior performance compared to conventional interpolation techniques. By effectively capturing temporal coherence and reducing motion-related artifacts, our algorithm achieved improved reconstruction quality and preservation of fine details. The experimental results underscored the significance of motion estimation in multispectral SR, as neglecting this crucial aspect significantly impacted the overall performance. Our research contributes to the understanding of multispectral SR techniques, highlighting the importance of leveraging temporal information to enhance the desired frame.

For future work, an interesting avenue would be to explore the integration of joint super-resolution and demosaicing. Combining these two tasks could lead to more comprehensive image enhancement, addressing both spatial and color resolution simultaneously. By leveraging the strengths of both techniques, it is possible to achieve further improvements in the visual quality and fidelity of spectral images.

References

[1] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, pp. 1423–1468, 2014.

[2] J. M. Amigo, "Practical issues of hyperspectral imaging analysis of solid dosage forms," *Analytical and bioanalytical chemistry*, vol. 398, pp. 93–109, 2010.

[3] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[4] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman, "Automatic estimation and removal of noise from a single image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 299–314, 2007.

[5] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE signal processing magazine*, vol. 13, no. 3, pp. 43–64, 1996.

[6] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson, "Super-resolved surface reconstruction from multiple images," *Maximum Entropy and Bayesian Methods: Santa Barbara, California, USA, 1993*, pp. 293–308, 1996.

	MAP		Bicubic		Bilinear		Nearest	
PSNR	w motion	o motion	w motion	o motion	w motion	o motion	w motion	o motion
VS 1	**35,79**	34.96	27.33	26.21	27.24	2565	26.36	24.52
VS 2	**34.6**	33.51	31.87	30.79	30.65	29.54	30.49	29.44
VS 3	**33.44**	32.27	29.46	26.34	29.08	25.49	28.93	24.25
VS 4	**31.89**	31.76	27.58	25.74	27.77	26.50	26.61	24.47
VS 5	**30.60**	30.41	27.32	26.21	27.23	25.65	26.36	24.52
SSIM								
VS 1	**0.95**	0.934	0.7427	0.6426	0.727	0.572	0.632	0.482
VS 2	**0.83**	0.78	0.76	0.73	0.75	0.7	0.74	0.59
VS 3	**0.87**	0.86	0.79	0.61	0.769	0.52	0.75	0.42
VS 4	**0.88**	0.87	0.70	0.55	0.74	0.64	0.62	0.46
VS 5	**0.87**	0.84	0.74	0.64	0.70	0.57	0.63	0.48
RMSE								
VS 1	**0.0162**	0.018	0.042	0.048	0.043	0.052	0.048	0.0599
VS 2	**0.02**	**0.02**	0.025	0.029	0.026	0.029	0.028	0.03
VS 3	**0.026**	0.024	0.033	0.048	0.035	0.053	0.035	0.061
VS 4	**0.025**	0.0258	0.041	0.051	0.04	0.047	0.046	0.059
VS 5	**0.029**	0.03	0.043	0.048	0.043	0.052	0.048	0.0593

TABLE II: Performance Metrics(PSNR, SSIM, RMSE) for each method with and without motion estimation and for each of the 5 VNIR video.

[7] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. IEEE, 2004, pp. I–I.

[8] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[9] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE transactions on image processing*, vol. 5, no. 6, pp. 996–1011, 1996.

[10] B. Bascle, A. Blake, and A. Zisserman, "Motion deblurring and super-resolution from an image sequence," in *Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings Volume II 4*. Springer, 1996, pp. 571–582.

[11] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[12] S. Baker and T. Kanade, "Super-resolution optical flow," Carnegie Mellon University, The Robotics Institute, Technical Report, 1999.

[13] R. Fransens, C. Strecha, and L. Van Gool, "Optical flow based super-resolution: A probabilistic approach," *Computer vision and image understanding*, vol. 106, no. 1, pp. 106–115, 2007.

[14] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*. Springer, 2004, pp. 25–36.

[15] S. Liu, Y. Yang, Q. Li, H. Feng, Z. Xu, Y. Chen, and L. Liu, "Infrared image super resolution using gan with infrared image prior," in *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2019, pp. 1004–1009.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[17] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 1958–1975, 2009.

[18] C. Liu and W. T. Freeman, "A high-quality video denoising algorithm based on reliable motion estimation," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*. Springer, 2010, pp. 706–719.

[19] M. Vega, J. Mateos, R. Molina, and A. K. Katsaggelos, "Super resolution of multispectral images using l1 image models and interband correlations," *Journal of Signal Processing Systems*, vol. 65, no. 3, pp. 509–523, 2011.

[20] Z.-W. Pan and H.-L. Shen, "Multispectral image super-resolution via rgb image fusion and radiometric calibration," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1783–1797, 2018.

[21] C. Lanaras, J. Bioucas-Dias, E. Baltsavias, and K. Schindler, "Super-resolution of multispectral multiresolution images from a single sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[22] M. Tipping and C. Bishop, "Bayesian image super-resolution," *Advances in neural information processing systems*, vol. 15, 2002.

[23] C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in *CVPR 2011*. IEEE, 2011, pp. 209–216.

