

# IMVIP 2023

25<sup>th</sup> Irish Machine Vision and Image Processing Conference  
30<sup>th</sup> August - 1<sup>st</sup> September, 2023  
Galway, Ireland



Sponsored by:



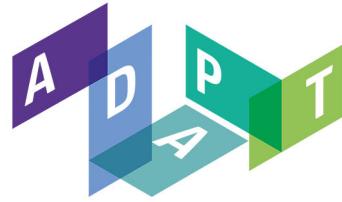
OLSCOIL NA  
GAILLIMHE  
UNIVERSITY  
OF GALWAY



Irish Pattern  
Recognition  
and Classification  
Society

XPERI

IEEE



Engaging Content  
Engaging People

Published by the Irish Pattern Recognition & Classification Society

[iprcs.org](http://iprcs.org)

ISBN 978-0-9934207-8-8

©2023

This work is distributed free of charge by the Irish Pattern Recognition & Classification Society on behalf of the Irish Machine Vision & Image Processing Conference, and the contributing authors to this conference. Both organisers and authors own the rights of their contribution to this book.

## Welcome Note from the Chair

On behalf of the Organizing Committees, we warmly welcome you to the 25<sup>th</sup> Irish Machine Vision and Image Processing Conference (IMVIP) held this year in Galway in the West of Ireland. It's our honor to celebrate this milestone event, which not only marks a quarter-century of contributions to research and societal impact but also provides a platform to discuss future advancements. The main event starts with a keynote from the CTO of Xperi Corporation, focusing on state-of-the-art developments in Driver Monitoring Systems and Next-Generation Perceptual AI.

In addition to our rich main program featuring eight themed sessions and over 20 poster presentations, this year's anniversary edition includes an extra day dedicated to Special Sessions and Training activities. These will explore cutting-edge topics like Data Generation and Augmentation, Deep Learning in Medical Image Processing, and Immersive Technologies for Health & User Experience. We'll cap off this additional day with a session discussing the impact of recent Data Privacy laws on Machine Vision, rounded out by an expert panel focused on Machine Vision Privacy Challenges.

To put a conference of this magnitude together is not a small task. To that end, we want to thank Professors Sonya Coleman and Richard Gault for providing their wisdom and guidance from their work on IMVIP 2022; Dr. Muhammad Ali Farooq (Program Chair) for his tireless efforts in organising the review process and all sessions and tracks; Dr. Hossein Javidnia (Publicity Chair) for his many contributions including support in setting up the registration process; Dr. Mariam Yiwere (Conference Website and Registrations); Dr. Claudia Costache (Finance Chair, Local Organisation & Logistics). We thank all members of the program committee for their work and contributions during the peer review process. In addition, we would like to offer special thanks to Xperi corporation and ADAPT research centre for their generous financial support and to the Schools of Computer Science and Engineering at University of Galway for supporting our extra day of Training and Special Sessions. Lastly, we would like to thank all of the conference participants for their contributions which are the foundation of this conference.

Prof. Peter Corcoran  
University of Galway, Ireland  
August 2023

# Committee Members

## Programme Chair

Prof. Peter Corcoran (University of Galway)

## Organizing Committee

Prof. Michael Schukat (University of Galway)

Prof. Niall Murray (Technical University of the Shannon)

Dr. Muhammad Ali Farooq (University of Galway)

Dr. Hossein Javidnia (Dublin City University)

Dr. Ihsan Ullah (University of Galway)

Dr. Claudia Costache (University of Galway)

Dr. Mariam Yiwere (University of Galway)

## Programme Committee

Alan Smeaton, Insight Centre for Data Analytics, Dublin City University

Aryan Singh, University of Limerick

Brian Mac Namee, University College Dublin

Bryan Gardiner, Ulster University

Cem Direkoglu, University of Limerick

Dane Brown, Rhodes University

Darragh Lydon, Queen's University Belfast

Dermot Kerr, Ulster University

Donald Bailey, Massey University

Francesco Bianconi, University of Perugia

Gabriel Costache, Xperi

Ganesh Sistu, Valeo Vision Systems

Gregory Balogh, Queen's University Belfast

Hossein Javidnia, Dublin City University

Huiyu Zheng, University of Ulster

Ihsan Ullah, University of Galway

James McDermott, University of Galway

Jane Courtney, TU Dublin

Jesus Martinez del Rincon, Queen's University Belfast

John McDonald, Maynooth University

Joseph Lemley, Xperi Corporation, Galway, Ireland

Kathleen Curran, University College Dublin

Kevin McGuinness, Dublin City University

Liam Kilmartin, University of Galway

Malika Bendechache, University of Galway

Mariam Yiwere, University of Galway

Martin Boyer, AIT Austrian Institute of Technology GmbH

Mehdi Sefidgar Dilmaghani, University of Galway

Michael Schukat, University of Galway

Michela Lorandi, Dublin City University

Muhammad Fahim, Queen's University Belfast

Muhammad Ali Farooq, University of Galway

Niall McLaughlin, Queen's University Belfast

Noel Connor, Dublin City University

Orla Sealy Phelan, University of Galway

Paul Whelan, Dublin City University

Paul McKevitt, Ulster University

Paul Cuffe, University College Dublin

Peter Corcoran, University of Galway  
Richard Gault, Queen's University Belfast  
Rishabh Jain, University of Galway  
Robert Ross, Technological University Dublin  
Rozenn Dahyot, Maynooth University  
Sally McClean, Ulster University  
Sean Mullery, IT Sligo  
Seyedalireza Khoshirat, University of Delaware  
Shubhajit Basak, University of Galway  
Sonya Coleman, University of Ulster  
Soumyabrata Dev, University College Dublin  
Stephen Foy, Technological University Dublin  
Sushil Sharma, University of Limerick  
Suzanne Little, Dublin City University  
Wang Yao, University of Galway  
Waseem Shariff, University of Galway

## Keynote Speakers

**Dr. Petronel Bigioi**  
*CTO, XPERI*



Petronel Bigioi serves as chief technology officer. Based in Ireland, he is responsible for leading the engineering team focused on developing image processing and audio solutions for the home, automotive and mobile markets. Petronel has more than 250 granted and published U.S. and international patents to date. He is an Institute of Electrical and Electronics Engineer fellow with more than 20 years of experience in the digital still camera and mobile phone industries, working in both signal processing and connectivity. His work has been recognized by the Romanian Academy of Science's Gheorghe Cartianu Award. A co-founder of several successful companies including FotoNation, which was later acquired by Xperi, he is also a pioneer of digital camera connectivity and is a co-author of the picture transfer protocol (PTP) and PTP-over-IP networks communication standards. Petronel obtained his Ph.D. in electronics, master's degree in application-specific integrated circuits design and bachelor's degree in electronics engineering from Transilvania University in Brasov. Petronel also holds a master's degree in networks and communications from the National University of Ireland, Galway.

### Driver Monitoring Systems and Next-Generation Perceptual AI

Monitoring drivers and passengers inside of vehicles is an increasingly critical capability. For example, driver monitoring is required in order for cars to obtain a top safety rating from NCAP. This presentation introduces XPERI's driver and in-cabin monitoring solutions and examines real-world use-cases in which these solutions are being deployed.

Dr. Bigioi illustrates the evolution of these technologies as they have been used in conventional cars, as they are increasingly being used in cars with partial self-driving capability, and how they are likely to be used in fully automated vehicles. This goes well beyond driver monitoring to include new types of safety features as well as non-safety uses such as entertainment, personalization, human-machine interfaces and even monitoring occupant health.

# Keynote Speakers

## Prof. Peter Corcoran

*University of Galway*



Prof. Corcoran is a 2021 SFI ADAPT-2 PI specializing in Edge-AI & Computer Vision. Other recent research activity includes : 2015-2019, SFI industry/academic partnership with Xperi generated more than 50 academic publications, 15 patent filings, and graduated 8 PhDs for public investment of 730k; 2019, successful participation in ECSEL Helius project, lead PI an NUIG on SFI Center for Research Training (D-real); 2020 successful participation in DTIF DAVID project to develop low-power smart-toy platform. He is recognized by Guide2Research as #1 researcher in the ICT/Electronic Engineering field in Ireland (2020) and is an IEEE Fellow (2010); More than 500 technical publications, 100+ peer-reviewed journal papers, 150+ International peer-reviewed conference papers; Co-inventor on 400+ granted US patents, 100+ granted European. University Professor & Lecturer for 30+ years; Member of IEEE Consumer Electronics Society 25+ years.

## Fake Children: Why we need them and how to make them

Researchers working with human-centric Machine Vision applications have found GDPR to be a huge challenge. Many of today's Machine Vision systems rely on neural network models and require large training datasets for optimal performance. But what do you do when your application is for a Smart-Toy and you need data from a vulnerable population, such as young children, in order to train your Edge-AI Machine Vision system? GDPR imposes many complexities in collecting, managing and processing data from real children.

Fortunately it is now quite feasible to leverage state-of-the-art GAN and other generative neural technologies to build data samples at scale. In this talk we'll get some insights into the power and potential of today's neural technologies to build a gender balanced dataset of child data, including controllable facial expressions, age variations, facial pose and even speech-driven animations with photo-realistic lip-synch. Learn more from our second keynote presented by Prof. Peter Corcoran and Dr. Muhammad Ali Farooq from the University of Galway.

## Table of Content

<b>Welcome Note from the Chair</b>	ii
<b>Committee Members</b>	iii
<b>Keynote Speaker: Dr. Petronel Bigioi</b>	v
<b>Keynote Speaker: Prof. Peter Corcoran</b>	vi
<b>Multimodal Data: Innovations in Data Generation and Augmentation</b>	
<b>1 Adapting the CycleGAN Architecture for Text Style Transfer</b>	1
<i>Michela Lorandi, Maram A. Mohamed, and Kevin McGuinness</i>	
<b>2 Assessment of Synthetic Turfgrass Dataset Generation for Divot Detection</b>	9
<i>Stephen Foy and Simon McLoughlin</i>	
<b>3 A Comparative Study of Image-to-Image Translation Using GANs for Synthetic Child Race Data</b>	17
<i>Wang Yao, Muhammad Ali Farooq, Joseph Lemley, and Peter Corcoran</i>	
<b>4 AudRandAug: Random Image Augmentations for Audio Classification</b>	25
<i>Teerath Kumar, Muhammad Turab, Alessandra Mileo, Malika Bendechache, and Takfarinas Saber</i>	
<b>Deep Learning in Medical Image Processing</b>	
<b>5 Compact &amp; Capable: Harnessing Graph Neural Networks and Edge Convolution for Medical Image Classification</b>	34
<i>Aryan Singh, Pepijn Van de Ven, Ciarán Eising, and Patrick Denny</i>	
<b>6 Assessing Intra-class Diversity and Quality of Synthetically Generated Images in a Biomedical and Non-biomedical Setting</b>	42
<i>Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairí O'Reilly</i>	
<b>7 Facial Camera-Based Heart Rate Estimation Using r-PPG Convolutional Neural Networks</b>	50
<i>Mohamed Moustafa, Joseph Lemley, and Peter Corcoran</i>	
<b>Immersive Technologies for Health, Rehabilitation, and User Experience</b>	
<b>8 A QoE and Visual Attention Evaluation on the Influence of Spatial Audio in 360° videos</b>	59
<i>Amit Hirway, Yuansong Qiao, Niall Murray</i>	
<b>9 An Expert Evaluation of a VR Intervention for Children with ASD</b>	67
<i>Yujing Zhang, Conor Keighrey, Niall Murray</i>	
<b>10 A Quality of Experience Evaluation of an Interactive Multisensory 2.5D Virtual Reality Art Exhibit</b>	75
<i>Chen Chen, Niall Murray, Conor Keighrey</i>	

<b>11</b>	<b>Virtual Rehabilitation for Patients with Osteoporosis: Translating Physiotherapy Exercises into Exergames</b>	81
	<i>Eléa Thuilier, John Carey, Bryan Whelan, John Dingliana, Mary Dempsey, Shane Biggins, Kenzo Thuilier, Attracta Brennan</i>	
<b>12</b>	<b>Designing the VR Probe: An Introductory Application to Virtual Reality for People Living with Dementia (PLWD)</b>	89
	<i>Gearóid Reilly, Gabriel-Miro Muntean, Aisling Flynn, Attracta Brennan, Sam Redfern</i>	

### Explainable AI and DeepFakes

<b>13</b>	<b>Evaluation of Explainable AI Localisation Performance Using Relevance F-Score</b>	96
	<i>Gregory Balogh, Niall McLaughlin, Austen Rainer</i>	
<b>14</b>	<b>Sampling Matters in Explanations: Towards Trustworthy Attribution Analysis Building Block in Visual Models through Maximizing Explanation Certainty</b>	104
	<i>Jiaolin Luo (Róisín), James McDermott, and Colm O'Riordan</i>	
<b>15</b>	<b>Learning from Exemplary Explanations</b>	112
	<i>Misgina Tsighe Hagos, Kathleen M. Curran, and Brian Mac Namee</i>	
<b>16</b>	<b>DF-Net: The Digital Forensics Network for Image Forgery Detection</b>	120
	<i>David Fischinger and Martin Boyer</i>	
<b>17</b>	<b>DF2023: The Digital Forensics 2023 Dataset for Image Forgery Detection</b>	128
	<i>David Fischinger and Martin Boyer</i>	

### Humans, Facial, Gesture, and Action Analysis

<b>18</b>	<b>Development of a Classification-based Eye Gaze estimation technique using an Integrated Laptop Camera: two models are better than one</b>	136
	<i>Jack Cribbin, Charles Markham</i>	
<b>19</b>	<b>A lightweight 3D dense facial landmark estimation model from position map data</b>	143
	<i>Shubhajit Basak, Sathish Mangapuram, Gabriel Costache, Rachel McDonnell, and Michael Schukat</i>	

### Vision-based Object Detection & Tracking

<b>20</b>	<b>YUDO: YOLO for Uniform Directed Object Detection</b>	151
	<i>Dorđe Nedeljković</i>	
<b>21</b>	<b>Accurate object detection using the sensor fusion of an event-based and a frame-based camera</b>	159
	<i>Orla Sealy Phelan, Dara Molloy, Roshan George, Edward Jones, Martin Glavin, Brian Deegan</i>	
<b>22</b>	<b>A Comparative Analysis of Deep Learning Mobile Networks on Marine Vessel Detection</b>	167
	<i>Dipak G Sharma, Michael O'Neill</i>	
<b>23</b>	<b>Towards the Use of Computer Vision Techniques on Streetscape Imagery to Empower Citizens in the Planning Enforcement Process</b>	175
	<i>Sam Lynch and Paul Cuffe</i>	
<b>24</b>	<b>Feature Based Approaches for Homography Estimation</b>	183
	<i>Samuel Venezia, Sonya Coleman, Dermot Kerr, and John Fegan</i>	

<b>25</b>	<b>Uncompromising Operator Safety: A Standalone Device Approach for Threat Immunity and Malfunction Prevention through Visual Cognition</b>	191
	<i>Mihai Penica, Eoin O'Connell, Reenu Mohandas, William O'Brien, Martin Hayes</i>	
<b>26</b>	<b>YOLOatr : Deep Learning Based Automatic Target Detection and Localization in Thermal Infrared Imagery</b>	199
	<i>Aon Safdar, Usman Akram, Waseem Anwar, Basit Malik1, Mian Ibad Ali</i>	

### Applications in Transportation

<b>27</b>	<b>Decisive Data using Multi-Modality Optical Sensors for Advanced Vehicular Systems</b>	207
	<i>Muhammad Ali Farooq, Waseem Shariff, Mehdi Sefidgar Dilmaghani, Wang Yao, Moazam Soomro, and Peter Corcoran</i>	
<b>28</b>	<b>Navigating Uncertainty: The Role of Short-Term Trajectory Prediction in Autonomous Vehicle Safety</b>	215
	<i>Sushil Sharma, Ganesh Sistu, Lucie Yahiaoui, Arindam Das, Mark Halton and Ciarán Eising</i>	
<b>29</b>	<b>Aerially Determined Dynamic Environment Mapping for Enhanced Road Vehicle Awareness</b>	223
	<i>Brendan Halligan, Dara Molloy, Edward Jones, Brian Deegan, Martin Glavin and Liam Kilmartin</i>	

### Medical Imaging, Healthcare & Assistive Technologies

<b>30</b>	<b>Empowering Visually Impaired Individuals: A Novel Use of Apple Live Photos and Android Motion Photos</b>	231
	<i>Seyedalireza Khoshirat and Chandra Kambhamettu</i>	
<b>31</b>	<b>Saliency Maps as an Explainable AI Method in Medical Imaging: A Case Study on Brain Tumor Classification</b>	239
	<i>Ayse Keles, Ozan Akcay, Halil Kul, and Malika Bendechache</i>	

### Transformer Based Methods

<b>32</b>	<b>Domain Generalisation with Bidirectional Encoder Representations from Vision Transformers</b>	247
	<i>Hamza Riaz and Alan F. Smeaton</i>	
<b>33</b>	<b>Dynamic Cost Volumes with Scalable Transformer Architecture for Optical Flow</b>	251
	<i>Vemburaj Yadav, Alain Pagani, Didier Stricker</i>	

### New Datasets and Misc. Applications

<b>34</b>	<b>DeepSky dataset: A new benchmark for ground-based cloud classification using all-sky images</b>	259
	<i>Dimitrios Tsourounis, Dimitris Kastaniotis, Panagiotis Tzoumanikas, George Andrianakos, Orestis Panagopoulos, Andreas Kazantzidis, Christos Theocharatos, and George Economou</i>	
<b>35</b>	<b>Haptic Gloves (SM-EXO) for Multi-Users in Pick-and-Place Collaborative Robot Simulated Environment</b>	267
	<i>Rupal Srivastava, Eber Lawrence Souza Gouveia, Niall Murray, Declan Devine</i>	
<b>36</b>	<b>Do the Frankenstein, or how to achieve better out-of-distribution performance with manifold mixing model soups</b>	274
	<i>Hannes Fassold</i>	

- 37 Plant Disease Detection on Multispectral Images using Vision Transformers** 282  
*Dane Brown and Malithi De Silva*

**Posters**

- 38 Quality of Multimedia Experience Prediction using Peripheral Physiological Signals** 290  
*Sowmya Vijayakumar, Ronan Flynn, Peter Corcoran, and Niall Murray*
- 39 Physiological Synchrony: A Novel Approach to Evaluating User Quality of Experience in Collaborative Distributed Virtual Reality Environments** 294  
*Bhagyabati Moharana, Dr. Conor Keighrey, Dr. Niall Murray*
- 40 Will your Doorbell Camera still recognize you as you grow old?** 298  
*Wang Yao, Muhammad Ali Farooq, Joseph Lemley, and Peter Corcoran*
- 41 Defect Classification in Additive Manufacturing Using CNN-Based Vision Processing** 302  
*Xiao Liu, Alessandra Mileo and Alan F. Smeaton*
- 42 The Impact of Glare on End of Production Line Camera Calibration Algorithms: A Brief Analysis** 306  
*Payal Bhattacherjee, Anbucchezhiyan Selvaraju, Sudarshan Paul, Arindam Das, Ishan Vermani*
- 43 Automatic Archery Scoring System Using Deep Learning and Image Processing** 310  
*Haozhe Ma, Michael G. Madden*
- 44 Deep Learning enabled Computer Vision in Remanufacturing and Refurbishment applications: Defect Detection and Grading for Smart Phones** 314  
*Reenu Mohandas, Martin Hayes, Colin Fitzpatrick, Mark Southern*
- 45 Improving GMM registration with class encoding** 318  
*Solmaz Panahi, Jeremy Chopin, Matej Ulicny & Rozenn Dahyot*
- 46 Calculating Breathing Rates from Remote PPG Signals Using Machine Learning Methods** 322  
*Adara Andonie, Timothy Hanley, Dara Golden, Robyn Maxwell, Joe Lemley, and Ashkan Parsi*
- 47 The Xperi 3D Full Body Photogrammetric Scanner** 326  
*Arpad Zoldi, Stefan Bigioi, Jakub Pawelec, Padraic Toomey, Victor Vlad and, Bogdan Basuc*
- 48 An Introduction to the Xperi Driving Simulation Environment: Hardware, Software and Data Acquisition** 330  
*Rachel Corcoran, Paul Kielty, Padraic Toomey, and Joe Lemley*
- 49 Neuromorphic Seatbelt State Detection for In-Cabin Monitoring with Event Cameras** 334  
*Paul Kielty, Cian Ryan, Mehdi Sefidgar Dilmaghani, Waseem Shariff, Joe Lemley, and Peter Corcoran*
- 50 Sign Language Recognition: Can depth cameras be used to correct Mediapipe errors?** 338  
*Frank Fowley and Anthony Ventresque*
- 51 Investigating the Interplay Between Cervical Spine Sagittal Balance and Lower Back Pain Using Computational Biomechanics and Biomedical Imaging** 342  
*Katherine Nery Rios Peralta, David MacManus, Michaela Davis, Kathleen M. Curran*
- 52 Towards a performance analysis on pre-trained Visual Question Answering models for autonomous driving** 346  
*Kaavya Rekanar, Ciarán Eising, Ganesh Sistu, Martin Hayes*
- 53 Evaluate Fine-tuning Strategies for Fetal Head Ultrasound Image Segmentation with U-Net** 350  
*Fangyijie Wang, Guénolé Silvestre, Kathleen M. Curran*

<b>54</b>	<b>Hardware Accelerators in Autonomous Driving</b>	354
	<i>Ken Power, Shailendra Deva, Ting Wang, Julius Li, Ciarán Eising</i>	
<b>55</b>	<b>Self-Supervised Online Camera Calibration for Autonomous Driving Applications</b>	358
	<i>Ciarán Hogan, Ganesh Sistu, Ciarán Eising</i>	
<b>56</b>	<b>An Ensemble Deep Learning Approach for COVID-19 Severity Prediction Using Chest CT Scans</b>	362
	<i>Sidra Aleem, Mayug Maniparambil, Suzanne Little, Noel O'Connor and Kevin McGuinness</i>	
<b>57</b>	<b>Explaining decisions of a light weight deep network model for Coronary Artery Disease classification in Magnetic Resonance Imaging</b>	366
	<i>Aaleen Khalid, Talha Iqbal, and Ihsan Ullah</i>	
<b>58</b>	<b>Non-Contact Breathing Rate Detection Using Optical Flow</b>	370
	<i>Robyn Maxwell, Timothy Hanley, Dara Golden, Adara Andonie, Joseph Lemley, and Ashkan Parsi</i>	
<b>59</b>	<b>Non-Contact NIR PPG Sensing through Large Sequence Signal Regression</b>	374
	<i>Timothy Hanley, Dara Golden, Robyn Maxwell, Joseph Lemley, and Ashkan Parsi</i>	
<b>60</b>	<b>Spurious Correlation Mitigation in CXR Images via Reinforcement learning and Self-Supervision</b>	378
	<i>Weichen Huang and Kathleen M. Curran</i>	

# Adapting the CycleGAN Architecture for Text Style Transfer

Michela Lorandi, Maram A.Mohamed, and Kevin McGuinness

*Dublin City University*

## Abstract

Text Style Transfer, the process of transforming text from one style to another, has gained significant attention in recent years due to its potential applications in various Natural Language Processing (NLP) tasks. In this paper, we present a novel approach for Text Style Transfer using a Cycle Generative Adversarial Network (CycleGAN). Our method utilizes the adversarial training framework of CycleGAN to learn the mapping between different text styles in an unsupervised manner, without the need for paired data. By leveraging the cycle consistency loss, our model is able to simultaneously learn style transfer mappings in both directions, allowing for bidirectional style transfer. We conduct experiments on the Yelp dataset to evaluate the effectiveness of our approach. Our results illustrate that our proposed TextCycleGAN achieves reasonable performance in terms of style transfer accuracy and fluency considering the simple architecture adopted in both generators and discriminators, while also providing bidirectional transfer capabilities (negative-positive and positive-negative).

**Keywords:** CycleGAN, Text Style Transfer, Text Generation

## 1 Introduction

CycleGAN is a type of generative adversarial network (GAN) that can be used for image-to-image translation tasks [Zhu et al., 2017]. CycleGAN is notable for being able to learn mappings between two domains without requiring paired examples of those domains during training. The basic idea behind CycleGAN is that it learns two mappings: one from domain  $A$  to domain  $B$  and another from domain  $B$  to domain  $A$ . These mappings are learned simultaneously by training two GANs, with each GAN learning to generate images in one of the two domains. The two GANs are trained in an adversarial manner, with one generator trying to generate realistic images in its domain, while the discriminator tries to discriminate between the generated images and the real images from its domain. One of the key benefits of CycleGAN is that it can learn to translate between domains without relying on paired examples, which can be difficult to obtain or create. It instead relies on the assumption that if an image in domain  $A$  can be translated to a realistic image in domain  $B$  and then translated back to a realistic image in domain  $A$ , the mapping is successful. This process is referred to as cycle consistency. CycleGAN has been used for a variety of image-to-image translation tasks, including style transfer, colorization, and image synthesis.

Language is a fundamental tool for human communication and plays a vital role in our ability to convey ideas, emotions, and experiences. However, communicating across different languages or styles can be challenging, as words, phrases, and even intonation can carry different meanings or cultural connotations. Accurate translation or transfer of meaning is essential to effective communication, but it requires a deep understanding of both the source and target languages or styles, as well as cultural and contextual factors. Despite the advancements in technology and the availability of machine translation tools, the nuances of human language and communication continue to pose challenges for translation and transfer of meaning, making it a complex and ongoing area of research and practice.

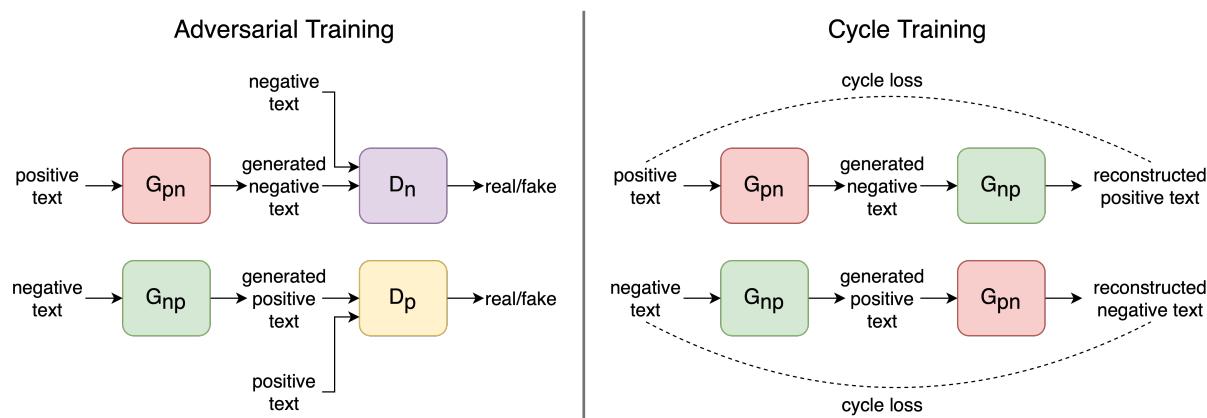
While CycleGAN was originally developed for image-to-image translation, its underlying principles can be extended to other domains, including text. Applying CycleGAN to text datasets could enable text-to-text

translation or style transfer, allowing for the generation of new text that preserves the content of the original text while adopting the style of a different author or language.

The potential of CycleGAN for text translation or style transfer lies in its ability to learn mappings between different domains without requiring paired examples. This is particularly useful for text datasets, where obtaining paired examples for training can be difficult or time-consuming. Instead, CycleGAN can use unpaired datasets in two different languages, for example, to learn the relationship between them and generate new text that preserves the meaning of the original while adopting the style of the other language. There have been some recent developments in the application of CycleGAN to text datasets, with promising results. For example, researchers have used CycleGAN to perform style transfer between different authors of poetry, generating new poems in the style of a different author while preserving the meaning and structure of the original [Vecchi et al., 2022]. While the application of CycleGAN to text datasets is still a relatively new field of research, its potential is significant. Text-to-text translation or style transfer could be useful in a variety of applications, including literature, marketing, and advertising. However, as with any machine learning application, it is important to ensure that the generated text is accurate, understandable, and free of bias.

The main contribution of this research work is that we investigate the impact of using CycleGAN to perform sentiment style transfer, for instance, going from positive to negative sentiment. For this investigation, we use the Yelp dataset.

This work is organised as follows: Section 1 introduces CycleGAN and the motivation behind this research work, while in Section 2, we discuss related works. In Section 3, we discuss the dataset that we used and the model details. We share our experimental setup and report the experimental results in Section 4. We conclude the work in Section 5 by discussing our findings and possible future work.



**Figure 1: Style Transfer CycleGAN.** On the **left**, the adversarial training of the two flows: positive to negative text and negative to positive text. On the **right**, the cycle training of the two flows. First, the positive text is converted into negative text, which is used to reconstruct the positive text. Positive and reconstructed positive texts are compared to compute the cycle loss. The same is done on the negative flow.<sup>1</sup>

## 2 Related Work

Text style transfer is a challenging problem in natural language processing, where the goal is to generate text in a target style while preserving the content and meaning of the original text. Different approaches have been implemented to tackle this problem, such as [Tikhonov et al., 2019], which explores different methods including retraining a pre-trained language model and adapting a pre-trained model using a small amount of labeled data. [Wang et al., 2019] proposes a method for text style transfer that allows for control over specific

<sup>1</sup>All diagrams have been created by the authors.

attributes of the transferred text, such as sentiment or formality. These works demonstrate the wide range of approaches being explored in this area and highlight the ongoing challenges in achieving high-quality text style transfer without sacrificing content and meaning.

In addition, [Shen et al., 2017] propose to use non-parallel texts assuming there is a shared latent content distribution across different corpora and propose to align latent content distributions to perform style transfer. The idea is to have an encoder that maps the input text into its content latent representation and a generator that recreates the original text using the learned content latent representation combined with the original style. Similarly, [Luo et al., 2019] use non-parallel data proposing a cycle reinforcement learning algorithm to enable fine-grained control text sentiment transfer by incorporating the intensity of the sentiment in the generation process. The cycle RL is composed of two rewards: a sentiment reward and a content reward. The sentiment reward evaluates how well the generated text matches the target sentiment, while the content reward is based on the idea that, if the model performs well, it is easy to reconstruct the original input, thus enabling a cycle RL.

Unlike these existing approaches, we propose to implement a CycleGAN architecture for text style transfer obtaining two specialised generators for sentiment style transfer. Inspired by the application of the CycleGAN architecture in computer vision, we apply the same architecture on text, adopting two generators and two discriminators that are specialised in the sentiment domain.

### 3 Methodology

#### 3.1 TextCycleGAN

For the style transfer task, we propose TextCycleGAN, which is built using two different generators and discriminators. The styles in sentiment transfer are defined as positive and negative, therefore TextCycleGAN is composed of two text GANs: one is going from positive to negative with a negative discriminator, which is a real/fake classifier for the negative sentence, and another negative-to-positive generator with a positive discriminator as a real/fake classifier for the positive sentence.

##### 3.1.1 Adversarial Training

TextCycleGAN (Figure 2) is composed of a generator, which is an encoder-decoder network with LSTM layers, and a discriminator, which is a binary LSTM classifier. In the generator, the input sentence is encoded to find its hidden representation. The obtained hidden representation is passed to the decoder together with the start-of-sentence token (SOS) in order to start generating the next token. At every step, we feed the previous token and the previously obtained hidden representation to the decoder in order to generate the next token.

The generator has the objective of transferring the sentiment of the input text into the target sentiment, while the discriminator has to identify whether the given text is real or fake (Figure 1 left). More formally, the generator  $G_{pn}$  aims to minimize the adversarial loss:

$$\mathcal{L}_{\text{GAN}}(G_{pn}, D_n) = \mathbb{E}_{n \sim N}[\log D_n(n)] + \mathbb{E}_{p \sim P}[\log(1 - D_n(G_{pn}(p)))] \quad (1)$$

while the discriminator  $D_n$  aims to maximize it. In the positive to negative, we have the generator  $G_{pn}$  that takes in input a positive text and transfers the content into negative text. The generated negative text is fed to the negative discriminator  $D_n$  together with real negative text so that the discriminator can predict whether each text is real or generated.

Since the softmax operation in the generator is not differentiable with respect to the discriminator, we explore two ways to solve the issue. Inspired by SeqGAN [Yu et al., 2017], we apply a pseudo-loss in which we compute policy gradient, while inspired by [Kusner and Hernández-Lobato, 2016] we apply the Gumbel softmax trick [Huang et al., 2021]. Finally, the discriminator loss is a binary cross-entropy loss.

### 3.1.2 Cycle Loss

The idea of CycleGAN is that the model should be able to reconstruct the input text using the generated text (Figure 1 right). Going from positive to negative, we have the generator  $G_{pn}$  that takes in input a positive text and transfers the content into negative text. The generated negative text is fed to the generator  $G_{np}$  to generate the reconstruction of the positive text. The same process is done in the negative-to-positive path. The cycle loss is computed as the cross entropy loss between the reconstructed positive text and the original positive text plus the cross entropy loss between the reconstructed negative text and the original negative text. The cycle loss for the CycleGAN is defined as:

$$\mathcal{L}_{\text{cyc}}(G_{pn}, G_{np}) = \mathbb{E}_{p \sim p_{\text{data}}(p)} [\|G_{np}(G_{pn}(p)) - p\|_1] + \mathbb{E}_{n \sim p_{\text{data}}(n)} [\|G_{pn}(G_{np}(n)) - n\|_1], \quad (2)$$

where  $G_{pn}$  and  $G_{np}$  are the generators for mapping the positive sentiment domain to the negative sentiment domain, and mapping the negative sentiment domain to the positive sentiment domain, respectively. The cycle loss measures the difference between the reconstructed text and the original text.

The final loss is the combination of adversarial training, i.e. positive-to-negative adversarial loss and negative-to-positive adversarial loss, and cycle training, given by:

$$\mathcal{L}(G_{pn}, G_{np}, D_p, D_n) = \mathcal{L}_{\text{GAN}}(G_{pn}, D_n) + \mathcal{L}_{\text{GAN}}(G_{np}, D_p) + \lambda \mathcal{L}_{\text{cyc}}(G_{pn}, G_{np}). \quad (3)$$

where  $D_p$  and  $D_n$  are the positive and negative sentiment discriminators.

### 3.1.3 Generator Pretraining

Since our objective is to modify the sentiment of the text while maintaining the original content, we pre-train the generator to reconstruct the input text. In this way, the generator should be able to learn to maintain the original content while modifying the style words. To achieve this objective, we use the cross entropy between input text and generated text.

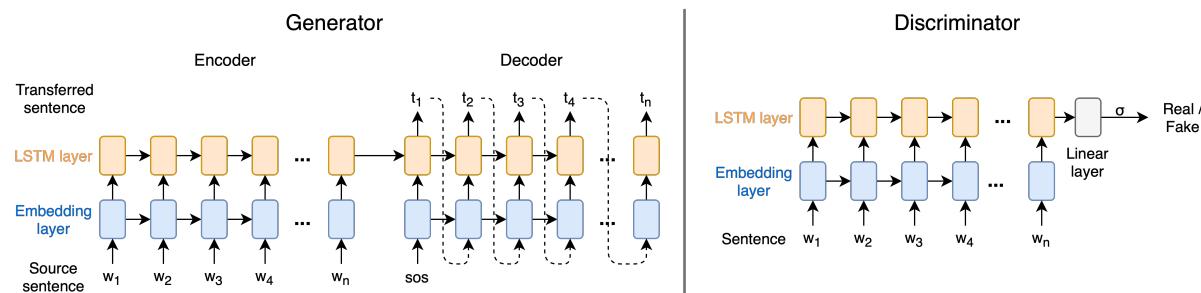


Figure 2: **Generator and discriminator architectures.** **Left:** architecture of the generator, i.e. an encoder-decoder network with LSTM layers. **Right:** architecture of the discriminator, i.e. a binary LSTM classifier.<sup>1</sup>

## 4 Experiments and Results

### 4.1 Dataset

The experiments for this work have been conducted using the Yelp dataset [Asghar, 2016]. The Yelp dataset is a large collection of customer reviews and ratings for businesses, including restaurants, hotels, and various services. It consists of reviews, each containing a rating, text description, and other metadata such as the date of the review and the business category. This dataset has been widely used in NLP research, particularly for sentiment analysis, text classification, and text generation tasks. In the context of text style transfer using CycleGAN, the Yelp dataset is used to train a model that can transfer the writing style of one group of reviews to another.

Table 1: Comparison between the proposed model and baselines. Accuracy (%) on a pre-trained classifier), Fluency (perplexity using GPT-2), BLEU score between input and transferred texts, and G-score (using accuracy and BLEU) are reported.

<b>Model</b>		<b>Accuracy ↑</b>	<b>Fluency ↓</b>	<b>BLEU ↑</b>	<b>G-score ↑</b>
CAAE [Shen et al., 2017]		82.7	-	11.2	30.43
ARAE [Zhao et al., 2018]		83.2	-	2.3	13.83
DRLST [John et al., 2019]		91.2	-	7.6	26.33
TextCycleGAN	Policy Gradient (PG)	68.04	2049	32.42	47.07
	PG + more epochs	69.24	1824	28.47	44.4
	Gumbel Softmax (GS) + more epochs	78.85	1920	20	39.71

## 4.2 Experimental setup

**Preprocessing stage** The experiments were conducted using the Yelp dataset by first filtering out the texts that exceed 20 words, meaning that we removed all texts longer than 20 tokens. As a result, we considered 444101 texts in train set, 63483 texts in dev set and 126670 texts in test set.