[24] N. Wang and D. Yeung, "Bayesian robust matrix factorization for image and video processing," in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 1785–1792. [Online]. Available: https://doi.org/10.1109/ICCV.2013.224

[25] M. F. T. B. C. Russell and W. T. Freeman, "Exploiting the sparse derivative prior for super-resolution and image demosaicing," in *Proceedings of the Third International Workshop Statistical and Computational Theories of Vision*, 2003, pp. 1–28.

[26] H. S. Mousavi and V. Monga, "Sparsity-based color image super resolution via exploiting cross channel constraints," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5094–5106, 2017.

[27] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[28] Y. Kwon, S. Kim, D. Yoo, and S.-E. Yoon, "Coarse-to-fine clothing image generation with progressively constructed conditional gan." in *VISIGRAPP (4: VISAPP)*, 2019, pp. 83–90.

[29] Y. Benezeth, D. Sidibé, and J.-B. Thomas, "Background subtraction with multispectral video sequences," in *IEEE International Conference on Robotics and Automation workshop on Non-classical Cameras, Camera Networks and Omnidirectional Vision (OMNIVIS)*, 2014, pp. 6–p.

[30] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

Centralized Sample Expansion and Prior Correlation Evaluation for Hyperspectral Image Classification with Fully Convolutional Network

Ningyang Li
Faculty of Computer Science and Technology
Hainan University
Haikou, China
fes_map@qq.com

Zhaohui Wang*
Faculty of Computer Science and Technology
Hainan University
Haikou, China
william_hig@163.com

***Abstract*—Hyperspectral remote sensing image classification based upon deep learning has attracted increasing attention of researchers from various fields. Recently, fully convolutional network, provides a different perspective to cope with the pixel-wise classification of hyperspectral image. However, due to the issue of the limited sample and the diversity of samples, the fully convolutional network-based models which are not trained sufficiently generally cannot extract the discriminating spectral-spatial features for classification. In this paper, a fully convolutional network with centralized spectral-spatial sample expansion and prior correlation evaluation is proposed for hyperspectral image classification. By focusing on the center pixel, even if the samples expanded by the centralized spectral-spatial sample expansion module are in different scales, which remains the consistency of spatial structures. The prior correlation evaluation module then emphasizes the important areas of the expanded samples to help the residual fully convolutional network extract the multi-scale discriminating spectral-spatial features for classification. Experimental results on two classic data sets confirm the structural rationality of the proposal and its outstanding classification performances compared with the state-of-the-arts.**

Keywords—hyperspectral image classification, centralized spectral-spatial sample expansion, prior correlation evaluation, fully convolutional network, deep learning

I. INTRODUCTION

Tremendous advancements of the spectral and spatial resolutions of imaging sensors enable hyperspectral remote sensing image to acquire subtle spectral reflectance energy and abundant spatial distributions from the surface of Earth [1]. Such rich information provides solid support for better classification. Thus, hyperspectral image (HSI) classification has drawn growing attention of researchers from various fields, such as agriculture [2], military [3], urban planning [4], etc.

Conventional machine learning algorithms, including *k*-nearest neighbor [5], support vector machine [6], and so on, were exploited to deal with HSI classification. Although they received good classification accuracy, there are still some deficiencies which are difficult to mitigate. On one hand, these methods are not good at modeling the deep feature representation which is important for precise recognition, especially when HSI contains hundreds of bands [7]. On the other hand, some of them process the spectral or spatial information merely, which curbs the complementation of different types of features.

During the past decade, with the continuous upgrade of computational hardware, deep learning algorithms have reaped unprecedented achievement and become the dominant methods for image analysis. Earlier deep learning models for HSI classification, such as deep brief network [8], long short-term network [9], were used to extract the high-level spectral features from each pixel. However, due to the scarcity of spatial features, the classification results of these models are still unsatisfactory. Thanks to the emergence of convolutional neural network (CNN) [10], the feature extraction of HSI has been made great improvements [11], [12]. Spectral features can be extracted by 1-D CNN with less parameters. Spatial features can also be extracted by 2-D CNN while preserving the original spatial layouts. To integrate both kinds of features, HSI cube/patch, which contains the center pixel to be classified and its neighborhoods, has become the popular sample for HSI classification. The power of CNN has received extensive recognition and many techniques, including residual connection [13], dense connection [14], and capsule units [15], were combined with CNN to elevate the feature representation for classification performances.

Recently, fully convolutional network (FCN) [16] was proposed to resolve the problem that the arbitrary sizes of images cannot be well handled with a CNN-based model. The FCN adopts convolutional layers in the whole model for feature extraction and classification, which received better results than CNN for HSI semantic segmentation [17]-[19]. For example, a well pretrained deep FCN was introduced to explore the multiscale spatial structural information of the whole data set [20], [21]. But due to the different characteristic of HSI data sets, it is generally hard to prepare an appropriate pretrained model. Considering the issue of the limited samples, a patch-based training pattern with sparse point labels was proposed to increase samples for training, which relieved the insufficient training of model to some extent [22]. To further improve the classification accuracy, data enhancement, such as flip, rotation, rearrangement, was applied to produce more patches to meet the requirement of thorough optimization [23], [24]. A context-aware module was designed to help the FCN focus on spatial context dependency existed in different region of land-cover [25]. Samples with multi-view, including sub-pixel view, pixel view, and supe-pixel view, were sent to FCN to obtain more accurate prediction [26].

Although these methods have received good classification performances, there are still some deficiencies to be addressed. First, the HSI patch-based FCN models need a number of samples to avoid the issue of over-fitting. Some common data augment method, such as rotation, rearrangement, exert

979-8-3503-4218-5/23/$31.00 ©2023 IEEE

usually small improvement on classification results. This is because that convolution is not sensitive for orientation and position and the new spatial structures generated by the rearrangement method may be mutual or meaningless. Hence, an effective data augment method which takes the characteristics of HSI patch into account is demand for training the model. Second, due to the diversity of the pixel in a sample, it is generally difficult for existing methods to pay attention to the relevant spatial areas in samples, thereby hampering the representation of the discriminating spectral-spatial features. To mitigate these drawbacks, this paper proposes an FCN model which adopts the centralized spectral-spatial sample expansion (CSSSE) method and the prior correlation evaluation (PCE) for HSI classification. The CSSSE can infer the virtual pixel by considering the spectral and spatial correlations between real neighboring pixels at the same time. An expanded HSI patch is centered on the center pixel, which maintains the consistency of spatial structures. The PCE aims to highlight the relevant areas by assessing the relevancy between the center pixel and its neighborhoods. With the help of the two modules, the subsequent FCN can be well trained to extract more discriminating spectral-spatial features from the multi-scale samples. Experimental results confirm the effectiveness of the proposal and its outstanding classification performance.

The remainder of the paper is organized as follow. Section II introduces the proposed model in detail. Experiments and analyses are presented in Section III. Finally, this paper is summarized in Section IV.

II. Methodology

A. Overview of the Proposed Model

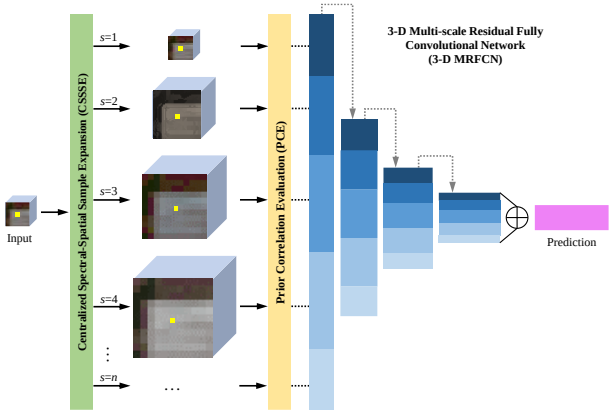


Fig. 1. Framework of the proposed model. Where "s", gray dotted arrows, and "⊕" denote the expansion scale, the residual connections, and element-wise addition, respectively.

The objective of the proposed model is to acquire the multi-scale discriminating spectral-spatial features for HSI classification. As shown in Fig. 1, the model consists of a CSSSE module, a PCE module, and a 3-D multi-scale residual fully convolutional network (3-D MRFCN). First, the input, an HSI patch $\mathcal{X} \in \mathbb{R}^{\omega\times\omega\times b}$, is sent to the CSSSE module to generate more samples in different scales, where ω and b indicate width and number of band, separately. ω is set to the optimal value 7. The expanded HSI patches focus on the center pixel and remain the original spatial structures of certain class. Next, the PCE module infers the relevant spatial areas of the multi-scale samples based on the correlation between the center pixel and neighborhoods. The meaningful pixels in the relevant areas are emphasized in the following stage. Finally, the 3-D MRFCN fuses the multi-scale discriminating spectral-spatial features extracted from the refined samples and assigns the most proper label for each pixel with softmax activation function. The more expanded samples and the residual connections enable the 3-D MRFCN to be optimized easily.

B. Centralized Spectral-Spatial Sample Expansion

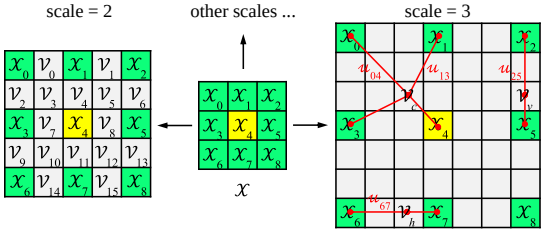


Fig. 2. Expansion processes of an HSI cube ($\omega = 3$) with the CSSSE method. Where yellow pixels, green pixels, and light gray pixels indicate the center pixels, neighboring pixels, and virtual pixels, respectively. Red lines denote the spatial distance between the virtual pixel and its referenced pixels.

An effective FCN model usually be supported by abundant samples. However, the limited samples and uneven number of samples are common for most HSI data sets. Some sample generation methods [23], [24], [27], such as flip, rotation, and rearrangement, may exert little improvement on convolution operation and disturb the special spatial textures of each class. Therefore, a novel CSSSE module is proposed to generate new samples in different scales. Taking the center pixel as the core, the CSSSE module gives enough thought to the spectral similarities and spatial correlations between real pixels during generating virtual pixels.