**Training** The styles in sentiment transfer are defined as positive and negative, therefore TextCycleGAN is composed of 2 text GANs: each of the text generators was built using a 2-layer LSTM encoder and decoder. The discriminator was also built on a LSTM network. We trained our sentiment transfer model with the following hyperparameters setup: maximum sequence length=20, batch size=64, generator pre-train epochs=10, training epochs=25 or 50, optimizer=Adam, learning rate= $2 \times 10^{-4}$ ,  $\beta = 0.5$ , embedding dimension=256, generator hidden dimension=512, discriminator hidden dimension=128, number of layers=2, LSTM dropout=0.5, cycle loss lambda=10.

**Evaluation Protocol** To assess the performance of text style transfer using TextCycleGAN, we use the following evaluation metrics:

- BLEU [Papineni et al., 2002] between input text and transferred text. We measure the semantic preservation between input and output texts to check if the content has been preserved during the style transfer.
- Accuracy of target sentiment. We use a pre-trained binary sentiment classifier<sup>2</sup> to obtain the sentiment of the transferred text and check whether it has been transferred in the correct sentiment or not.
- Fluency. We check whether the generated texts are written in fluent English by computing the perplexity [Jelinek et al., 1977] with GPT-2.
- G-score [Hu et al., 2022]. We compute the geometric mean of BLEU and accuracy, which represent the overall performance of the model combining different evaluation metrics.

**Baselines** We compare our method with three models that implement Implicit Style-Content Disentanglement [Hu et al., 2022]: CAAE, ARAE, and DRLST. CAAE [Shen et al., 2017] model is a model that implicitly disentangles text style trained in an adversarial manner. The model is based on the assumption that different corpora share a latent content distribution and it is possible to align latent distributions to perform text style transfer. ARAE [Zhao et al., 2018] is a language generation model based on adversarial learning with the objective to modify the specific attributes in text. DRLST [John et al., 2019] is an adversarial learning model based on the incorporation of auxiliary multi-tasks for style prediction and adversarial objectives for bag-of-words prediction in order to perform text style transfer.

<sup>2</sup><https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english>

Table 2: Analysis on positive and negative sentiment separately. Accuracy (%) on a pre-trained classifier), Fluency (perplexity using GPT-2), BLEU score between input and transferred texts are reported.

	Positive			Negative		
	Acc ↑	BLEU ↑	Fluency ↓	Acc ↑	BLEU ↑	Fluency ↓
TextCycleGAN - PG	70.74	34.83	1755	65.34	30.01	2343
TextCycleGAN - PG + more epochs	70.28	30.61	1403	68.2	26.34	2218
TextCycleGAN - GS + more epochs	84.80	20.70	1586	72.89	19.31	2254

### 4.3 Results

Table 1 compares our model variantions with the baseline models from [Hu et al., 2022]. We did not use the same sentiment classifier so there may be differences in accuracy due to the usage of a different classifier. First, we observed that the proposed model with Gumbel softmax reaches a moderate good accuracy (78.85%) in transferring the texts in the target sentiment. Furthermore, the proposed model variations show a high BLEU score, which means that they are able to maintain the content of the input text. It also means that in some cases some sentiment words of the original text are maintained instead of changing them or neutral text is generated. For example, the input positive text “*so i liked the service my mom and i received today .*” is transferred to “*i appreciate the service my mom and i received today .*”, in which the sentiment is maintained as positive. Another example is considering the input positive text “*awesome breakfast!*” transferred to “*breakfast!*”. The transferred text is not correctly transferred to negative, but it is translated into a neutral sentiment.

In addition, we decided to separately analyze positive and negative generated texts in order to understand if the trained generators present differences, as shown in Table 2. We observe that the accuracy of negative-to-positive transfer is much higher than positive-to-negative transfer, thus demonstrating that the generated negative texts are not transferred correctly and may need more training to achieve better performance. In addition, looking at BLEU score and Fluency we see that the negative-to-positive generator achieves higher scores in all three settings of the proposed TextCycleGAN.

Table 3 shows some examples in which the model was able to correctly transfer the input text into the target sentiment.

## 5 Conclusions and Future Work

In this work, we have explored adapting CycleGAN to Text Style Transfer using Yelp dataset and modifying the original CycleGAN to align with text dataset resulting in our proposed model TextCycleGAN. The proposed approach does not outperform the compared models in terms of accuracy, but does outperform the others in terms of BLEU score and G-score, while only being trained for a maximum of 50 epochs. Furthermore, the results on the generated positive texts show that the model was able to correctly transfer the sentiment from negative to positive.

In the future, we aim at exploring different improvements, such as adding an attention mechanism or changing the generator architecture to be a Transformer model, for example. Furthermore, we will explore different style attributes, such as formality, and we will further analyse the differences between positive and negative generated texts. In addition to that, diffusion models also offer an alternative approach to text style transfer without relying on explicit paired data. By modelling the data distribution through a diffusion process, these models can be trained on unpaired text samples, allowing for flexible and diverse style transfer. This method leverages the sequential nature of text and the latent space diffusion to generate diverse and high-quality text outputs while avoiding the need for parallel training data.

Table 3: Examples of transferred sentences using the proposed TextCycleGAN.

<b>From negative to positive</b>	<b>From positive to negative</b>
<p><i>Source text</i> it was over cooked , mushy , and surprising , it was cold !</p> <p><i>Gradient Policy</i> it was perfectly cooked , soft , and yes , it was delicious !</p> <p><i>Gradient Policy + more epochs</i> it was perfectly cooked , tender , and fresh , it was delicious !</p> <p><i>Gumbel Softmax</i> it was perfectly cooked , crispy , and spiced , it was delicious !</p>	<p><i>Source text</i> the office is well maintained and clean at all times .</p> <p><i>Gradient Policy</i> the office is poorly maintained and dirty at all times .</p> <p><i>Gradient Policy + more epochs</i> the office is poorly maintained and dirty at all times .</p> <p><i>Gumbel Softmax</i> the office is not managed and clean at all times .</p>
<p><i>Source text</i> this morning however , i had a very bad customer service experience .</p> <p><i>Gradient Policy</i> this morning upon , i had a very good customer service experience .</p> <p><i>Gradient Policy + more epochs</i> this morning however , i had a very good customer service experience .</p> <p><i>Gumbel Softmax</i> this morning yesterday , i had a very good customer service experience .</p>	<p><i>Source text</i> it 's so comfortable , clean , and the air works great .</p> <p><i>Gradient Policy</i> it 's so loud , dirty , and the air smelled horrible .</p> <p><i>Gradient Policy + more epochs</i> it 's old , dirty , and the air looks horrible .</p> <p><i>Gumbel Softmax</i> it 's loud , dirty , and the workers smelled horrible .</p>

## Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) at Dublin City University under Grant No. 18/CRT/6224 and Insight SFI Centre for Data Analytics, Dublin City University under Grant No. SFI/12/RC/2289 P2.

## References

- [Asghar, 2016] Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.
- [Hu et al., 2022] Hu, Z., Lee, R. K.-W., Aggarwal, C. C., and Zhang, A. (2022). Text style transfer: A review and experimental evaluation. *SIGKDD Explor. Newsl.*, 24(1):14–45.
- [Huang et al., 2021] Huang, F., Chen, Z., Wu, C. H., Guo, Q., Zhu, X., and Huang, M. (2021). Nast: A non-autoregressive generator with word alignment for unsupervised text style transfer. *arXiv preprint arXiv:2106.02210*.

- [Jelinek et al., 1977] Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.
- [John et al., 2019] John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. (2019). Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.
- [Kusner and Hernández-Lobato, 2016] Kusner, M. J. and Hernández-Lobato, J. M. (2016). Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.
- [Luo et al., 2019] Luo, F., Li, P., Yang, P., Zhou, J., Tan, Y., Chang, B., Sui, Z., and Sun, X. (2019). Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022, Florence, Italy. Association for Computational Linguistics.
- [Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- [Shen et al., 2017] Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [Tikhonov et al., 2019] Tikhonov, A., Shibaev, V., Nagaev, A., Nugmanova, A., and Yamshchikov, I. (2019). Style transfer for texts: Retrain, report errors, compare with rewrites. pages 3927–3936.
- [Vecchi et al., 2022] Vecchi, L. P., Maffezzolli, E. C., and Paraiso, E. C. (2022). Transferring multiple text styles using cyclegan with supervised style latent space. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.
- [Wang et al., 2019] Wang, K., Hua, H., and Wan, X. (2019). Controllable unsupervised text attribute transfer via editing entangled latent representation. *Advances in Neural Information Processing Systems*, 32.
- [Yu et al., 2017] Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- [Zhao et al., 2018] Zhao, J., Kim, Y., Zhang, K., Rush, A., and LeCun, Y. (2018). Adversarially regularized autoencoders. In *International conference on machine learning*, pages 5902–5911. PMLR.
- [Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

# Assessment of Synthetic Turfgrass Dataset Generation for Divot Detection

Stephen Foy<sup>1,2</sup> and Simon McLoughlin<sup>2</sup>

<sup>1</sup>*School of Engineering, Atlantic Technological University, Galway*

<sup>2</sup>*School of Informatics and Engineering, Technological University Dublin*

## Abstract

Turfgrass divot maintenance is a vital component of sports grounds and golf courses. Precision Turfgrass Management (PTM) uses a variety of smart technologies to manage turfgrass in a sustainable and efficient manner. Deep Learning is an effective way of developing data inspection but requires a large, annotated dataset. This work describes the development of a photo realistic dataset generated using Blender for a turfgrass environment to train deep learning networks to identify anomalies on turfgrass surfaces. Additionally, two colour transfer techniques are analysed, to ensure the synthetic data can closely resemble the real domain data.

**Dataset:** <https://www.zenodo.org/record/8030582>

**Code:** <https://github.com/stevefoy/DeepTurfSynth>

**Keywords:** Imaging, Synthetic data, Turfgrass, Machine Vision, Mobile robotics

## 1 Introduction

Turfgrass is defined as a plant system that remains green for at least six months of the year while maintaining a continuous dense ground cover during its dormancy period [Carlson et al., 2022]. This type of grass system can be found in numerous locations, from domestic gardens to public parks, sports grounds and golf courses. The benefits of turfgrass systems can be described in terms of functional, recreational, and aesthetic components [Beard and Green, 1994]. One commercial use of turfgrass is in the golf course industry. According to the R&A report in 2021, worldwide there are 148 million acres used for golf with 38,081 golf courses spread among 206 of the 251 countries of the world [Klein, 2021].

Precision Turfgrass Management (PTM) is a data driven approach for the maintenance of Turfgrass resources which includes mowing practices, irrigation, nutrient, pest, and cultivation management [Bell et al., 2013]. A divot refers to a small area of turf or grass that has been displaced from the sports ground [Braun et al., 2020]. A survey paper by Carlson looked at the accuracy of sensors to detect turfgrass performance indicators and the stressors before or during visual symptoms occur on plants. Their finding suggest that there is insufficient training and expertise among superintendents on PTM technology, with a recommendation that future research should focus on automating data collection, processing, and interpretation of data [Carlson et al., 2022].

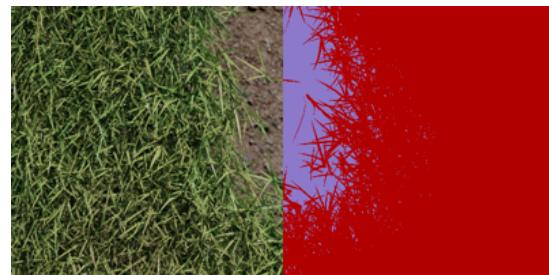


Figure 1: Turfgrass and semantic mask

This paper explores the generation of turfgrass surface data in a 3-dimensional simulation environment called Blender. A photo realistic image dataset with matching synthetic segmentation annotations has been generated and is openly shared in Zenodo. Secondly, several deep learning benchmarks were developed to measure against prior art using a manually annotated test divot dataset. Furthermore, two augmentation and domain information transfer techniques that can increase the performance of this synthetic dataset to a real domain were proposed. The goal of such work is for smart edge learning and deployment, where turfgrass repairs can be autonomously executed on a robotic platform.

## 2 Related work

The variability in turfgrass between locations is considerable with multiple factors affecting it, most notably grass species, plant health/soil fertility and cut height. Divots can be defined as texture anomalies; the topic of image-level anomaly detection is not new and extends to many domains and industries. Image-level anomaly detection is a technique used in computer vision to find anomalies or abnormalities in images [Hendrycks et al., 2019]. In 2016 Ding presented some novel work using vision for the grading of turfgrass. Conventionally turfgrass grading on championship golf courses is usually carried out by experts, whereas this project used an Unmanned Aerial Vehicle (UAV) to capture the image data. An offline computer vision algorithm was used with a machine learning model to calculate a grade of the turfgrass, unfortunately no data was made public to reproduce the results [Ding et al., 2016].

The collection and annotation of image data is an important process, but it is a labour-intensive task particularly for small objects such as plants and foliage. The work of Waldchen presents a summary of publicly available datasets [Waldchen et al., 2018]. In this work it may be observed that most public datasets are for image classification tasks. The Flourish Project under the TA European Horizon 2020 project had objectives to combine Unmanned Aerial Vehicle (UAV) and Unmanned Ground Vehicle (UGV) capabilities for a range of farm management activities. An interesting outcome of the project was a semantically annotated dataset released by the Bonn university Photogrammetry and Robotics group, which used a NDVI camera to help automate the annotation process. This project was focused on sugar beet plants and weeds rather than turfgrass [Liebisch et al., 2016].

In the highly cited work by Skovsen, a novel semi-synthetic dataset called the GrassClover dataset was created with semantic segmentation level annotations. The annotated dataset was developed from several manually collected images. The data has a ground sampling area of 4–8 pixels per millimetre. The generation procedure consisted of cropping plants from images and randomly overlapping these cutouts onto different soil image textures [Skovsen et al., 2019]. The data in the GrassClover dataset was influential to the work in this paper where critical classes (Grass and Soil) could be used to develop baseline algorithm performances. An interesting semi-synthetic (photo realistic) approach was seen in the project titled VisionBlender which was for medical imaging. The project used Blender software as a 3-dimensional (3D) modelling tool for creating scenes and Blender cycles (ray tracing engine) to generate photo realistic data, with accompanying semantic level masks, and depth maps [Cartucho et al., 2021]. An open-source framework for Blender known as the BlenderProc framework, has presented some interesting workflows for creating large-scale urban scenes [Denninger et al., 2023].

## 3 Approach taken

Taking inspiration from the GrassClover dataset and VisionBlender, the objective of this project was to develop a photo realistic dataset with annotations for turfgrass. A 3D turfgrass scene in its simplistic form consisted of a plane of soil with grass objects distributed across it. When rendering images of the dataset at 4k resolution, a full grass scene had in the region of ~200,000 grass leaves in a scene. Blender's new node-based workflow allowed for efficient and randomised rotations and scale and density

of grass assets on the various soil textures. To achieve photo realism both proprietary and custom grass leaf assets were used in the generation. Additionally soil and dirt textures were from 4K to 8K resolution [Price, 2023]. Distribution of this open database keeps within copyright restrictions. The texture incorporated in the rendered scenes used RGB textures, normal maps, displacement maps, and specular maps. The Blender Cycles ray tracing engine was used and resulted in rendering times per image of ~1 minute on a NVIDIA RTX4090 GPU with OptiX enabled. The virtual camera in the scene was setup with a 26mm focal length at 4K resolution to mimic both a modern smart phone and the GrassClover dataset with the objective to validate this data against the state of the art.

An additional consideration in these scenes was the creation of divots. On a golf course the divot width can vary depending on the golf club, the swing technique, and the type of turf. A divot caused by iron shots with standard-sized golf clubs tend to range from around 2.5 to 7.6 centimetres in width [Turgeon, 1991]. Based on this knowledge, a weight map or alternatively known as a vertex weight map were created in blender for the creation of divots. These widths were considered when creating the divots, so nothing was created less than 2.5 cm. These weight textures were incorporated with a geometry node workflow. When divots occur on a grass turf there is an additional deformation to the planer topology, and this was modelled in Blender as a 10mm offset.

### 3.1 Semantic Segmentation data

To generate a semantic segmentation mask in blender, a custom blender composition Nodetree was required. The Nodetree can take each rendered pixel in the rendered image and associate a Pass Index number with a colour to build up a 16bit colour mask (semantic segmentation mask). Objects created in blender are made up of meshes and have a Material Shader associated that give the mesh colour. A feature in blender software is the ability to associate a custom Pass Index number for each Material Shader. The strategy taken in this work was to add a sub-string that was representative of the annotation group, as a Blender object can have multiple Material Shaders. Let us take for example a grass Material Shader named as *Grass\_Rye\_Leaves*. This means the sub-string *Grass* can be found in the Material Shader name. A Blender Python script was developed to associate lists of Material Shader names or sub-strings of the Shader names as class definitions. An additional post processing step using a Python script converted the 16 bit RGB mask into a simple 8 bit monochrome image. These images are more efficient in deep learning frameworks such as PyTorch.

## 4 Datasets

### 4.1 Turfgrass test dataset

Images of divots were collected from a golf course with a smart phone camera (Samsung Galaxy A70); the images had a resolution of 4032x3024 and a focal length of 26mm. The images were captured at a handheld height of one meter over the turfgrass surface with the camera facing parallel. A simple calibration procedure using a small target 56 x 87mm was used to estimates that there was 3.6 pixels per millimetre ground coverage. The manual annotation process was carried out using GIMP imaging software. During the initial research a number of semantic segmentation networks were trained on the GrassClover Dataset, see Figure 3. The positive results

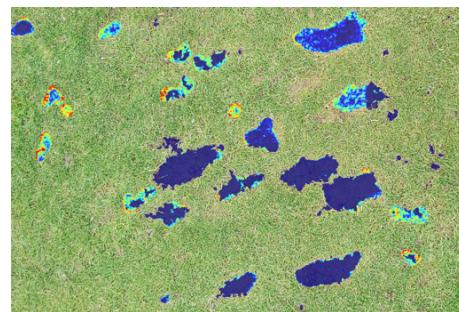


Figure 2: Probability mask in GIMP

achieved were utilized to support the manual annotation process, where the soil class matrix of the Soft-Max layer was imported into GIMP. A coloured visualization of this map can be seen in the Figure 2. A total of 90 images were annotated and saved as a multi-layer TIFF file which could be processed by a Python script to split the images into 400x400 pixel patches to form the test set for a synthetic generated turfgrass. This resulted in a test dataset of 3864 images. Although the results obtained from testing on this dataset may exhibit a slight bias towards the GrassClover dataset, the primary objective was to establish a robust test set for our own synthetic dataset.