As shown in Fig. 2, the expanded HSI cubes are composed of the pixels from original HSI patch and virtual pixels $\mathcal{V} \in \mathbb{R}^{1\times1\times b}$. Specifically, there are three kinds of virtual pixels, crossing pixel $\mathcal{V}_c$, horizontal pixel $\mathcal{V}_h$, and vertical pixel $\mathcal{V}_v$, in each expanded HSI cube. For crossing pixel $\mathcal{V}_c$, four closest real pixels ($\mathcal{X}_0$, $\mathcal{X}_1$, $\mathcal{X}_3$, $\mathcal{X}_4$) are chose as the references for generation. While for the other two, $\mathcal{V}_h$ and $\mathcal{V}_v$, four closest real pixels ($\mathcal{X}_6$, $\mathcal{X}_7$) and ($\mathcal{X}_2$, $\mathcal{X}_5$) are chose, respectively.

To ensure the reality of virtual pixel, the spectral similarity $\mathcal{s}$ and spatial correlation $\mathcal{u}$ between virtual pixel and its referenced pixels are considered. Cosine distance [28], as an amplitude invariant metric, is utilized to measure the spectral similarity. The spatial correlation is evaluated based upon the spatial positions of virtual pixel and its referenced pixels. The two kinds of attributes can be described as follows.

$$\mathcal{s}_{ij} = \frac{\sum \mathcal{X}_i \cdot \mathcal{X}_j}{\sqrt{\sum \mathcal{X}_i^2}\cdot\sqrt{\sum \mathcal{X}_j^2}} \tag{1}$$

$$\mathcal{u}_{ij} = 1 - \sqrt{\frac{(r_v - r_i)^2 + (c_v - c_i)^2}{(r_j - r_i)^2 + (c_j - c_i)^2}} \tag{2}$$

Where $\mathcal{X}_i$, $\mathcal{X}_j$, r_v, and c_v denote ith and jth real pixels, the row and column of virtual pixel, separately.

Then, the virtual pixel $\mathcal{V}_{ij}$ is generated with the assistance of spectral similarity and spatial correlation

$$\mathcal{V}_{ij} = \frac{\mathcal{u}_{ij} + \mathcal{u}_{ij}\cdot\mathcal{s}_{ij}}{2}\cdot\mathcal{X}_i + \frac{(1-\mathcal{u}_{ij}) + (1-\mathcal{u}_{ij})\cdot\mathcal{s}_{ij}}{2}\cdot\mathcal{X}_j. \tag{3}$$

During the process, spectral similarity is weighted with the operation "$u \cdot s$", which assigns different spectral similarity for the pixels in different positions to relieve the unnatural spatial textures caused by same similarity. Then, the spectral similarity and spatial correlation are fused by the operation "$u + u \cdot s$" to make full use of the benefits for generation.

There are some disparities among the generation of three kinds of virtual pixels. For crossing virtual pixel v_c, its two spectral similarities, s_{04} and s_{13}, and two spatial correlations, u_{04} and u_{13}, are obtained first. Then the mean of virtual pixels, v_{04} and v_{13}, is set to the final result. For horizontal and vertical virtual pixels, v_h and v_v, they can be deduced directly by Equa. (1)-(3). The only discrepancy is that their spatial correlations, u_{67} and u_{25}, are simplified as follows.

$$u_{67} = 1 - (c_v - c_6)/(c_7 - c_6) \tag{4}$$

$$u_{25} = 1 - (r_v - r_2)/(r_5 - r_2) \tag{5}$$

The expanded samples in different scales behave similar spatial structures with the original HSI patch, which maintains the consistency of the spatial contexts of each class. Thus, the inter-class similarity and intra-class distance will not increase during the feature extraction of FCN.

C. Prior Correlation Evaluation

The PCE module is designed to capture the salient spatial areas based on the spectral correlation between the center pixel and its neighboring pixels, thereby improving the effectiveness of subsequent feature extraction.

Considering the instability of reflective energy of HSI, cosine distance, as an amplitude invariant measurement, is adopted to evaluate the spectral similarity between pixels. Then, a softmax activation function is utilized to convert the similarity into the importance coefficient. These processes can be summarized as follows.

$$s_i = \frac{\sum x_c \cdot x_i}{\sqrt{\sum x_c^2} \cdot \sqrt{\sum x_i^2}} \tag{6}$$

$$m_i = \frac{e^{s_i}}{\sum_{k=1}^{\omega \times \omega} e^{s_k}} \tag{7}$$

where x_c, x_i, s_i, and m_i represent the center pixel, ith neighborhood, spectral similarity of between the center pixel and ith neighborhood, and the importance coefficient of ith neighborhood, separately. This kind of spectral-similarity-based strategy has achieved good effects in exploring the relevant areas for HSI classification [28], [29].

D. 3-D Multi-scale Residual Fully Convolutional Network

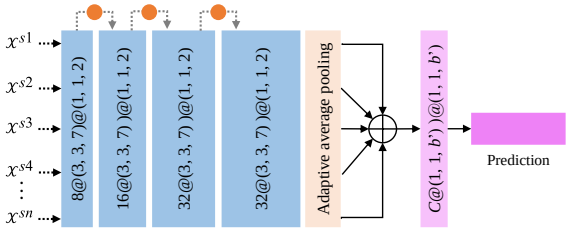


Fig. 3. Architecture of 3-D MRFCN.

To extract the multi-scale features from the expanded samples in different sizes, a 3-D MRFCN which can fuse the features in different scales is constructed. As shown in Fig. 3, the samples in n scales share the identical architecture, which enables the network to handle strong feature representation with less parameters. The 3-D MRFCN contains four 3-D convolutional layers with rectified linear unit activation function for feature extraction and a prediction layer. To relieve the vanishing gradient, the residual connection (gray dotted arrows) is introduced between the feature extraction layers. The orange circles denote the average pooling layers which can adjust the shape of feature maps to realize the transition of gradients. An adaptive average pooling layer is applied to squeeze the spatial sizes of feature maps in n scales and then they are integrated by the element-wise addition. At last, the prediction layer which equipped with C kernels with the size of $1 \times 1 \times b'$ and softmax activation function decides the final label, where b' denotes the number of bands of the feature maps output by the fourth convolutional layer.

III. Experiments and Analyses

A. Data Sets and Configuration

To prove the effectiveness of the proposed model, two publicly available HSI data sets [30], including Indian Pines and Botswana, are chose for experiments. Indian Pines data set was gathered by the Airborne Visible/Infrared Imaging Spectrometer sensor over Indian Pines test site. It consists of 145×145 pixels and 200 available bands. There are 16 categories in 10249 labeled pixels. Botswana data set was acquired by the Hyperion sensor mounted on the Earth Observing-1 satellite over the Okavango Delta, Botswana. It consists of 1476×256 pixels and 145 bands. There are 14 categories in 3248 available labeled pixels.

Before optimizing the proposed network, all samples are normalized by dividing the max grayscale value of data set and the parameters of all layers are initialized with *Xavier* normal distribution [31]. During the back propagation, the RMSprop optimizer [32] with the parameters (*learning rate*, *beta1*, *beta2*) = (0.001, 0.9, 0.999) is employed to reduce the error between true label and prediction. The batch-size and the number of iteration are set to 32 and 200, respectively.

B. Classification Results

In this section, the classification performance of proposed model is compared with other methods, including 3-D CNN [11], deep multi-scale spatial-spectral FCN (DMS^{3}FCN) [20], spectral-spatial 3-D FCN (SS3FCN) [23], and nonlocal-dependent learning network (NDL-Net) [25]. Each model is re-implemented according to the original article and shares the same samples as the proposed model.

1) Quantitative Evaluation

The numbers of training/test samples N, recall of each class, overall accuracy (OA), average accuracy (AA), kappa coefficient (κ), and training/test times T of different methods on two data sets are reported in Tables I and II. First, Compared with other four methods, the classification results of 3-D CNN are the lowest, but it cost the least time to finish training and test procedures with its simple architecture. Second, among three FCN-based models, DMS^{3}FCN aims to extract the deep multi-scale spectral-spatial features and receives better classification accuracy than 3-D CNN. By flipping and rotating the training samples during model

training, SS3FCN obtain rich features when predict new samples. NDL-Net is good at capturing the long-distance dependency between different regions of a class. But the OA of it on Indian Pines data set is a little lower. The possible reason is that some classes, such as No. 1, No. 7, and No. 9, only have one subregion, which results in the invalid non-local dependency. Third, our proposed model achieves the best classification performances on two data sets, which gains the most number of the highest recalls, OAs, AAs, and κs. For the classes having more samples of Indian Pines data set, e. g. No. 3, No. 10, and No. 11, the predictions of the proposal are satisfying. Moreover, though the proposal spends relatively long time for training, the test time of it on two data sets are the least compared with other three FCN-base models. This is because the proposal adopts the share architecture for different scales of samples, which reduces the number of parameters.

TABLE I. CLASSIFICATION PERFORMANCES OF DIFFERENT METHODS ON INDIAN PINES DATA SET WITH 15% TRAINING SAMPLES

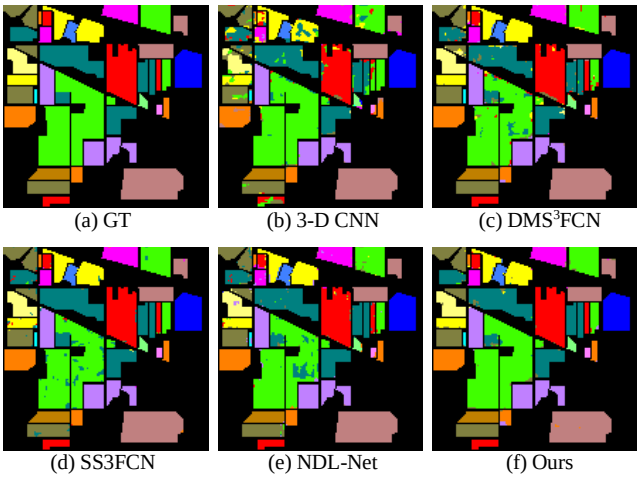
No.	*N*	3-D CNN	DMS^{3}FCN	SS3FCN	NDL-Net	Ours
1	7/39	64.86	90.00	**100.00**	83.33	**100.00**
2	214/1214	85.74	90.53	**98.71**	96.02	94.62
3	124/706	88.55	95.55	94.06	94.06	**99.81**
4	35/202	68.25	**100.00**	98.05	94.81	98.05
5	73/410	89.15	95.54	**99.68**	98.73	94.59
6	109/621	99.49	99.16	99.16	**99.79**	**99.79**
7	5/23	52.17	94.44	**100.00**	88.89	**100.00**
8	72/406	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
9	3/17	62.50	**100.00**	**100.00**	69.23	**100.00**
10	145/827	78.76	95.89	97.47	94.94	**99.53**
11	368/2087	94.55	89.29	91.73	90.85	**97.12**
12	88/505	80.80	**97.40**	95.06	92.21	95.32
13	31/174	**100.00**	**100.00**	**100.00**	99.25	**100.00**
14	190/1075	99.01	99.88	99.51	**100.00**	98.91
15	58/328	88.35	96.41	**99.20**	95.62	96.81
16	14/79	83.78	**100.00**	96.67	98.33	98.33
OA		93.57	94.73	96.59	95.29	**97.64**
AA		83.50	96.51	98.08	93.50	**98.30**
κ		88.90	95.75	97.05	94.64	**97.32**
T		295.57/1.56	482.69/4.22	546.35/5.55	466.84/4.85	494.39/3.54