## 4.2 Synthetic Turfgrass dataset

The blender virtual camera was rendered at 4032x3024 and placed at 1 metre above the surface resulting in four pixels per millimetre, with no perspective distortion in the lens. The generated data created in blender had matching annotations as detailed in the earlier sections. A Python script split the images into 400x400 pixel patches, resulting in a total of 6510 images, with 10 % used as a validation set during training.

## 5 Training

In this section a quantitative evaluation of the experiments and a benchmark on our dataset using a number of common DL architectures is presented [Wang et al., 2022]. The initial baseline created used the public GrassClover dataset. The Grassclover dataset paper mentions the use of FCN8 and a subset of their dataset with no explicit details. The paper's class intersection over union (IoU) scores are as follows grass, white clover, red clover, weeds and soil, [64.4, 59.5, 72.6, 39.1, 39.0]. Similar results were established using a FCN8 + VGG16 encoder [0.53, 0.59, 0.71, 0.56, 0.64, 0.50] with a PyTorch version on a validation set. The focus was on two classes - grass and soil, in the training of our synthetic dataset.

Table 1: Trained on GrassClover and deployed on a real world turfgrass dataset

Model	Object Class					
	grass	white clover	red clover	weeds	soil	clover other
DeeplabV3+	0.58	0.0	0.0	0.0	0.54	0.0
PSPNet	0.13	0.0	0.0	0.0	0.18	0.0
FCN8	0.49	0.0	0.0	0.0	0.37	0.0
UNet	0.35	0.0	0.0	0.0	0.17	0.0

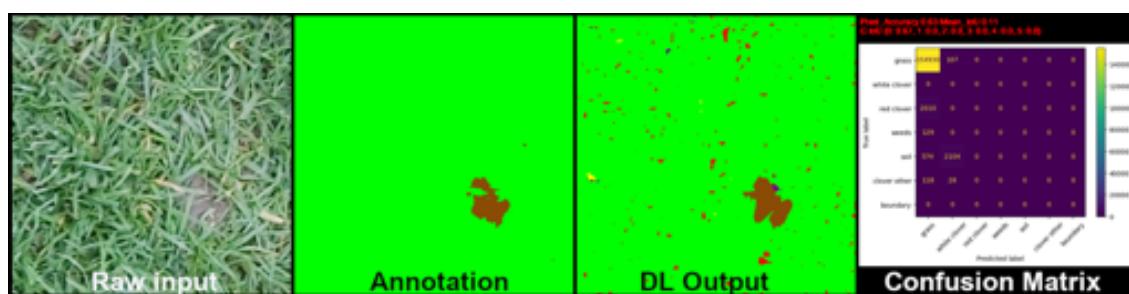


Figure 3: Result on test image using DeepLabV3+, trained on GrassClover dataset

## 5.1 Colour augmentation

We observed that a number of test images encountered performance issues resulting in low class IOU scores during the development work. These images had a slight colour variation and by visualizing the histograms of these problematic images, we could prove that there were auto-white balance (AWB) abnormalities. The AWB feature is active by default on most cameras, giving a colour temperature to the captured raw data and a neutral appearance to the human eye. AWB, or Auto White Balance, strives to emulate the phenomenon of colour constancy observed in the human visual system, particularly under different lighting conditions. Consequently, one can observe the objects in the scene as accurately reflecting the chromatic properties of the captured scene [Buzzelli et al., 2023]. To mitigate against these issues is to capture the image using a manual white balance mode or alternatively capturing in raw mode and correcting after.



Figure 4: Colour transfer, Raw to target synthetic

Trials were conducted using the PyTorch ColourJitter transformation module for brightness (0.1-0.5), contrast (0.1-0.5), saturation (0.1-0.5), and hue (0-0.1). Unfortunately, no significant improvements on class IOU scores were achieved on the problematic test images. The informative work of Afifi and Brown on issues related to AWB was used to provide a colour augmentation strategy to our synthetic dataset [Afifi and Brown, 2019]. The augmented images had a noticeable realistic colour compared to using PyTorch colour jitter augmentations. Another colour augmentation explored in this paper was a colour transfer technique pioneered by Reinhard, which involves the transfer the colour distribution from one image (source) to training images (destination) [Reinhard et al., 2001]. Several real images captured from the smart phone that consisted of just turfgrass were used as the source set and this augmentation strategy was distributed into the training at a rate of 10 % of the dataset, and only on images that had > 95 % pixels in the foliage class.

## 5.2 Training and Evaluation

The segmentation models were trained on synthetic data and tested on the manually annotated divot dataset. Four common DL architectures were selected to establish our baselines: DeeplabV3+ with Resnet101 backbone, PSPNet with ResNet101 backbone, FCN8 with VGG16 backbone and UNet. Networks with Resnet and VGG16 backbones were initialized with PyTorch ImageNet weights. During the training PSPNet and DeepLabV3+ architectures had the encoder and batch normalization layers frozen, resulting in the decoder layers learning the segmentation classes. The mean RGB image and standard deviation were calculated in the synthetic data set and used to normalise FCN8 and UNet models during training and deployment on the test set, while other defaults used were ImageNets. All the findings pre-

sented in this paper were obtained using the PyTorch 2 framework, incorporating the **crossEntropyLoss** function and employing a class weighting strategy. The assigned weights for grass and soil, [1.5, 9.5] respectively, were computed in accordance with the Enet article [Paszke et al., 2016]. One challenge with generating a high-quality annotation is that they can be too accurate with small objects annotations, some of which looks like salt and pepper noise. An optional parameter in the loss function used was **label\_smoothing** at 0.2 which improved generalization. Adam was used as the optimiser on all the training. Table 2 presents the performance results achieved after training these networks only on synthetic data for a duration of 25 epochs, followed by their deployment on the test dataset.

Table 2: Comparisons of semantic segmentation results on turfgrass dataset, (ST: Source to destination image transfer augmentation, AWB: auto-white balance augmentation )

Model	Classes			
	Pixel Accuracy	MIoU	Foliage	Divot
DeeplabV3+	0.97	0.72	0.97	0.46
DeeplabV3+ + ST	0.98	0.81	0.98	0.64
DeeplabV3+ + AWB	0.99	0.82	0.98	0.66
PSPNet	0.95	0.64	0.95	0.34
PSPNet + ST	0.97	0.68	0.97	0.39
PSPNet + AWB	0.97	0.71	0.97	0.45
FCN8	0.97	0.70	0.97	0.44
FCN8 + ST	0.98	0.77	0.98	0.55
FCN8 + AWB	0.98	0.84	0.99	0.68
UNET	0.98	0.79	0.98	0.59
UNET + ST	0.98	0.81	0.98	0.64
UNET + AWB	0.99	0.83	0.98	0.68

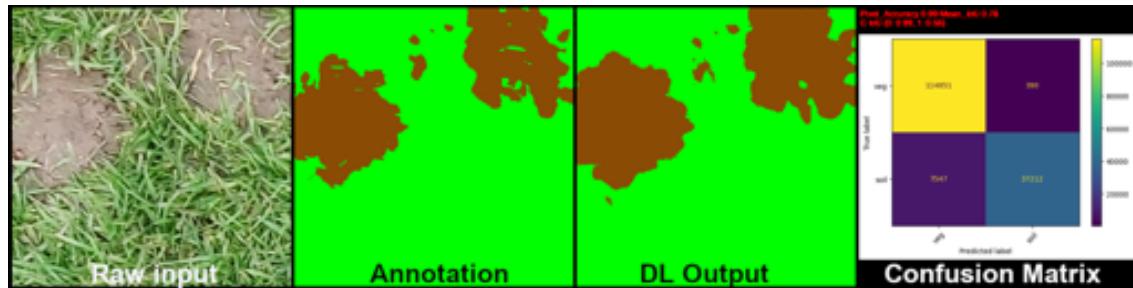


Figure 5: Result on test image using UNET + ST, trained on our synthetic dataset

## 6 Discussion

This paper described the development of a test dataset and training of a number of segmentation networks to create a baseline using the public GrassClover data. A common smartphone camera was used as the test system; this choice imposes a hardware limitation but this is common across many similar vision systems. The synthetic dataset were trained across these similar DL models and presented positive results. A number of colour augmentation strategies were presented to address colour inconsistency that exists between the synthetic dataset and the real world test dataset. AWB augmentation strategy

has shown positive increases in performance, furthermore the colour transfer algorithm has produced visually impressive results and has led to an interesting increase in performance.

## 7 Conclusions and Future Work

This work has shown that this blender simulation workflow can be used to address the challenges of annotation where there is a high object count such as grass in a turfgrass scene. The study benchmarks have additionally proven that this synthetic data can be used in the training of deep learning networks, with benchmarks presented against a real dataset. There are several directions possible in research but a natural progression with this blender pipeline is for the analysis for turfgrass scenes at different camera angles to explore the possibilities of turfgrass and divot topology mapping.

## Acknowledgments

The authors are grateful for the advice on Turfgrass maintenance practices from Damian McLaverty, currently chair of the Association of Turfgrass Professionals Ireland Association (ATPI).

## References

- [Afifi and Brown, 2019] Afifi, M. and Brown, M. S. (2019). What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance. arXiv:1912.06960 [cs].
- [Beard and Green, 1994] Beard, J. B. and Green, R. L. (1994). The Role of Turfgrasses in Environmental Protection and Their Benefits to Humans. *Journal of Environmental Quality*, 23(3):452–460. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.2134/jeq1994.00472425002300030007x>.
- [Bell et al., 2013] Bell, G. E., Kruse, J. K., and Krum, J. M. (2013). The Evolution of Spectral Sensing and Advances in Precision Turfgrass Management. In *Turfgrass: Biology, Use, and Management*, pages 1151–1188. John Wiley & Sons, Ltd. Section: 30 \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.2134/agronmonogr56.c30>.
- [Braun et al., 2020] Braun, R. C., Patton, A. J., Braithwaite, E. T., and Kowalewski, A. R. (2020). Establishment of low-input turfgrass from seed with patch and repair mixtures: Mulch and starter fertilizer effects. *Crop Science*, 60(6):3362–3376. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/csc2.20266>.
- [Buzzelli et al., 2023] Buzzelli, M., Zini, S., Bianco, S., Ciocca, G., Schettini, R., and Tchobanou, M. K. (2023). Analysis of biases in automatic white balance datasets and methods. *Color Research & Application*, 48(1):40–62. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/col.22822>.
- [Carlson et al., 2022] Carlson, M. G., Gaussoin, R. E., and Puntel, L. A. (2022). A review of precision management for golf course turfgrass. *Crop, Forage & Turfgrass Management*, 8(2):e20183. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cft2.20183>.
- [Cartucho et al., 2021] Cartucho, J., Tukra, S., Li, Y., S. Elson, D., and Giannarou, S. (2021). VisionBlender: a tool to efficiently generate computer vision datasets for robotic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9(4):331–338. Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1080/21681163.2020.1835546>.

- [Denninger et al., 2023] Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K. H., Humt, M., and Triebel, R. (2023). BlenderProc2: A Procedural Pipeline for Photorealistic Rendering. *Journal of Open Source Software*, 8(82):4901.
- [Ding et al., 2016] Ding, K., Raheja, A., Bhandari, S., and Green, R. L. (2016). Application of machine learning for the evaluation of turfgrass plots using aerial images. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping*, volume 9866, pages 84–96. SPIE.
- [Hendrycks et al., 2019] Hendrycks, D., Mazeika, M., and Dietterich, T. (2019). Deep Anomaly Detection with Outlier Exposure. arXiv:1812.04606 [cs, stat].
- [Klein, 2021] Klein, B. S. (2021). Golf Around the World 2021. Technical Report Fourth Edition, R&A Royal and Ancient Golf Club of St Andrews.
- [Liebisch et al., 2016] Liebisch, F., Pfeifer, J., Khanna, R., Lottes, P., Stachniss, C., Falck, T., Sander, S., Siegwart, R., Walter, A., and Galceran, E. (2016). Flourish-A robotic approach for automation in crop management. In *Workshop Computer-Bildanalyse und Unbemannte autonom fliegende Systeme in der Landwirtschaft*, volume 21, page 2016. Wernigerode Harz University Wernigerode, Germany.
- [Paszke et al., 2016] Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. (2016). ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. arXiv:1606.02147 [cs].
- [Price, 2023] Price, A. (2023). Poliigon - Textures, Models and HDRIs for 3D rendering.
- [Reinhard et al., 2001] Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P. (2001). Color Transfer between Images. *IEEE Computer Graphics and Applications*, 21:34–41.
- [Skovsen et al., 2019] Skovsen, S., Dyrmann, M., Mortensen, A. K., Laursen, M. S., Gislum, R., Eriksson, J., Farkhani, S., Karstoft, H., and Jorgensen, R. N. (2019). The GrassClover Image Dataset for Semantic and Hierarchical Species Understanding in Agriculture. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2676–2684, Long Beach, CA, USA. IEEE.
- [Turgeon, 1991] Turgeon, A. J. (1991). Turfgrass management. *Turfgrass management.*, 1(Ed. 3). Publisher: Prentice-Hall Inc.
- [Wang et al., 2022] Wang, Y., Ahsan, U., Li, H., and Hagen, M. (2022). A Comprehensive Review of Modern Object Segmentation Approaches. *Foundations and Trends® in Computer Graphics and Vision*, 13(2-3):111–283. arXiv:2301.07499 [cs].
- [Wäldchen et al., 2018] Wäldchen, J., Rzanny, M., Seeland, M., and Mäder, P. (2018). Automated plant species identification—Trends and future directions. *PLOS Computational Biology*, 14(4):e1005993.

# A Comparative Study of Image-to-Image Translation Using GANs for Synthetic Child Race Data

Wang Yao<sup>1</sup>, Muhammad Ali Farooq<sup>1</sup>, Joseph Lemley<sup>2</sup>, and Peter Corcoran<sup>1</sup>

<sup>1</sup>*School of Engineering, University of Galway, Ireland.*

<sup>2</sup>*Xperi Corporation, Galway.*

## Abstract

The lack of ethnic diversity in data has been a limiting factor of face recognition techniques in the literature. This is particularly the case for children where data samples are scarce and presents a challenge when seeking to adapt machine vision algorithms that are trained on adult data to work on children. This work proposes the utilization of image-to-image transformation to synthesize data of different races and thus adjust the ethnicity of children's face data. We consider ethnicity as a style and compare three different Image-to-Image neural network based methods, specifically pix2pix, CycleGAN, and CUT networks to implement Caucasian child data and Asian child data conversion. Experimental validation results on synthetic data demonstrate the feasibility of using image-to-image transformation methods to generate various synthetic child data samples with broader ethnic diversity.

**Keywords:** Image to Image Translation, Synthetic Data, Children Race, GAN, GDPR

## 1 Introduction

In recent years face authentication has witnessed significant advancements and has become widely deployed in various applications, such as authentication systems, immigration management, and financial security. Studies [Abdurrahim et al., 2018, Cavazos et al., 2020] have shown that these face authentication systems exhibit biases, especially when it comes to recognizing faces from certain racial or ethnic groups. This may lead to discrimination and injustice against specific groups, for instance by wrongly identifying them as suspects, restricting their access rights, or other such issues. Thus, race imbalance is a pressing issue that demands attention in face authentication applications.

However, collecting large amounts of effective race/ethnicity data in the real world is laborious and challenging because the process of data acquisition is expensive and time-consuming, especially when it comes to human subjects. Considering the General Data Protection Regulations (GDPR) in European Union (EU) region [European Parliament and Council of the European Union, 2016], when collecting any video, image, or audio data from human subjects, the scope of usage of such data and any subsequent processing of the data must be clearly defined and explained to the subject, which normally requires explicit consent. In addition, it is important that data must be stored securely and that it supports a series of rights, for example, the right of the subject to retract the stored data at any time. This becomes more complex in the case of engaging with child subjects, as the consent of the legal guardian is required, and it is preferable to inform the subject in plain language about the scope of collecting and further using this data.

Our work with the DAVID smart-toy platform [C3I, 2021] has motivated us to explore the generation of synthetic facial data, which is not subject to data protection regulations. In this work, we consider ethnicity as a style and employ style-to-style transformation to synthesize data from different races. This research adopts the potential of Image-to-Image (I2I) translation approaches to generate synthetic child race data, which will benefit the diversity of training data, reducing the ethnic bias of facial recognition systems, and improving the robustness of machine vision algorithms trained on this type of synthetic data. In this work, we aim to

generate synthetic Caucasian child data and Asian child data by training three I2I translation methods including pix2pix [Isola et al., 2017], CycleGAN [Zhu et al., 2017], and CUT [Park et al., 2020]. Then we qualitatively and quantitatively validated the synthetic child racial samples using machine vision algorithms.

## 2 Related Works

Image-to-Image translation methods refer to converting input images from a source domain to a target domain while preserving the content representations of the input image. I2I algorithms can solve many problems in computer vision tasks, such as image registration, image segmentation, and image restoration. It can be categorized into supervised I2I and unsupervised I2I based on whether the source domain images and target domain images are aligned image pairs. Isola et al. [Isola et al., 2017] proposed pix2pix to solve various supervised I2I tasks by adopting a conditional GAN framework, which is also a baseline for the image translation framework. However, training supervised translation in real-world scenes is impractical because it is difficult to create a paired dataset. CycleGAN [Zhu et al., 2017] and its variants such as TraVeL-GAN [Amodio and Krishnaswamy, 2019] employ cycle-consistency loss, which was proven effective in solving this problem. Study [Pang et al., 2021] reveals that most methods using cycle consistency constraints tend to directly synthesize a new domain with a global target style translation and rarely consider local objects or fine-grained instances during translation. CUT [Park et al., 2020] employs contrastive learning in a patch-based way instead of learning the entire images. LPTN [Liang et al., 2021] uses a Laplacian pyramid to decompose the input and achieve I2I translation.

Even though these I2I methods perform well in many tasks, one major challenge of these approaches is to achieve robust results for race-to-race child facial transformation applications. It is difficult to find a large-scale real child dataset with multiple race classes. Existing large-scale face datasets such as VGGFace2 [Cao et al., 2018], MS-Celeb-1M [Guo et al., 2016], and FFHQ [Karras et al., 2019] are generally focused on adult data. Although children's faces are present in some age datasets, the amount of data is small and the resolution varies widely [Moschoglou et al., 2017, Zhang et al., 2017]. Moreover, few studies have been conducted to generate synthetic data for different races [Yucer et al., 2020, Ba et al., 2021]. One study proposed by Yucer et al. [Yucer et al., 2020] is focused on generating synthetic race data through CycleGAN and using this data to improve face recognition accuracy. Another recent study [Ba et al., 2021] generates synthetic skin tones while retaining their pulsatile signals for exploring physiological signals. Both studies are focused on generating adult data. In this work, we explore generating synthetic child race data by using I2I translation methods.

## 3 Methodology

In this section, datasets, I2I translation methods, and evaluation metrics that were utilized in this study are detailed.

### 3.1 Datasets

In this work, we collect a synthetic child race dataset by finetuning a pre-trained StyleGAN2 [Karras et al., 2020] network. This dataset comprises of data from 2400 Asian children (1200 girls +1200 boys) and 2400 Caucasian children (1200 girls +1200 boys), which are generated by using the latent space editing technique [Wu et al., 2021]. We divided the dataset into groups of boys and girls and paired Asian children with Caucasian children one by one. The paired examples are shown in Figure 1.

To the best of our knowledge, there is no existing large-scale dataset focused on child race facial data. As mentioned in related work, recording real-world large-scale child

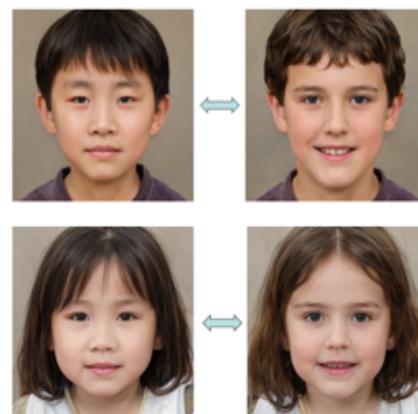


Figure 1: Example of two image pairs.

race data is a laborious task that will also require ethical approvals and must be performed according to GDPR regulations therefore in this work we have mainly focused on using synthetic child data for preserving personal identification data (PID). Moreover, considering the comparison with the supervised I2I method, we want to collect race data in image pairs. For this purpose, we have focused on using latent space editing in StyleGAN2 [Karras et al., 2020] to get the image pairs. Our goal is to generate child facial data with diversified ethnicity. To achieve this we have first fine-tuned StyleGAN2 [Karras et al., 2020] to produce synthetic child images in previous work [Farooq et al., 2023]. This paper serves as an initial study to compare the feasibility of different I2I methods for synthesizing child racial face data.

### 3.2 I2I translation methods

In this work, we have incorporated three typical I2I translation methods which are discussed below.

**Pix2pix:** The pix2pix model [Isola et al., 2017] is a conditional GAN, which means that the output image is generated conditionally on the input image. The discriminator gets the source image and the target image and determines whether the target image is a reasonable transformation of the source image. It is a supervised I2I model and requires many aligned image pairs for training.

**CycleGAN:** CycleGAN [Zhu et al., 2017] consists of two generators and two discriminators and designed cycle-consistent adversarial networks for unpaired I2I translation. A cycle-consistency loss is designed to measure the difference between the synthetic image produced by the second generator and the source image. CycleGAN uses cycle-consistent constraints to train unsupervised image translation models through the GAN architecture, which enables the conversion between two unpaired image sets.

**CUT:** CUT [Park et al., 2020] is unsupervised one-side translation beyond cycle-consistency constraint. It utilizes a contrastive learning framework to maximize the mutual information between two patches. A multi-layer, patch-based approach is used to encourage the generated images to be similar to the source images. This model avoids the use of cycle-consistency loss, and only one set of GANs is needed for image transformation.

### 3.3 Evaluation metrics

Three well-known quantitative evaluation metrics have been selected for validating the synthetic child race facial data. These are listed below with a brief discussion on what image quality each of these measures.

**FID:** Fréchet inception distance (FID) [Heusel et al., 2017] calculates the distance of the distribution between the synthetic images and the source images. The lower FID score means the model has a better performance.

**PSNR:** Peak signal-to-noise ratio (PSNR) measures the intensity differences between the reference image and the test image. The higher PSNR score indicates a higher quality of the test image.

**SSIM:** Structural similarity index (SSIM) [Wang et al., 2004] computes the perceptual distance between the reference image and the test image according to luminance, contrast, and structure. The higher SSIM score implies that the greater the similarity between the reference image and the test image.

## 4 Experiment

**Training Setting:** Four mappings {Caucasian boy → Asian boy, Asian boy → Caucasian boy, Caucasian girl → Asian girl, Asian girl → Caucasian girl} are adopted in our work. As pix2pix and CUT uses one directional transformation, four models are trained separately during the experiments. CycleGAN has bidirectional

S.no	Parameter	Pix2Pix/CycleGAN Value	CUT Value
1	Preprocess	Scale width	Scale width
2	Load size	256	256
3	Batch size	8	8
4	Learning Rate	0.0002	0.0002
5	Learning Rate Policy	linear	linear
6	Learning Rate Decay Iters	50	50
7	Dropout	False	False
8	Discriminator	PatchGAN	PatchGAN
9	Generator	resnet_9blocks	resnet_9blocks
10	Mirror augment	False	True
11	Training epochs	200	400

Table 1: Hyper-parameter Selection

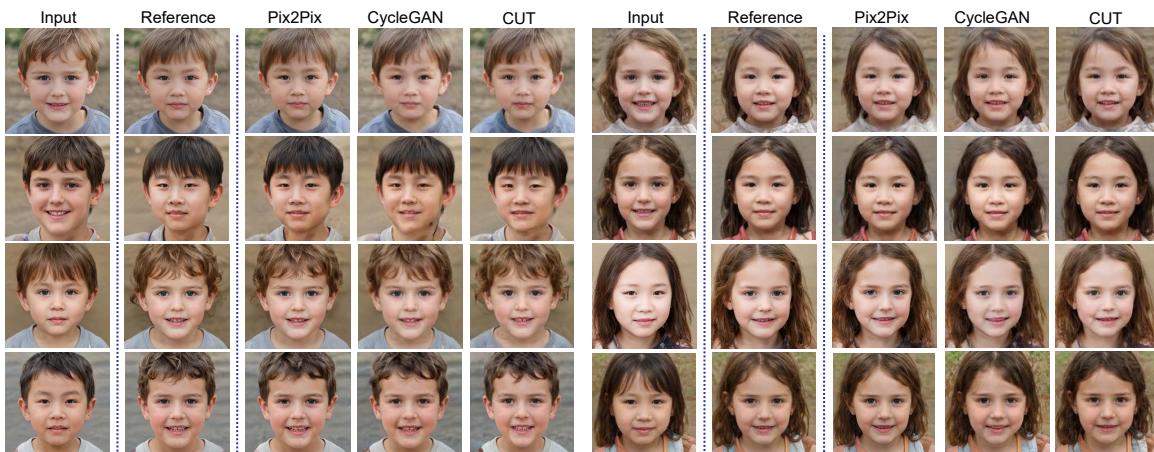


Figure 2: Qualitative results of synthetic boys and girls race data.

transformations, and thus only two models are trained during training. All the child faces are aligned and resized to  $256 \times 256$ . Table 1 shortlisted the optimal set of network hyperparameters during the training process. The complete experiment was performed on a server-grade machine equipped with 2 NVIDIA GeForce GTX TITAN.

#### 4.1 Qualitative Evaluation

Figure 2 shows the qualitative evaluation results from three I2I translation methods. The first two rows show the conversion from Caucasian to Asian. The last two rows show the conversion from Asian to Caucasian. The results from pix2pix show the best visual results, which are most similar to the reference image. The next most effective result is the image synthesized by the CUT. Figure 3 shows the details of different facial features of generated girl's Caucasian faces. The shape of the eyes has changed significantly in Figure 3, especially for the girl image generated by pix2pix. Secondly, we can observe the hair color

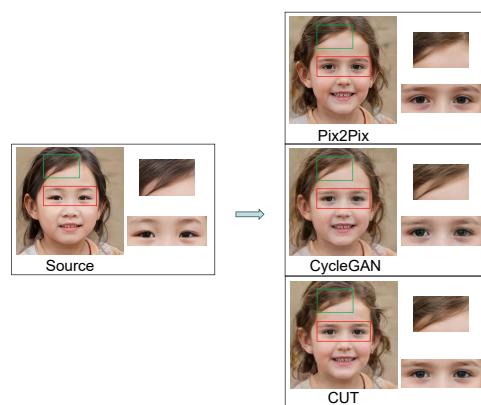


Figure 3: Example of converting an Asian girl to a Caucasian girl

Asian Boy → Caucasian Boy			
Evaluation Metrics	<b>Pix2Pix</b> [Isola et al., 2017]	<b>CycleGAN</b> [Zhu et al., 2017]	<b>CUT</b> [Park et al., 2020]
<b>FID</b> (↓)	49.59	31.46	26.36
<b>PSNR</b> (↑)	24.41	22.09	22.10
<b>SSIM</b> (↑)	0.75	0.65	0.65
Caucasian Boy → Asian Boy			
Evaluation Metrics	<b>Pix2Pix</b> [Isola et al., 2017]	<b>CycleGAN</b> [Zhu et al., 2017]	<b>CUT</b> [Park et al., 2020]
<b>FID</b> (↓)	29.62	29.54	25.06
<b>PSNR</b> (↑)	25.38	22.64	22.74
<b>SSIM</b> (↑)	0.78	0.68	0.67
Asian Girl → Caucasian Girl			
Evaluation Metrics	<b>Pix2Pix</b> [Isola et al., 2017]	<b>CycleGAN</b> [Zhu et al., 2017]	<b>CUT</b> [Park et al., 2020]
<b>FID</b> (↓)	26.31	42.27	25.09
<b>PSNR</b> (↑)	24.02	21.42	21.58
<b>SSIM</b> (↑)	0.70	0.60	0.62
Caucasian Girl → Asian Girl			
Evaluation Metrics	<b>Pix2Pix</b> [Isola et al., 2017]	<b>CycleGAN</b> [Zhu et al., 2017]	<b>CUT</b> [Park et al., 2020]
<b>FID</b> (↓)	44.47	43.24	34.87
<b>PSNR</b> (↑)	24.29	21.15	21.73
<b>SSIM</b> (↑)	0.70	0.62	0.63

Table 2: The average FID, PSNR, and SSIM scores of different I2I methods.

transformation since it becomes light brown and the shape of the hair is more curly when compared to the original (reference) image.

We will conduct a MOS study to provide further evidence on the quality of generated images but are currently waiting for ethical approvals. In the meantime, several people in our group have assisted with a small informal study where they are asked to distinguish between the original data and the generated data. From the results of this informal study, indications are that synthetic data are difficult to distinguish from original data samples.

## 4.2 Quantitative Evaluation

In order to evaluate the performance of these tuned GAN models, we have generated 100 Caucasian boys, 100 Asian boys, 100 Caucasian girls, and 100 Asian girls through Pix2Pix, CycleGAN, and CUT using unseen test data. We have calculated the average FID, PSNR, and SSIM scores through the reference images and generated images. The results are shown in Table 2. The synthetic images generated from pix2pix have the highest scores on PSNR and SSIM, which indicate that the synthetic images generated from pix2pix comprises robust quality facial features from human perception. CUT has the best FID value through three models, which means the distribution of generated images from CUT is close to the distribution of the source images.

## 4.3 Synthetic Facial Race Analysis

To analyze the ethnicity attribute of synthetic child data, a pre-trained Deepface [Serengil and Ozpinar, 2021] model is used to classify the race of generated child data. This model is trained on a large-scale real balanced ethnicity dataset, which has six race labels including Asian, white, middle eastern, Indian, Latino Hispanic, and black. Since this work focuses on two types of race data i.e. Asian and Caucasian/white; thus we only focus on the classification accuracy of Asian and Caucasian/white. Table 3 shows the results of the racial classification. From Table 3, it can be observed that synthetic child race data were classified with robust accuracy levels of 99% for Asian boys using pix2pix and CUT whereas we achieved 97% for Caucasian girls using pix2pix.

Model	Test Data	Asian	Caucasian	Other Race
pix2pix	Caucasian Boy	4%	76%	20%
CycleGAN	Caucasian Boy	5%	76%	19%
CUT	Caucasian Boy	8%	72%	20%
pix2pix	Asian Boy	99%	1%	0
CycleGAN	Asian Boy	96%	2%	2%
CUT	Asian Boy	99%	1%	0
pix2pix	Caucasian Girl	0	97%	3%
CycleGAN	Caucasian Girl	4%	85%	11%
CUT	Caucasian Girl	0	94%	6%
pix2pix	Asian Girl	89%	10%	1%
CycleGAN	Asian Girl	96%	3%	1%
CUT	Asian Girl	89%	9%	2%

Table 3: The average accuracy of race classification.

## 5 Discussion

In this work, we have focused on using synthetic child data for generating different race variations. The main reason for us to explore using synthetic data over real child data is due to the DAVID embedded smart toy platform [C3I, 2021] where child data is required to fine-tune and compress computer vision models originally optimized on adult subjects. In this project, we have tried to gather child data by using a 3D scanner. When we engaged with the data protection office, we realized that the complexity of managing and providing access to original child data made this approach impractical.

This is reflected in several aspects as below.

- Personally Identifiable Data (PID) Protection: Synthetic data can be used to avoid collecting sensitive personally identifiable Data (PID) from a vulnerable population of data subject, such as children. State-of-the-art data synthesis techniques enable data from adult subjects to be adapted to generate corresponding child data samples and in future work we can validate these data against original dataset when, and if, these become publicly available.
- Cost Effective: Utilizing synthetic data methods is far more cost-effective when compared to collecting and labelling large volumes of real child data. Collecting real data can involve significant expenses, such as data collection infrastructure, data storage, and data labelling costs. Thus, synthetic data generation and further utilizing it for experimental analysis avoids these costs, making it an attractive option for training advanced machine learning models.
- Controllable Data: Our work provides a range of tools, based on state-of-the-art data transformation models, to add expression, lighting, age, gender and pose variations to the original data samples. This work adds the additional capability to provide ethnic variations, an important tools to help diversify machine vision algorithms based on neural-AI models.
- GDPR Compliance: In situations where it is necessary to use and further share the data for research analysis while ensuring the anonymity of individuals, synthetic data can be freely employed. By replacing real data with synthetic data, the privacy of individuals can be preserved, thus addressing ethical concerns and complying with general data protection regulations (GDPR).

Note that a full validation of our proposed use of models based on generated data will require testing on original child subjects. This work is part of the DAVID roadmap and one of our industry partners has collected suitable original data from c.500 child subjects. Unfortunately, due to GDPR, such data can only be shared within the DAVID research consortium.

## 6 Conclusion

We have presented a comparative study on image-to-image translation methods including pix2pix, CycleGAN, and CUT to generate child race data which aims to solve the diversity issues in child data. As an initial work, we have trained ten models to explore the translation between Caucasian child faces and Asian child faces in this research and the experimental results show that it is feasible to synthesize child race faces through image-to-image translation. Three evaluation metrics have been adopted in our experiment. The results show that CUT has the highest FID value of all models, while pix2pix has the highest PSNR and SSIM scores.

A pretrained ethnic classification model is introduced to evaluate the synthetic race data, which shows that synthetic child ethnicity can be classified accurately. Since the dataset used in our work was manipulated by the latent code of a finetuned StyleGAN2 model, each pair of images (Asian  $\leftrightarrow$  Caucasian) has some underlying similarity. This may introduce some limitations but this work should be regarded as early-stage proof-of-concept rather than a comprehensive study. Future work will extend to handle more diversified race generation, and combine this work with state-of-the-art text-to-image frameworks.

## Acknowledgments

This research is supported by (i) Irish Research Council Enterprise Partnership Ph.D. Scheme (Project ID: EPSPG/2020/40), (ii) Xperi Corporation, Ireland, and (iii) the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF).

## References

- [Abdurrahim et al., 2018] Abdurrahim, S. H., Samad, S. A., and Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34:1617–1630.
- [Amodio and Krishnaswamy, 2019] Amodio, M. and Krishnaswamy, S. (2019). Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8983–8992.
- [Ba et al., 2021] Ba, Y., Wang, Z., Karinca, K. D., Bozkurt, O. D., and Kadambi, A. (2021). Overcoming difficulty in obtaining dark-skinned subjects for remote-ppg by synthetic augmentation. *arXiv preprint arXiv:2106.06007*.
- [C3I, 2021] C3I (2021). David - smart toys. <https://www.universityofgalway.ie/c3i/datasets/datacollectionactivities/3dscanner/david-smarttoys/>.
- [Cao et al., 2018] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.
- [Cavazos et al., 2020] Cavazos, J. G., Phillips, P. J., Castillo, C. D., and O’Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111.
- [European Parliament and Council of the European Union, 2016] European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council.
- [Farooq et al., 2023] Farooq, M. A., Yao, W., Costache, G., and Corcoran, P. (2023). Childgan: Large scale synthetic child facial data using domain adaptation in stylegan. *arXiv preprint arXiv:2307.13746*.
- [Guo et al., 2016] Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III* 14, pages 87–102. Springer.

- [Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- [Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.
- [Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.
- [Karras et al., 2020] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.
- [Liang et al., 2021] Liang, J., Zeng, H., and Zhang, L. (2021). High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9392–9400.
- [Moschoglou et al., 2017] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59.
- [Pang et al., 2021] Pang, Y., Lin, J., Qin, T., and Chen, Z. (2021). Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881.
- [Park et al., 2020] Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer.
- [Serengil and Ozpinar, 2021] Serengil, S. I. and Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.
- [Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.
- [Wu et al., 2021] Wu, Z., Lischinski, D., and Shechtman, E. (2021). Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872.
- [Yucer et al., 2020] Yucer, S., Akçay, S., Al-Moubayed, N., and Breckon, T. P. (2020). Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19.
- [Zhang et al., 2017] Zhang, Z., Song, Y., and Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818.
- [Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

# AudRandAug: Random Image Augmentations for Audio Classification

Teerath Kumar<sup>1</sup>, Muhammad Turab<sup>2</sup>, Alessandra Mileo<sup>3</sup>, Malika Bendechache<sup>4</sup>, and Takfarinas Saber<sup>4</sup>

<sup>1</sup>*CRT-AI Centre, School of Computing, Dublin City University, Ireland*

<sup>2</sup>*Norwegian University of Science and Technology (NTNU), Gjovik, Norway*

<sup>3</sup>*INSIGHT Research Centre, School of Computing, Dublin City University, Ireland*

<sup>4</sup>*Lero Research Centre, School of Computer Science, University of Galway, Ireland*

## Abstract

Data augmentation has proven to be effective in training neural networks. Recently, a method called RandAug was proposed, randomly selecting data augmentation techniques from a predefined search space. RandAug has demonstrated significant performance improvements for image-related tasks while imposing minimal computational overhead. However, no prior research has explored the application of RandAug specifically for audio data augmentation, which converts audio into an image-like pattern. To address this gap, we introduce AudRandAug, an adaptation of RandAug for audio data. AudRandAug selects data augmentation policies from a dedicated audio search space. To evaluate the effectiveness of AudRandAug, we conducted experiments using various models and datasets. Our findings indicate that AudRandAug outperforms other existing data augmentation methods regarding accuracy performance.