TABLE II. CLASSIFICATION PERFORMANCES OF DIFFERENT METHODS ON BOTSWANA DATA SET WITH 15% TRAINING SAMPLES

No.	*N*	3-D CNN	DMS^{3}FCN	SS3FCN	NDL-Net	Ours
1	41/229	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
2	15/86	97.53	**100.00**	97.53	**100.00**	**100.00**
3	38/213	**100.00**	97.00	94.50	97.50	**100.00**
4	32/183	**100.00**	90.12	94.19	98.26	**100.00**
5	40/229	90.28	88.43	99.86	97.69	**100.00**
6	40/229	75.81	**98.60**	94.88	95.81	96.74
7	39/220	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
8	31/172	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
9	47/267	**100.00**	99.20	99.20	94.02	**100.00**
10	37/211	95.48	99.50	97.49	96.98	**100.00**
11	46/259	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
12	27/154	95.17	**100.00**	**100.00**	**100.00**	99.31
13	40/228	**100.00**	88.84	94.42	99.07	**100.00**
14	14/81	89.47	**100.00**	**100.00**	**100.00**	**100.00**
OA		96.61	97.00	97.96	98.27	**99.69**
AA		95.98	97.26	98.01	98.52	**99.72**
κ		95.87	96.75	97.79	98.12	**99.67**
T		157.81/0.90	269.21/1.66	393.25/1.92	300.18/1.77	354.58/1.22

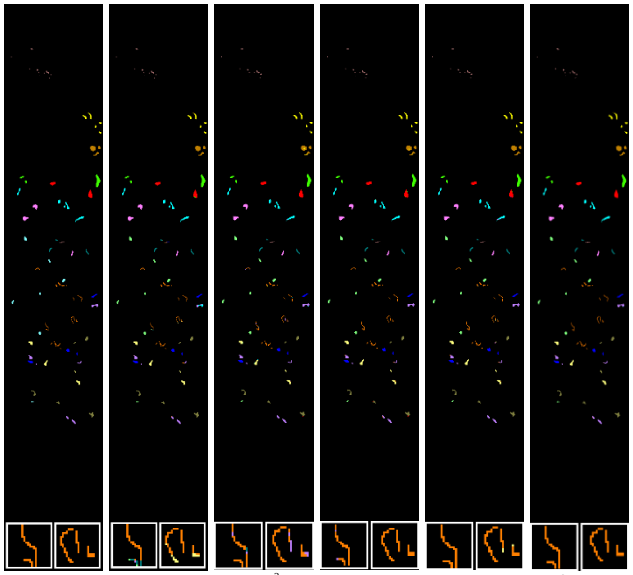
2) Qualitative Evaluation

The classification maps of different methods on two data sets are presented in Fig. 4 and 5. Compared with ground truth (GT) maps, the predictions of the proposal are purer. There are less dotted noises and speckles in each region of class. Some details are amplified in white squares in Fig. 5. For the class No. 11 which has the most number of samples of Indian Pines data set, the proposal obtains the absolutely same prediction, including edges and interiors, as the GT map. According to Table II, only two classes, No. 6 and No. 12, are not predicted with 100% probabilities, which is a superior result.



(a) GT (b) 3-D CNN (c) DMS^{3}FCN (d) SS3FCN (e) NDL-Net (f) Ours

Fig. 4. Classification maps of different methods on Indian Pines data set.



(a) GT (b) 3-D CNN (c) DMS^{3}FCN (d) SS3FCN (e) NDL-Net (f) Ours

Fig. 5. Classification maps of different methods on Botswana data set.

C. Ablation Study

TABLE III. OAS OF DIFFERENT COMBINATIONS OF MODULES

		CSSSE	PCE	3-D MRFCN	OA
Indian Pines	Comb_1	×	×	✓	95.59
	Comb_2	✓	×	✓	97.12
	Comb_3	×	✓	✓	96.33
	Comb_4	✓	✓	✓	**97.64**
Botswana	Comb_1	×	×	✓	96.98
	Comb_2	✓	×	✓	98.85
	Comb_3	×	✓	✓	97.47
	Comb_4	✓	✓	✓	**99.69**

To verify the rationality of each module of the proposal, different combinations of them are built for ablation study. As shown in Table III, Comb_4, also is the proposal, receives the highest OAs on two data sets. When the CSSSE and PCE modules are removed at the same time, the 3-D MRFCN is remained and achieves unsatisfying results. Compared with Comb_3, the CSSSE module will exert more improvement. This is because the PCE is designed to emphasize the relevant areas following the CSSSE module, which is a supplementary. If there is no CSSSE module ahead, the number of expanded samples will be reduced, which may weaken the positive influence of the PCE module.