**Keywords:** Audio Classification, Data Augmentation, Random Audio Augmentation

## 1 Introduction

Deep learning (DL) has successfully addressed complex problems, proving proficiency in managing large datasets and discerning intricate patterns. Consequently, DL has become an indispensable tool for various tasks, including image recognition [Aleem et al., 2022, Wu and Chen, 2015, Fujiyoshi et al., 2019, Kumar et al., 2021a, Kumar et al., 2023b, Kumar et al., 2023a, Ranjbarzadeh et al., 2023], natural language processing [Torfi et al., 2020, Hirschberg and Manning, 2015, ?], and audio processing [Hershey et al., 2017, Fu et al., 2010, Turab et al., 2022, Park et al., 2020, Chandio et al., 2021]. Notably, DL has demonstrated impressive performance in the field of audio data analysis. Extensive research has been conducted on numerous tasks such as audio classification, music generation, and environmental sound classification [Lee et al., 2017b].

Previous studies [Fu et al., 2010, Turab et al., 2022, Park et al., 2020, Kumar et al., 2020] have highlighted the challenge of training neural networks directly on raw audio data, as it can be difficult for them to learn essential features. To overcome this limitation, researchers have shown that neural networks can achieve significantly improved performance by training them on audio-specific features [Palanisamy et al., 2020]. Convolutional Neural Networks (CNNs) have been widely employed for audio content analysis, utilizing various features and methods [Turab et al., 2022, Palanisamy et al., 2020, Li et al., 2019].

Despite the accuracy achieved through feature extraction methods, there remains room for improvement due to limited availability of labeled data. Deep learning models require large-scale labeled data to learn more accurate features. However, the process of labeling data on a large scale is tedious, time-consuming, and expensive [Kumar et al., 2021b]. To address this challenge, various data augmentation (DA) techniques can be applied to existing data by increasing the diversity and size of data, allowing the model to learn from different

perspectives of each sample. The objective is to train the network on additional distorted data, enabling the network to become invariant to these distortions and generalize better to unseen data. Several studies have explored data augmentation methods in the audio domain [Ko et al., 2015, Nanni et al., 2020, Jain et al., 2021]. In line with the principles of image randAug [Cubuk et al., 2020], we propose a novel approach for audio classification called AudRandAug, which is demonstrated in Figure 1. Our work contributes in the following ways:

- Inspired by RandAug, we introduce a novel data augmentation technique named AudRandAug.
- We perform several experiments to select the most effective augmentation methods for inclusion in the search space of AudRandAug
- To validate the proposed approach, we perform several experiments on different datasets.
- We provide code in GitHub repository: <https://github.com/turab45/AudRandAug.git>

The rest of the paper is organized as, section 2 discusses the related work, section 3 explains the proposed methodology, section 4 provides experimental details such as datasets, training setup and results, and finally section 6 concludes the work.

## 2 Related Work

This section discusses relevant data augmentation work in the audio domain. Deep learning methods have been widely applied to audio/sound data, such as music genre classification [Dong, 2018, Choi et al., 2017, Zhang et al., 2016], audio generation [Oord et al., 2016, Roberts et al., 2018], environmental sound classification [Guzhov et al., 2021, Aytar et al., 2016, Demir et al., 2020], and more [Chachada and Kuo, 2014, Dandashi and AlJaam, 2017]. From an architectural perspective, various methods have been explored for audio classification. Models using 1-D Convolution, such as EnvNet [Tokozume and Harada, 2017] and Sample-CNN [Lee et al., 2017b], have been proposed for raw audio waveform classification. However, recent work has primarily focused on utilizing CNN on spectrogram (an image pattern), which has led to state-of-the-art (SOTA) results. Dong et al. [Dong, 2018] proposed a CNN-based method for music genre classification, and Palanisamy et al. [Palanisamy et al., 2020] showed that an ImageNet pre-trained model could be a strong baseline network for audio classification.

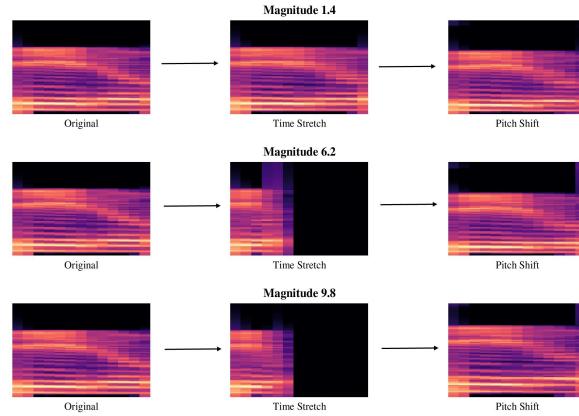
In addition to architectural considerations, data augmentation has shown promising results in various audio tasks. For convenience, audio data augmentation can be broadly divided into two levels: (i) data augmentation on the raw audio level and (ii) data augmentation on the feature level.

### 2.1 Data augmentation on raw audio level

Extensive research has been carried out on using deep learning techniques for raw audio data analysis. Various models have been developed specifically for classifying raw audio waveforms using 1-D Convolutions. For instance, EnvNet [Tokozume and Harada, 2017] and Sample-CNN [Lee et al., 2017b] are notable examples of models that leverage raw audio waveforms as their inputs. These models have demonstrated significant advancements in achieving SOTA performance across different sound categories [Lee et al., 2017a].

### 2.2 Data augmentation on features level

Recent research has emphasized employing CNNs on spectrograms to achieve SOTA outcomes. Dong et al. [Dong, 2018] proposed a CNN-based method for music genre classification, achieving accuracy of 70%. Additionally, Palanisamy et al. [Palanisamy et al., 2020] demonstrated that a pre-trained ImageNet model can serve as a strong baseline network for audio classification. To further enhance generalization, a few studies have explored feature extraction [Turab et al., 2022, Su et al., 2019, Liu et al., 1998] and data augmentation



**Figure 1: Example audio mel-spectrograms augmented by AudRandAug.** In these examples,  $N = 2$  and different  $M$  magnitude values are shown. As the magnitude of the distortion rises, so does the strength of the augmentation.

approaches [Ko et al., 2015, Nanni et al., 2020, Kim et al., 2021]. In the work by Turab et al., [Turab et al., 2022], different feature selection methods were investigated for audio using ensemble techniques. The search for optimal augmentation policies was explored in [BSG, 2020], while Kumar et al. [Kumar et al., 2020] proposed a novel intra-class random erasing data augmentation to enhance network robustness. Furthermore, Kim et al. [Kim et al., 2021] introduced Specmix, a novel audio data augmentation technique specifically designed for time-frequency domain features. This approach improved the performance of various neural network architectures by up to 2.7%. Salamon et al. [Salamon and Bello, 2017] proposed a deep neural network architecture coupled with audio augmentations to address the challenge of data scarcity in their work. Among all these approaches, none has explored image randAugment approach for audio data. To the best of our knowledge, we are the first to explore it.

### 3 Proposed Methodology

In this section, we explain the proposed method and used data augmentation methods in the search space.

#### 3.1 Method

Inspired by the RandAug [Cubuk et al., 2020] in the domain of images, we introduce a random data augmentation technique for audio classification called AudRandAug. This approach involves determining the optimal parameters for each specific data augmentation operation. Subsequently, we apply a total of  $N$  data augmentations, each with its corresponding optimal magnitude or parameter(s), like an optimal magnitude for time stretch augmentation is 1.4. The proposed algorithm is provided in **algorithm 1**.

**Require:**  $N$  (Integer),  $M$  (List)

**Ensure:** Selected augmentations and magnitudes

- 1:  $\text{augmentations} \leftarrow ["\text{Noise}", "\text{Pitch}", "\text{Time}", "\text{Padding}", "\text{Clip}", "\text{Trim}", "\text{Reverse}", "\text{BPF}", "\text{BSF}"]$
- 2:  $\text{selected\_augmentations} \leftarrow \text{random choice from augmentations, size } N$
- 3:  $\text{selected\_magnitude} \leftarrow M$  where  $\text{selected\_augmentation}$  equals  $\text{augmentation}$
- 4: **return**  $[(\text{aug}, m) \text{ for } (m, \text{aug}) \text{ in } \text{zip}(\text{selected\_magnitude}, \text{selected\_augmentations})]$  where  $m$  is magnitude of particular augmentation

**Algorithm 1:** AudRandAug

In our proposed approach, we adopt the same algorithm as described in RandAug [Cubuk et al., 2020]. The selection of data augmentation from the search space is performed with a uniform probability. We investigate several essential audio data augmentations, keeping in mind that all augmentations are applied to audio waveforms and subsequently converted into Mel-spectrograms. Finally, these augmented spectrograms are used as inputs to the CNN model. Table 1 provides a detailed overview of each used data augmentation technique.

Augmentation	Description
Noise Injection	This involves introducing additive white Gaussian noise (AWGN) to the original audio recording through element-wise addition.
Pitch Shifting	This alters the pitch of an audio recording without impacting its duration or timing.
Time Stretching	This modifies the speed or duration of an audio recording while preserving its pitch and tonal characteristics. This is achieved by utilizing the Short-time Fourier transform (STFT) technique.
Padding	Padding in audio refers to the technique of enhancing the sound quality of a recording by replacing a fraction of the beginning or end of the audio with padded sections.
Clip	Clipping removes excessive audio signal to prevent distortion and ensure a clean sound.
Reverse	Reversing an audio signal involves inverting its polarity, commonly used to create a reversed playback effect or special audio effects.
Band Pass Filter	A band pass filter is an electronic filter designed to permit a specific range of frequencies to pass through while attenuating all others. This filter is frequently employed in audio applications to eliminate unwanted noise and interference, ensuring optimal sound quality.
Gain	To enhance the model's resilience to variations in input gain, it is beneficial to multiply the audio by a random amplitude factor. By doing so, the model becomes less dependent on specific gain values and exhibits more consistent performance across a diverse range of input signals.
Time Masking	This is an audio technique where a randomly selected portion of the audio is made silent, effectively removing unwanted noises or creating unique effects. This method is commonly employed to enhance audio quality and achieve specific audio effects.

Table 1: All the used data augmentation methods

## 4 Experiment Design

In this section, we explain the training setup, dataset, and results.

### 4.1 Training Set up

We used custom CNN and pre-trained VGG model, 0.001 learning rate, Adam optimizer, 100 epoch. The custom CNN is 2 convolutional layers network. First convolutional layer followed by max pooling, drop with 0.2. Second convolutional layer is followed by max pooling then flatten. Then two dense layers are used.

### 4.2 Datasets

We use Free Spoken Digits Dataset (FSDD) [Jackson, 2016] which is a simple audio dataset consisting of English spoken digit recordings in .wav files at 8khz. It contains 3,000 recordings from 6 speakers (50 of each digit per speaker) and English pronunciations, and it has 10 classes (0-9) and duration of the recordings is 1-2 seconds. Another dataset UrbanSound8K dataset [Salamon et al., 2014] contains 8732 labeled urban sound recordings in .wav format. All recordings are of a duration of 4 seconds from 10 classes. The files are sorted by 10 folds (folders called fold1-fold10)

Custom CNN Model Results				
Augmentation	FSDD dataset		UrbanSound8K dataset	
	Performance	Change ( $\Delta D$ )	Performance	Change ( $\Delta D$ )
Baseline	92.00	-	95.00	-
+ Noise Injection	94.5	2.5	97.27	2.27
+ Pitch Shifting	94.83	2.83	97.26	2.26
+ Time Stretching	92.50	0.5	97.23	2.23
+ Padding	92.50	0.5	97.13	2.13
+ Clip	93.33	1.33	93.11	1.89
+ Reverse	93.83	1.83	93.13	1.87
+ Band Pass Filter	93.00	1.0	97.31	2.31
+ Gain	96.50	4.5	97.32	2.32
+ Time Mask	92.16	0.16	96.56	1.56
+ Ours	97.16	5.16	96.37	1.37

VGG Model Results				
Augmentation	Performance	Change ( $\Delta D$ )	Performance	Change ( $\Delta D$ )
Baseline	95.95	-	96.37	-
+ Noise Injection	98.16	2.15	97.89	1.52
+ Pitch Shifting	98.66	2.71	98.19	1.82
+ Time Stretching	94.18	1.77	91.17	5.2
+ Padding	93.87	2.39	98.34	1.97
+ Clip	98.66	2.71	98.42	2.05
+ Reverse	93.83	2.12	95.92	0.45
+ Band Pass Filter	93.00	2.95	98.25	1.88
+ Gain	98.66	2.71	97.09	0.72
+ Time Mask	94.39	1.56	97.11	0.74
+ Ours	98.92	2.97	98.63	2.26

Table 2: Result using custom CNN and Pre-trained VGG models

### 4.3 Pre-processing

First, we apply augmentation on signal level, as the mentioned augmentation methods perform better on signal level rather than mel-spectrogram. We resize the mel-spectrogram to 32 x 32 as an image-like feature before feeding to the network. RandAug applied before training as a data preprocessing step.

## 5 Results

To evaluate the effectiveness of our proposed approach, we conducted experiments using various models on two datasets: FSDD and UrbanSound8K. It is important to note we included all the techniques in the search space that perform better than the baseline. We present the experimental results in Table 2, where a custom CNN was used as the baseline for both datasets. Accuracy served as the evaluation metric. The table reports the difference between each data augmentation (DA) technique and the baseline accuracy, denoted as  $\Delta D$ . A green  $\Delta D$  indicates an improvement in accuracy compared to the baseline, while a red  $\Delta D$  signifies a decrease. Only the data augmentation techniques that demonstrated improved accuracy are included in the table.

Our proposed data augmentation technique using custom CNN exhibited a significant absolute improvement of 5.16% on the FSDD dataset and 1.37% on the UrbanSound8K dataset. The 5.16% absolute improvement over the baseline on the FSDD dataset is particularly noteworthy, as it represents the highest accuracy perfor-

mance among all the utilized DAs. Although the 1.37% improvement on the UrbanSound8K dataset is not the highest, it still demonstrates a competitive enhancement in accuracy.

For the pre-trained VGG model, we conducted a similar set of experiments as with the CNN model. However, we observed that fewer data augmentation methods showed improved performance compared to the CNN case. Therefore, we excluded those augmentations with lower accuracy performance compared to the baseline from the search space. Our proposed method exhibited superior accuracy performance compared to all other data augmentation methods across both datasets. For the FSDD dataset, the proposed method showed an absolute improvement of nearly 3% over the baseline, while for the UrbanSound8K dataset, it demonstrated an absolute improvement of approximately 2.30%. Overall, our proposed method achieved the best accuracy performance among all the methods employed.

## 6 Conclusion

This paper introduces AudRandAug, a novel data augmentation technique specifically designed for audio data. AudRandAug selects data augmentation policies from a dedicated audio search space and demonstrates remarkable performance improvements compared to the baseline. Through extensive experiments on FSDD and UrbanSound8K datasets, using various models, AudRandAug consistently outperforms other data augmentation methods. The results validate the effectiveness and potential of AudRandAug in enhancing the performance of audio-related models. By addressing the specific needs of audio data, this research contributes to the advancement of audio tasks within the computer vision field. In future, AudRandAug can be used as a powerful technique for audio data augmentation, demonstrating significant accuracy improvements. This work opens up possibilities for further research and development of tailored data augmentation methods to optimize audio-related applications.

## 7 Acknowledgment

This research was supported by Science Foundation Ireland under grant numbers 18/CRT/6223 (SFI Centre for Research Training in Artificial intelligence) and 13/RC/2094/P\_2 (Lero SFI Centre for Software). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [BSG, 2020] (2020). Search for optimal data augmentation policy for environmental sound classification with deep neural networks. *Journal of Broadcast Engineering*, 6(6).
- [Aleem et al., 2022] Aleem, S., Kumar, T., Little, S., Bendechache, M., Brennan, R., and McGuinness, K. (2022). Random data augmentation based enhancement: a generalized enhancement approach for medical datasets. *arXiv preprint arXiv:2210.00824*.
- [Aytar et al., 2016] Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29.
- [Chachada and Kuo, 2014] Chachada, S. and Kuo, C.-C. J. (2014). Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, 3.
- [Chandio et al., 2021] Chandio, A., Shen, Y., Bendechache, M., Inayat, I., and Kumar, T. (2021). Audd: audio urdu digits dataset for automatic audio urdu digit recognition. *Applied Sciences*, 11(19):8842.

- [Choi et al., 2017] Choi, K., Fazekas, G., Sandler, M., and Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2392–2396. IEEE.
- [Cubuk et al., 2020] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.
- [Dandashi and AlJaam, 2017] Dandashi, A. and AlJaam, J. (2017). A survey on audio content-based classification. In *2017 International conference on computational science and computational intelligence (CSCI)*, pages 408–413. IEEE.
- [Demir et al., 2020] Demir, F., Abdullah, D. A., and Sengur, A. (2020). A new deep cnn model for environmental sound classification. *IEEE Access*, 8:66529–66537.
- [Dong, 2018] Dong, M. (2018). Convolutional neural network achieves human-level accuracy in music genre classification. *arXiv preprint arXiv:1802.09697*.
- [Fu et al., 2010] Fu, Z., Lu, G., Ting, K. M., and Zhang, D. (2010). A survey of audio-based music classification and annotation. *IEEE transactions on multimedia*, 13(2):303–319.
- [Fujiyoshi et al., 2019] Fujiyoshi, H., Hirakawa, T., and Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *IATSS research*, 43(4):244–252.
- [Guzhov et al., 2021] Guzhov, A., Raue, F., Hees, J., and Dengel, A. (2021). Esresnet: Environmental sound classification based on visual domain models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4933–4940. IEEE.
- [Hershey et al., 2017] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.
- [Hirschberg and Manning, 2015] Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.
- [Jackson, 2016] Jackson, Z. (2016). Free spoken digit dataset (fsdd). *Retrieved February*, 1:2020.
- [Jain et al., 2021] Jain, A., Samala, P. R., Mittal, D., Jyoti, P., and Singh, M. (2021). Spliceout: A simple and efficient audio augmentation method. *arXiv preprint arXiv:2110.00046*.
- [Kim et al., 2021] Kim, G., Han, D. K., and Ko, H. (2021). Specmix: A mixed sample data augmentation method for training withtime-frequency domain features. *arXiv preprint arXiv:2108.03020*.
- [Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*.
- [Kumar et al., 2023a] Kumar, T., Mileo, A., Brennan, R., and Bendechache, M. (2023a). Rsmda: Random slices mixing data augmentation. *Applied Sciences*, 13(3):1711.
- [Kumar et al., 2021a] Kumar, T., Park, J., Ali, M. S., Uddin, A., and Bae, S.-H. (2021a). Class specific autoencoders enhance sample diversity. *Journal of Broadcast Engineering*, 26(7):844–854.
- [Kumar et al., 2021b] Kumar, T., Park, J., Ali, M. S., Uddin, A. S., Ko, J. H., and Bae, S.-H. (2021b). Binary-classifiers-enabled filters for semi-supervised learning. *IEEE Access*, 9:167663–167673.

- [Kumar et al., 2020] Kumar, T., Park, J., and Bae, S.-H. (2020). Intra-class random erasing (icre) augmentation for audio classification. In *Proceedings Of The Korean Society Of Broadcast Engineers Conference*, pages 244–247. The Korean Institute of Broadcast and Media Engineers.
- [Kumar et al., 2023b] Kumar, T., Turab, M., Raj, K., Mileo, A., Brennan, R., and Bendechache, M. (2023b). Advanced data augmentation approaches: A comprehensive survey and future directions. *arXiv preprint arXiv:2301.02830*.
- [Lee et al., 2017a] Lee, J., Kim, T., Park, J., and Nam, J. (2017a). Raw waveform-based audio classification using sample-level cnn architectures. *arXiv preprint arXiv:1712.00866*.
- [Lee et al., 2017b] Lee, J., Park, J., Kim, K. L., and Nam, J. (2017b). Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv preprint arXiv:1703.01789*.
- [Li et al., 2019] Li, X., Chebiyyam, V., and Kirchhoff, K. (2019). Multi-stream network with temporal attention for environmental sound classification. *arXiv preprint arXiv:1901.08608*.
- [Liu et al., 1998] Liu, Z., Wang, Y., and Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI signal processing systems for signal, image and video technology*, 20(1):61–79.
- [Nanni et al., 2020] Nanni, L., Maguolo, G., and Paci, M. (2020). Data augmentation approaches for improving animal audio classification. *Ecological Informatics*, 57:101084.
- [Oord et al., 2016] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
- [Palanisamy et al., 2020] Palanisamy, K., Singhania, D., and Yao, A. (2020). Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*.
- [Park et al., 2020] Park, J., Kumar, T., and Bae, S.-H. (2020). Search for optimal data augmentation policy for environmental sound classification with deep neural networks. *Journal of Broadcast Engineering*, 25(6):854–860.
- [Ranjbarzadeh et al., 2023] Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Tataei Sarshar, N., Tirkolaee, E. B., Ali, S. S., Kumar, T., and Bendechache, M. (2023). Me-ccnn: Multi-encoded images and a cascade convolutional neural network for breast tumor segmentation and recognition. *Artificial Intelligence Review*, pages 1–38.
- [Roberts et al., 2018] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, pages 4364–4373. PMLR.
- [Salamon and Bello, 2017] Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283.
- [Salamon et al., 2014] Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM ’14, page 1041–1044, New York, NY, USA. Association for Computing Machinery.
- [Su et al., 2019] Su, Y., Zhang, K., Wang, J., and Madani, K. (2019). Environment sound classification using a two-stream cnn based on decision-level fusion. *Sensors*, 19(7):1733.

- [Tokozume and Harada, 2017] Tokozume, Y. and Harada, T. (2017). Learning environmental sounds with end-to-end convolutional neural network. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2721–2725. IEEE.
- [Torfi et al., 2020] Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.
- [Turab et al., 2022] Turab, M., Kumar, T., Bendechache, M., and Saber, T. (2022). Investigating multi-feature selection and ensembling for audio classification. *arXiv preprint arXiv:2206.07511*.
- [Wu and Chen, 2015] Wu, M. and Chen, L. (2015). Image recognition based on deep learning. In *2015 Chinese Automation Congress (CAC)*, pages 542–546. IEEE.
- [Zhang et al., 2016] Zhang, W., Lei, W., Xu, X., and Xing, X. (2016). Improved music genre classification with convolutional neural networks. In *Interspeech*, pages 3304–3308.

# Compact & Capable: Harnessing Graph Neural Networks and Edge Convolution for Medical Image Classification

Aryan Singh, Pepijn Van de Ven, Ciarán Eising, and Patrick Denny

*University of Limerick*

## Abstract

Graph-based neural network models are gaining traction in the field of representation learning due to their ability to uncover latent topological relationships between entities that are otherwise challenging to identify. These models have been employed across a diverse range of domains, encompassing drug discovery, protein interactions, semantic segmentation, and fluid dynamics research. In this study, we investigate the potential of Graph Neural Networks (GNNs) for medical image classification. We introduce a novel model that combines GNNs and edge convolution, leveraging the interconnectedness of RGB channel feature values to strongly represent connections between crucial graph nodes. Our proposed model not only performs on par with state-of-the-art Deep Neural Networks (DNNs) but does so with 1000 times fewer parameters, resulting in reduced training time and data requirements. We compare our Graph Convolutional Neural Network (GCNN) to pre-trained DNNs for classifying MedMNIST dataset classes, revealing promising prospects for GNNs in medical image analysis. Our results also encourage further exploration of advanced graph-based models such as Graph Attention Networks (GAT) and Graph Auto-Encoders in the medical imaging domain. The proposed model yields more reliable, interpretable, and accurate outcomes for tasks like semantic segmentation and image classification compared to simpler GCNNs. Code available at AnonRepo.

**Keywords:** Medical Imaging, Machine Vision, GNN, GCNN, Image classification.

## 1 Introduction

Medical image classification and segmentation play critical roles in the field of medical imaging. Although there have been considerable advancements in image classification, medical image classification faces unique challenges due to the diverse dataset modalities, such as X-ray, Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Ultrasound (US), and Computed Tomography (CT). Variations within and between modalities, mainly stemming from the inherent differences in imaging technologies, complicate the classification process. Additionally, obtaining labeled training data is costly in the medical domain. Pre-trained DNNs address these issues through transfer learning techniques, yielding impressive results. However, DNNs exhibit limitations, including inductive bias, inefficient capture of spatial and local-level associations, and inconsistent performance across modalities [Rajpurkar et al., 2017, Zhou et al., 2021].

GNNs offer a solution to these complexities, handling variations in data with embedded relationships effectively and accommodating heterogeneous graph nodes[Kim et al., 2023]. The successful application of Knowledge-Based Graph Methods in medical diagnosis supports this notion. We have compared GNN architecture with Convolutional Neural Networks (CNNs) and discussed various types of Graph Convolutional Neural Networks (GCNNs). We propose a GCNN model integrated with Edge Convolution [Wang et al., 2018](GCNN-EC) for medical image classification. By performing graph convolution and edge convolution for edge prediction. Edge convolution overcomes the limitations of vanilla GCNN thus improving classification. Our method enhances model performance with reduced training time and data requirements. This research validates graph-based learning's efficiency for medical image data. In this study, we focus on classifying the

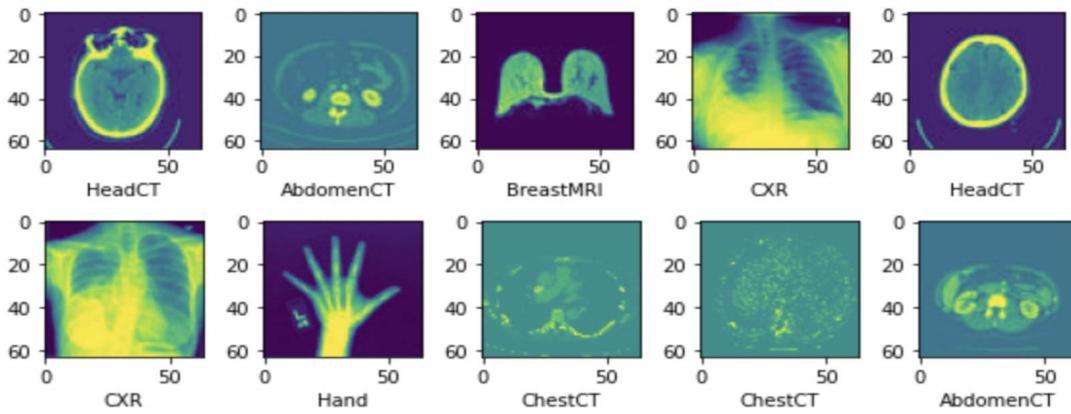


Figure 1: Sample MedMNIST data.

MedMNIST dataset [Yang et al., 2021], featuring 10 pre-processed datasets from various sources and modalities, with 708,069 images in 12 2D datasets. We narrow our research scope to six categories/classes containing 58,954 images with dimensions 28x28 as these classes represent diverse modalities, reflecting the compilation of images from various imaging techniques. These classes are AbdomenCT, BreastMRI, CXR, ChestCT, Hand, and HeadCT. The subsets are balanced.

We observe that our simple GCNN-EC outperforms leading state-of-the-art DNNs for specific MedMNIST dataset classes. Proposed model required less training than compared DNNs while using 100 times fewer parameters.

## 2 Prior Art

In this section, we will delve into the technicalities of CNNs, and compare their mechanisms with GCNNs. We will also shed light on three contemporary, state-of-the-art CNN models that have been utilized for the task of medical image classification. Furthermore, this section will introduce the diverse variants of GNNs, providing a comprehensive comparison from a technical standpoint. It will also cover the various applications of these models in medical domains.

### 2.1 CNNs

CNNs[Lecun et al., 1998] owe their name to the convolution operation, which involves overlaying a kernel onto the image grid and sliding it across the grid to extract local information, such as details from neighboring pixels. Technically, the convolution operation involves performing a dot product between the filter's elements and the corresponding elements of the image grid, then storing the result in an output matrix (often termed a feature map or convolved feature). As illustrated in Figure 2, the dot product employed in the convolution process is an aggregation operation. The main objective of this operation is to consolidate image data into a compressed form, making it feasible to extract global-level features from an image.

Thus, convolution as a process systematically extracts spatial hierarchies or patterns, starting from local pixel interactions (low-level features) to more abstract concepts (high-level features) as we progress deeper into the network. Finally, the hierarchical feature extracted from the preceding convolution and pooling operations is compressed into a compact and linear representation. The flattened feature vector derived is used for various tasks such as classification, segmentation, or feature localization.

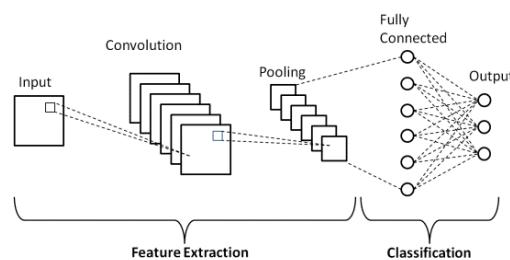


Figure 2: CNN architecture.

We have chosen three state-of-the-art DNNs that have demonstrated robust performance in image classification in the medical imaging domain for further discussion in this paper. The effectiveness of our proposed model is assessed in relation to these distinguished DNN models in the later section, thereby offering a comparative study.

**ResNet** [He et al., 2015] is a deep neural network architecture with a varying number of hidden layers, including a large number of convolutional layers to work efficiently by using residual blocks [He et al., 2015] that allow the network to effectively learn the residual or the difference between the input and output features. ResNets has been one of the best-performing models on the ImageNet dataset [Deng et al., 2009] for classification tasks. It has served as a skeleton for several DNNs that continue to use similar skip connection methods for achieving state-of-the-art performance. It has been applied for the classification of medical image data and has proved to produce state-of-the-art results [Alom et al., 2018], the ResNets-based model showed 99.05% and 98.59% testing accuracy for binary and multi-class breast cancer recognition.

**DenseNet** [Huang et al., 2016] is one of the densely connected deep-layered neural network architectures that also use residual blocks. They exploit the potential of the deep network by feature reuse, producing more condensed models that are easy to train and highly parameter efficient. Concatenating feature maps learned by different layers increases variation in the input of subsequent layers and improves efficiency. DenseNets uses the Network in Network architecture [Lin et al., 2014] which uses multi-layer perceptrons in the filters of convolutional layers to extract more complicated features. DenseNet has increasingly been applied as the backbone model for various medical imaging tasks[Zhou et al., 2021] from image registry to image embedding generation which has further been used for tasks like segmentation and classification.

**EfficientNet** [Tan and Le, 2019] makes use of techniques like compound scaling, that enable the efficient scaling of deep neural network architectures to meet specific requirements regarding data or resource limitations. Unlike other deep neural networks, EfficientNet achieves improved model performance without increasing the number of floating-point operations per second (FLOPS), resulting in enhanced efficiency. The method introduces the concept of efficient compound coefficients to uniformly scale the depth, width, and resolution of the network. When scaling a model by a factor of  $2^N$  in terms of computational resources, EfficientNet scales the network depth by  $\alpha^N$ , network width by  $\beta^N$ , and image size by  $\gamma^N$ . These coefficients, namely  $\alpha$ ,  $\beta$ , and  $\gamma$ , are determined through a grid search on the base model. By employing this compound scaling approach, EfficientNet strikes a balance between network accuracy and efficiency. EfficientNet has been proven to produce excellent results for medical image classification [Zhou et al., 2021] even in resource-constrained environments.

## 2.2 GNNs

A GNN is a specialized kind of neural network tailored for handling graph-structured data. It exploits the attributes of nodes and edges to learn representations for nodes, edges, and the overall graph. Its working principle is iterative message passing, where features from neighboring nodes are gathered by each node to update its own feature set. CNNs demonstrate limitations in capturing the associations between features within an image [Defferrard et al., 2017]. However, these intricate interconnections can be effectively captured by representing images as graphs and then utilizing GNNs to comprehend these intricate interdependencies. Also, GNNs better capture topological data features compared to CNNs[Bronstein et al., 2017]

The intricate complexities of graph data, ranging from structures as varied as protein sequences and chemical molecules to pixels in medical images serving as nodes, necessitate the transformation of these structures into suitable vector spaces. This transformation is crucial for performing computations and analyses. However, this comes with the daunting task of handling graph isomorphism issues, which involve identifying topological similarities between different graphs. In solving these intricate problems, the significance of GNNs becomes especially apparent, specifically in conjunction with the Weisfeiler-Lehman (WL) Isomorphism algorithm[Weisfeiler and Leman, 1968].

WL algorithm, a prevalent technique in this context, generates graph embeddings by aggregating colors, or more general features (image features, etc), of proximate nodes, culminating in a histogram-like representation

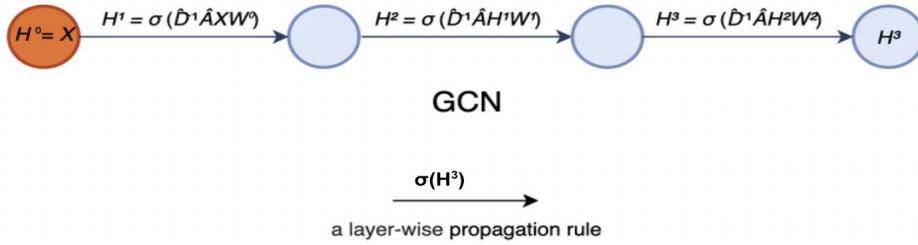


Figure 3: A simple GCN with 3 hidden layers.

for each graph. The conventional GNN architecture mirrors a neural rendition of the 1-WL algorithm, where '1' represents a single neighborhood hop. In this transformation, discrete colors evolve into continuous feature vectors, and neural network mechanisms are harnessed to aggregate information from the neighborhood of each node.

By virtue of this design, GNNs inherently embody a continuous variant of graph-based message passing similar to the WL algorithm. In this paradigm, details from a node's immediate surroundings are accumulated and relayed to the node, thereby facilitating learning from local graph structures[Errica et al., 2019]. This characteristic is at the core of the utility of GNNs in various domains requiring graph-based data analysis. Furthermore, we elaborate in detail on the specific type of GNN employed in this study, namely the GCNN.

There are two types of GCNNs:

1. GCNNs based on **spectral methods** (using convolutions via the convolution theorem [Bruna et al., 2013]). Spectral methods fall into the category of transductive learning, where learning and inference take place on the entire dataset. Spectral CNNs (SCNN) [Bruna et al., 2013] was the first implementation of CNNs on graphs, leveraging the graph Fourier transform [Wu et al., 2021] and defining the convolution kernel in the spectral domain. Examples include the Dynamic Graph Convolutional Network (DGCN), which has been effectively applied to detect relation heat maps in images for pose and gesture detection. HACT-Net [Dong et al., 2022] is a further example, which has been applied for the classification of Histopathological images.
2. GCNNs based on **spatial methods**. These GCNNs fall into the category of inductive learning, where learning and inference can be performed on a test and train dataset. They define convolution as a weighted average function over the neighborhood of the target vertex. For example, GraphSAGE [Hamilton et al., 2017] takes one-hop neighbors as neighborhoods and defines the weighting function as various aggregators over the neighborhood. The spatial GCNN is extremely robust due to its inductive learning which makes spatial GCNNs highly scalable.

We use a simple spectral GCNN  $f(X, A)$  that takes input  $X$  which is a vector of node features and an adjacency matrix  $A$ , along with a layer-wise propagation rule. The matrix  $A$  is normalized using methods mentioned in [Kipf and Welling, 2016] as multiplication of  $X$  and  $A$  will change the scale of feature vectors, which leads to disproportional learning from neighbors. The equation 1 defines the aggregation operation from one layer to another.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right). \quad (1)$$

$\tilde{A}$  is the adjacency matrix of the graph  $\mathcal{G}$ ,  $\tilde{D}$  is the degree matrix and  $W^{(l)}$  is a trainable weight matrix. The result is passed to an activation function  $\sigma(\cdot)$ . The output is then concatenated to get a new hidden state  $H^{(l+1)}$  for hidden layer  $l + 1$  as shown in Figure 3.

GCNN has its inherent limitations. They exhibit poor performance when confronted with dynamic graph structures, causing over-fitting to the training set. Furthermore, GCNNs are susceptible to the issue of over-

smoothing[Magner et al., 2020], whereby the addition of more convolutional layers results in an indistinguishable final embedding.

In this section, we defined and elaborated on the different types of CNN used in this study along with the types of GNNs and an explanation of spectral GCNN which has been used in the proposed method. In the next section, we explain the proposed method that leverages the power of GNNs, while also overcoming its limitations.

### 3 Our work

In this work, we present GCNN-EC which resolves identified issues around the limitation of CNNs in capturing the inherent connections between features within an image. The proposed method aims to leverage both local and global inter-pixel relationships by incorporating edge convolution along with graph convolution. Our procedure involves three stages: edge convolution, graph convolution, and classification. We merge RGB values into a node feature and compute edge features using a dynamic filter, the system processes the graph representation through multiple convolution layers before flattening the embedding for final classification. We explain these steps in detail in this section.