D. *Impact of Expansion Scale*

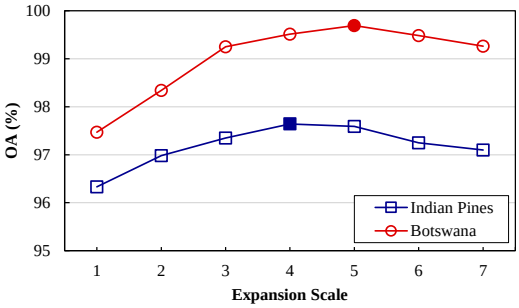


Fig. 6. OAs of the proposed model in different expansion scales. Where solid markers denote the default options.

The CSSSE module generates different sizes of samples in different expansion scales. To seek for the proper expansion scale, the proposal is trained with seven expansion scales. As shown in Fig. 6, the trends of the OAs on two data sets both increase first and then decrease. The best expansion scales for two data sets are 4 and 5, separately. Expansion scales smaller than the optimal values enable the model to learn robust feature representation with limited parameters. However, it may be difficult for the model to fit the data distribution if a bigger expansion scale is adopted. Because the number of samples and the size of samples are increased simultaneously.

E. *Impact of Data Enhancement Method*

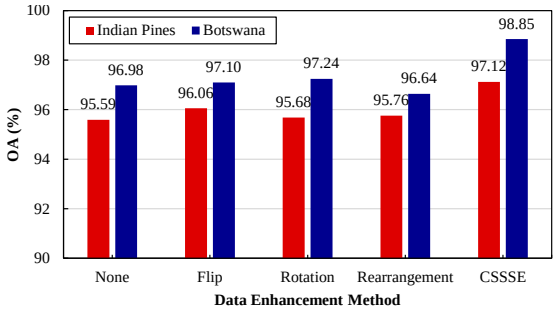


Fig. 7. OAs of the 3-D MRFCN with different data enhancement methods.

To confirm the effectiveness of the CSSSE module, it is compared with other data enhancement methods, including flip, rotation, and rearrangement. It can be seen clearly from Fig. 7 that the combination of the CSSSE module and the 3-D MRFCN receives the highest OAs on two data sets compared with other three strategies. This is because the flip and rotation operations are not always effective for the FCN-based model. Rearrangement is also a unstable method as it may disturb the individual spatial distributions of each class, which will cause smaller inter-class distance. Therefore, the proposed CSSSE module is an effective way for data enhancement.

F. *Impact of Training Sample Proportion*

To investigate the classification performances of different methods further, they are trained with different proportions of training samples. As shown in Fig. 8, the more the samples is trained, the higher the OAs of all methods are. When the training proportion is set to 15%, the growth rates of all curves decrease and remain a relatively stable level. Thus, the default training sample proportions of two data sets are all set to 15%. From this figure, the OAs of the proposal are higher than those of other methods even the training proportion is set to 5%.

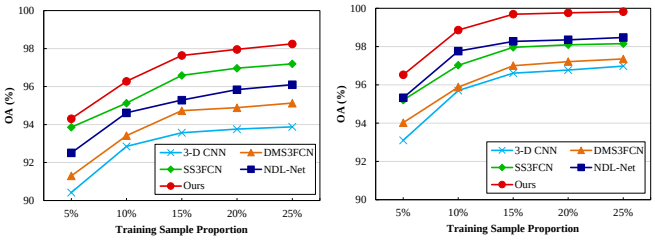


Fig. 8. OAs of different methods with different training sample proportions on Indian Pines (left) and Botswana (right) data sets.

IV. Conclusion

In this paper, an FCN model with the CSSSE and PCE modules is proposed to extract the multi-scale discriminating spectral-spatial features for HSI classification. By considering the spectral correlation and spatial distance, the expanded samples generated by the CSSSE module have high credibility. The peculiar spatial structures of each class is well maintained in expanded multi-scale samples. The PCE module aims to emphasize the relevant areas of samples with the spectral similarity between the center pixel and neighborhoods before feature extraction. With the assistance of the two modules, the subsequent 3-D MRFCN can represent the salient spectral-spatial features hidden in different scales of expanded samples. Experimental results on two classic HSI data sets demonstrate the rationality of the proposal and its superior classification performance compared with other state-of-the-arts.

Acknowledgment

This work was supported in part by the Framework of the Norwegian Research Council INTPART Project under Grant 309857 International Network for Image-Based Diagnosis (INID), and in part by the Hainan Key Research and Development Plan for Scientific and Technological Collaboration Projects under Grant GHYF2022015 - Research on Medical Imaging Aided Diagnosis of Infant Brain Development Diseases.

References

[1] M. Ahmad *et al*., "Hyperspectral image classification-traditional to deep models: A survey for future prospects," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 968-999, Jan. 2022.

[2] A. Khan, A. D. Vibhute, S. Mali, and C. H. Patil, "A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications," Ecological Informatics, vol. 69, pp. 101678, 2022.

[3] M. Shimoni, R. Haelterman, and C. Perneel, "Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques," IEEE Geosci. Remote Sens. Mag., vol. 7, no. 2, pp. 101-117, June 2019.

[4] R. Hänsch and O. Hellwich, "Fusion of multispectral LiDAR, hyperspectral, and RGB data for urban land cover classification," IEEE Geosci. Remote Sens. Lett., vol. 18, no. 2, pp. 366-370, Feb. 2021.

[5] K. Huang, S. Li, X. Kang, and L. Fang, "Spectral-spatial hyperspectral image classification based on KNN," Sens. Imaging, vol. 17, no. 1, pp. 1-13, 2016.

[6] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," Pattern Recognit., vol. 103, pp. 107298, Jul. 2020.

[7] D. L. Donoho, “High-dimensional data analysis: The curses and blessings of dimensionality,” AMS Math Challenges Lect., vol. 1, p. 32, Aug. 2000.

[8] Y. Chen, X. Zhao, and X. Jia, “Spectral-spatial classification of hyperspectral data based on deep belief network,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 6, pp. 2381-2392, Jun. 2015.

[9] F. Zhou, R. Hang, Q. Liu, and X .Yuan, “Hyperspectral image classification using spectral-spatial LSTMs,” Neurocomput., vol. 328, pp. 39-47, Feb. 2019.

[10] A. Krizhevsky *et al.*, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS), 2012, pp. 1097-1105.

[11] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, “Deep feature extraction and classification of hyperspectral images based on convolutional neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 10, pp. 6232-6251, Oct. 2016.

[12] N. Li and Z. Wang, “Hyperspectral image ship detection based upon two-channel convolutional neural network and transfer learning,” in Proc. IEEE 5th Int. Conf. Signal Image Process. (ICSIP), Oct. 2020, pp. 88-92.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comp. Vis. Patt. Rec. (CVPR), Jun. 2016, pp. 770-778.

[14] G. Huang, L. Zhuang, L. Maaten, and K. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comp. Vis. Patt. Rec. (CVPR), Jul. 2017, pp. 2261-2269.

[15] S. Sabour, N. Frosst, and G. Hinton, “Dynamic routing between capsules,” in Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS), 2017, pp. 3859-3869.

[16] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, Apr. 2017.

[17] J. Cui, K. Li, J. Hao, F. Dong, S. Wang, A. Rodas-González, *et al.*, “Identification of near geographical origin of wolfberries by a combination of hyperspectral imaging and multi-task residual fully convolutional network,” Foods, vol. 11, no. 13, pp. 1936, Jun. 2022.

[18] Z. Fang, G. Zhang, Q. Dai, B. Xue, and P. Wang, “Hybrid attention-based encoder-decoder fully convolutional network for PolSAR image classification,” vol. 15, no. 2, pp. 526, Jan. 2023.

[19] J. Gutiérrez-Zaballa *et al.*, “On-chip hyperspectral image segmentation with fully convolutional networks for scene understanding in autonomous driving,” Journal of Systems Architecture, vol. 139, pp. 102878, Jun. 2023.

[20] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, “Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 10, pp. 5585-5599, Oct. 2017.

[21] G. Jiang, Y. Sun, and B. Liu, “A fully convolutional network with channel and spatial attention for hyperspectral image classification,” Remote Sensing Letters, vol. 12, no. 12, pp. 1238-1249, Dec. 2021.

[22] X. Zhang, Z. Zheng, P. Xiao, Z. Li, and G. He, “Patch-based training of fully convolutional network for hyperspectral image classification with sparse point labels,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 8884-8897, 2022.

[23] L. Zou, X. Zhu, C. Wu, Y. Liu, and L. Qu, “Spectral-spatial exploration for hyperspectral image classification via the fusion of fully convolutional networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 659-674, 2020.

[24] H. Sun, X. Zheng, and X. Lu, “A supervised segmentation network for hyperspectral image classification,” IEEE Trans. Image Process., vol. 30, pp. 2810-2825, 2021.

[25] B. Tu, W. He, Q. Li, Y. Peng and S. Chen, “Fully convolutional network-based nonlocal-dependent learning for hyperspectral image classification,” IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-14, 2022, Art no. 5023414.

[26] M. K. S. Varma, R. Kulasekaran, and N. K. K. Rao, “HSIS-Net: Hyperspectral image segmentation using multi-view active learning based FCSN,” International Journal of Intelligent Engineering and Systems, vol. 16, no. 2, pp. 14-23, 2023.

[27] X. Kang, C. Li, S. Li, and H. Lin, “Classification of hyperspectral images by Gabor filtering based deep network,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 4, pp. 1166-1178, April 2018.

[28] N. Li, Z. Wang, F. A. Cheikh, and M. Ullah, “S^3AM: A spectral-similarity-based spatial attention module for hyperspectral image classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 5984-5998, 2022.

[29] N. Li and Z. Wang, “Spatial attention guided residual attention network for hyperspectral image classification,” IEEE Access, vol. 10, pp. 9830-9847, 2022.

[30] Hyperspectral Remote Sensing Scenes - Grupo de Inteligencia Computacional (GIC), http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes, accessed on 2023-05-22.

[31] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural network,” in Proc. 13th Int. Conf. Art. Intell. Stat. (AISTATS), pp. 249-256, 2010.

[32] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” COURSERA: Neural Net. Mach. Learn, vol. 4, no. 2, pp. 26-31, 2012.