#### 3.1 Edge convolution

We begin by creating a node feature vector by combining RGB channel values. This is a vital step in transforming the RGB image data into a grid format, where each pixel is a node connected to adjacent pixels. Next, we use a dynamic filter[Brabandere et al., 2016] to learn edge features. This filter incorporates the node features from the immediate neighbors and also those at a two-hop distance, meaning it considers not just the nearest node but also the nodes that are connected to these nearest nodes. This filter is unique for each input and is learned by the network based on node features and the Euclidean distance between the node feature vectors. This learned filter is stored as a registered buffer in the network and not used during back-propagation. Rather, it is used as an edge feature while being passed through convolution layers. It is through this process that we generate an abstract understanding of the relationships between various components of the image. The graph augmented with node and edge features is further improved by an edge convolution layer [Wang et al., 2019], which performs convolution operations on the graph using the edge features. The edge convolution layer detects and enhances the edges or boundaries in an image (grid of pixels). It focuses on identifying sharp transitions in intensity/color, which are indicative of object edges. Finally giving a more distinguishable graph representation.

#### 3.2 Graph convolution

The graph representation enriched by edge convolution is passed through graph convolution layers, capturing features of the graph by incorporating features of nodes and their connections to finally create a more accurate graph-level embedding. One of the strengths of the graph convolution layer is its ability to incorporate both local and global information. By considering the neighborhood relationships between nodes, it captures local patterns and structures within the graph. Additionally, by aggregating information from neighboring nodes iteratively, it gradually incorporates global information, allowing for a comprehensive understanding of the overall graph. Graph convolution alone suffers from the problem of over-smoothing, This limitation can be overcome by enhancing the graph representation quality by edge convolution to capture meaningful edge information.

#### 3.3 Classification

The graph embedding, obtained by flattening the output of the graph convolution layers, is classified using a dense layer. This layer transforms the embedding by matrix multiplication and bias addition thus learning weighted connections and finally applying non-linear activation. Enabling accurate predictions based on the learned representation of the graph.

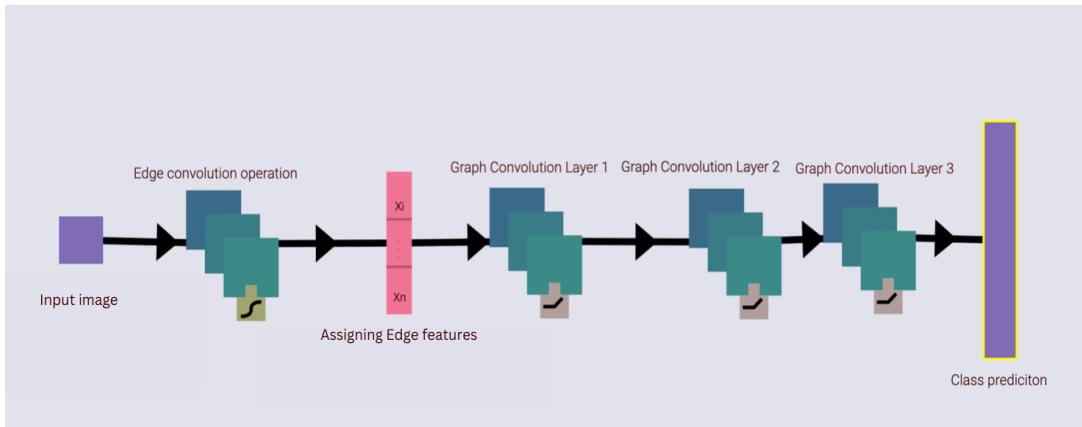


Figure 4: GCNN-EC architecture.

Table 1: Comparison of DNNs with proposed method for MedMNIST

Methods	AbdomenCT		BreastMRI		CXR		ChestCT		Hand		HeadCT		Parameters
	AUC	ACC											
ResNet18	0.800	0.839	0.897	0.899	0.832	0.842	0.901	0.940	0.915	0.921	0.733	0.762	11,689,512
EfficientNet-B0	0.901	0.907	0.905	0.918	0.958	0.960	<b>0.913</b>	<b>0.948</b>	0.907	0.911	0.874	0.894	4,014,658
DenseNet121	<b>0.936</b>	<b>0.942</b>	0.961	0.971	<b>0.972</b>	<b>0.985</b>	0.887	0.901	<b>0.916</b>	<b>0.925</b>	<b>0.899</b>	<b>0.914</b>	7,978,856
GCN-EC (ours)	0.876	0.882	<b>0.983</b>	<b>0.985</b>	0.957	0.965	0.748	0.813	0.886	0.905	0.869	0.874	<b>24,967</b>

We have employed PyTorch to implement our pipeline, while the Monai framework has been utilized for medical image processing. We generate a graph using MedMNIST images as the input and incorporate the Dynamic Edge Convolution layer[Wang et al., 2019] to perform edge convolution. The resulting learned representation then undergoes 3 graph convolution layers. We fine-tuned the model using Optuna [Akiba et al., 2019], obtaining a learning rate of 0.001 and a weight decay of 0.01. The parameter count in our models ranged from 24,967 to 67,938. We have used the Cross-Entropy loss function with Adam Optimizer and have trained our models for 4 epochs. The batch size used was 64, and the training, testing, and validation splits were 80%, 10%, and 10% respectively. GCNN-EC model architecture is shown in Figure4.

## 4 Result

In this section, we present the results achieved by our model on the 6 classes of the MedMNIST dataset to demonstrate the efficacy of simple GNN when compared with sophisticated DNNs. GCNN-EC model converges to stable loss value within 4 epochs. Our results are presented in Table 1 comparing the Area under the Curve (AUC) and Accuracy (ACC) of our model with the DNN models. From the plot in Figure 5, it is evident that our method is comparable to DenseNet and outperforms ResNet and EfficientNet, showing it as an effective classifier. The proposed method demonstrates superior performance compared to ResNet18 and EfficientNet-B0 while performing on par with DenseNet121. Notably, the GCNN-EC model utilizes 100 times fewer parameters than the three DNN models considered in this study. Furthermore, our model achieves a remarkable accuracy of 99.13% on the MNIST dataset (as indicated in the "gcnn-ec-mnist.py" file in the code).

## 5 Conclusion

Our model exhibited superior performance when compared to renowned CNNs like ResNet18 and EfficientNet-B0 while achieving comparable results to DenseNet121 on MedMNIST dataset. Notably, our model achieved

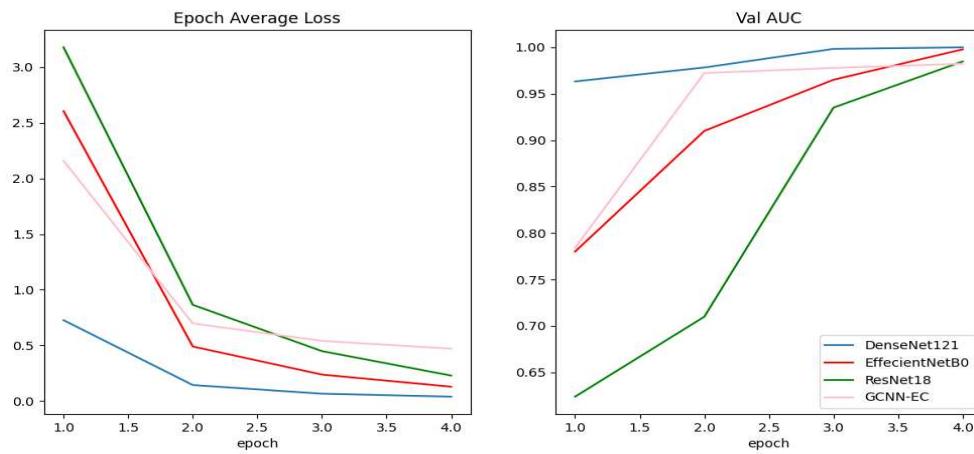


Figure 5: Loss and Area under the curve (AUC) over epochs for test dataset.

this with significantly fewer parameters (GCNN-EC: 24,967 vs. ResNet: 11.68M, EfficientNet: 4.01M, DenseNet: 6.95M), highlighting its efficiency and effectiveness in capturing meaningful features. This efficiency suggests the possibility of training our GCNN with significantly fewer data, an important factor in the medical field where properly labeled data is scarce and expensive.

## Acknowledgments

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6049. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Akiba et al., 2019] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework.
- [Alom et al., 2018] Alom, M. Z., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network.
- [Brabandere et al., 2016] Brabandere, B. D., Jia, X., Tuytelaars, T., and Gool, L. V. (2016). Dynamic filter networks.
- [Bronstein et al., 2017] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.
- [Bruna et al., 2013] Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs.
- [Defferrard et al., 2017] Defferrard, M., Bresson, X., and Vandergheynst, P. (2017). Convolutional neural networks on graphs with fast localized spectral filtering.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

- [Dong et al., 2022] Dong, Y., Liu, Q., Du, B., and Zhang, L. (2022). Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 31:1559–1572.
- [Errica et al., 2019] Errica, F., Podda, M., Bacciu, D., and Micheli, A. (2019). A fair comparison of graph neural networks for graph classification.
- [Hamilton et al., 2017] Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs.
- [He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.
- [Huang et al., 2016] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2016). Densely connected convolutional networks.
- [Kim et al., 2023] Kim, S., Lee, N., Lee, J., Hyun, D., and Park, C. (2023). Heterogeneous graph learning for multi-modal medical data analysis.
- [Kipf and Welling, 2016] Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.
- [Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- [Lin et al., 2014] Lin, M., Chen, Q., and Yan, S. (2014). Network in network.
- [Magner et al., 2020] Magner, A., Baranwal, M., and au2, A. O. H. I. (2020). Fundamental limits of deep graph convolutional networks.
- [Rajpurkar et al., 2017] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.
- [Tan and Le, 2019] Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks.
- [Wang et al., 2018] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2018). Dynamic graph CNN for learning on point clouds. *CoRR*, abs/1801.07829.
- [Wang et al., 2019] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds.
- [Weisfeiler and Leman, 1968] Weisfeiler, B. Y. and Leman, A. A. (1968). The reduction of a graph to canonical form and the algebra which appears therein.
- [Wu et al., 2021] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.
- [Yang et al., 2021] Yang, J., Shi, R., and Ni, B. (2021). MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE.
- [Zhou et al., 2021] Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Ginneken, B. V., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*.

# Assessing Intra-class Diversity and Quality of Synthetically Generated Images in a Biomedical and Non-biomedical Setting\*

Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairí O'Reilly

*Munster Technological University, Cork, Ireland.*

## Abstract

In biomedical image analysis, data imbalance is common across several imaging modalities. Data augmentation is one of the key solutions in addressing this limitation. Generative Adversarial Networks (GANs) are increasingly being relied upon for data augmentation tasks. Biomedical image features are sensitive to evaluating the efficacy of synthetic images. These features can have a significant impact on metric scores when evaluating synthetic images across different biomedical imaging modalities. Synthetically generated images can be evaluated by comparing the diversity and quality of real images. Multi-scale Structural Similarity Index Measure and Cosine Distance are used to evaluate intra-class diversity, while Frechet Inception Distance is used to evaluate the quality of synthetic images. Assessing these metrics for biomedical and non-biomedical imaging is important to investigate an informed strategy in evaluating the diversity and quality of synthetic images. In this work, an empirical assessment of these metrics is conducted for the Deep Convolutional GAN in a biomedical and non-biomedical setting. The diversity and quality of synthetic images are evaluated using different sample sizes. This research intends to investigate the variance in diversity and quality across biomedical and non-biomedical imaging modalities. Results demonstrate that the metrics scores for diversity and quality vary significantly across biomedical-to-biomedical and biomedical-to-non-biomedical imaging modalities.

**Keywords:** X-rays and Optical Coherence Tomography (OCT), Deep Convolutional GAN (DCGAN), Intra-class Diversity and Quality, Multi-scale Structural Similarity Index Measure (MS-SSIM) and Cosine Distance (CD), Frechet Inception Distance (FID).

## 1 Introduction

In image analysis, data imbalance affects datasets with asymmetrical distribution of samples for classes within a dataset. In the domain of biomedical imaging, this data imbalance is particularly common for classes of rarer diseases and conditions [Qin et al., 2022]. When a model is exposed to an imbalanced distribution during training, it may focus on the majority class and fail to learn important features and patterns from the minority class. Consequently, when the model encounters new, unseen data, it may struggle to classify samples from the underrepresented classes accurately. To address this issue, data augmentation is utilized to generate new samples from existing data. The intent is to improve the model's ability to generalize and therefore, accuracy, thus contributing to better clinical outcomes [Zhao et al., 2021].

Generative Adversarial Networks (GANs) are particularly useful for data augmentation because they can generate synthetic images that contain a similar distribution of diversified features, such as shapes, textures, or colors, compared to real images [Allahyani et al., 2023]. For deep learning models, GANs can increase the diversity of the training dataset and improve classifiers' performance by generating these synthetic images. GANs are popular in the biomedical imaging domain due to their ability to synthesize realistic images from a

\*This work is supported by the Munster Technological University's Risam Scholarship Award

given distribution [Zhao et al., 2021]. GANs consist of two models based on neural networks: a generator and a discriminator, which work together in a game-theoretic setting to produce realistic images. In biomedical image analysis, diverse image features play a significant role in training a model to produce better clinical diagnostic outcomes [Segal et al., 2021]. The architecture of a GAN is designed such that it can learn these diversified features and synthesize them in synthetic images.

The diversity and quality of synthetic images are important considerations throughout the training of GANs. In the context of GANs, diversity refers to the variation and range of generated image samples produced by a GAN model. It indicates a GAN's ability to generate a diverse set of synthetic images that cover different aspects, styles, and variations present in the original training images [Allahyani et al., 2023]. Intra-class diversity refers to the diversity assessment within a single class of images. Similarly, the quality of synthetic images refers to the fidelity, realism, and perceptual similarity of the generated image samples compared to the real data. It indicates a GAN's ability to produce high-quality synthetic image data that closely resembles the distribution of the original training data [He et al., 2020].

The intra-class diversity of synthetic images is evaluated using the Multi-scale Structural Similarity Index Measure (MS-SSIM) [Odena et al., 2017] and Cosine Distance (CD) [Borji, 2019]. The quality of synthetic images is evaluated using the Frechet Inception Distance (FID) [Borji, 2019].

It is critical to assess the intra-class diversity and quality of synthetically generated biomedical images. Biomedical images contain complex and diverse features indicating vital information about the subject. Evaluating these synthetic images should consider examining the diverse and high-quality features across different biomedical imaging modalities.

MS-SSIM, CD, and FID are significantly dependent on the salient features inherent in an imaging modality. These metrics use image features such as textures, luminance, and orientation of objects for quantification of the diversity and quality of synthetic images. Consequently, score values for these metrics vary across different biomedical imaging modalities and non-biomedical images. Furthermore, quantifying the diversity and quality of synthetic images can be impacted by the number of image samples. Therefore, the assessment of MS-SSIM, CD, and FID metrics is important for biomedical and non-biomedical imaging domains. It enables the efficacy of evaluating GAN architectures in generating diversified and high-quality synthetic images.

The contribution of this work is as follows:

- Study the effect of sample size on evaluation metrics used for assessing the diversity and quality of synthetic images, in both non-biomedical and biomedical imaging fields.
- Examine the inconsistency in evaluation metric scores when evaluating the diversity and quality of synthetic images across two biomedical imaging modalities: X-ray and Optical Coherence Tomography (OCT).
- Analyze the variability in evaluation metric scores when assessing the diversity and quality of synthetic images in both non-biomedical and biomedical imaging domains.

## 2 Related Work

Several studies on GANs have been proposed in the biomedical imaging domain, as indicated in Table 1. The majority of these studies have used FID scores to evaluate the quality of synthetic images. Few of which have used MS-SSIM to evaluate the intra-class diversity of synthetic images. Of those that quantify the FID, some studies such as [Qin et al., 2022] and [Tajmirriahi et al., 2022] do not specify the sample size used to evaluate it. Significant variance in FID scores is analyzed for X-ray images, see first four rows of Table 1. It is important to investigate the impact of the FID score, whether it depends on image size, image modality, GAN architecture, or sample size. Furthermore, none of the works reviewed considered whether different sample sizes of synthetic images could impact the evaluation of intra-class diversity and quality for these metrics. Similarly, in the non-biomedical imaging domains, studies reporting advances in GANs have failed to include details relating to the quantification of evaluation metrics, as indicated in Table. 2.

**Table 1: Evaluation of synthetic images using MS-SSIM, CD, and FID scores for assessing the quality and diversity of biomedical images. Sample size values refer to both real and synthetic images.**

Reference	Med. Image Res.	GANs	MS.	CD	FID			Sample Size		
					MS.	CD	FID	MS.	CD	FID
[Saad et al., 2023]	X-ray <b>128x128</b>	MSG-SAGAN	0.5, 0.47	N/A	139.6	3616	N/A	3616		
[Saad et al., 2022]	X-ray <b>128x128</b>	DCGAN	0.5, 0.529	N/A	0.687	1340	N/A	1340		
[Qin et al., 2022]	X-ray <b>256x256</b>	DCGAN	N/A	N/A	293.26	N/A	N/A	N/S		
[Segal et al., 2021]	X-ray <b>256x256</b>	PGGAN	N/A	N/A	8.02	N/A	N/A	0.1 M		
[Tajmirriahi et al., 2022]	OCT <b>128x128</b>	DDFA-GAN	0.17, 0.19	N/A	51.30	N/S	N/A	N/S		
[He et al., 2020]	OCT <b>224x224</b>	LSGAN	N/A	N/A	N/A	N/A	N/A	N/A		
[Segato et al., 2020]	MRI <b>64x64</b>	AEGAN	0.99, 0.60	N/A	N/A	2000	N/A	N/A		

Med: Medical; N/S: Not Specified; MS: MS-SSIM (Real, Synthetic); M: Million

In the literature, the research gap is highlighted in Table 1 and Table 2 to investigate the impact of sample size in evaluating the intra-class diversity and quality of synthetic images for biomedical and non-biomedical imaging domains. Prior studies also lack in demonstrating the variance in MS-SSIM, CD, and FID scores across biomedical-to-biomedical, and biomedical-to-non-biomedical imaging modalities.

**Table 2: Evaluation of synthetic images using MS-SSIM, CD, and FID scores for assessing the quality and diversity of synthetic images in the non-biomedical domain.**

Reference	Img. Type Res.	GANs	MS.	CD	FID			Sample Size		
					MS.	CD	FID	MS.	CD	FID
[Allahyani et al., 2023]	F-MNIST <b>28x28</b>	DivGAN	N/A	N/A	9.809	N/A	N/A	N/A	N/S	
[Sánchez et al., 2023]	CelebA <b>128x128</b>	REGAN	N/A	N/A	36.57	N/A	N/A	5K		
[Lee et al., 2022]	LSUN- Church <b>256x256</b>	StyleGAN2-GGDR	N/A	N/A	3.15	N/A	N/A	50K		
[Yu et al., 2022]	CelebA <b>64x64</b>	HSGAN	N/A	N/A	17.49	N/A	N/A	N/S		
[Zhang et al., 2021]	F-MNIST <b>28x28</b>	TWGAN	N/A	N/A	10.3	N/A	N/A	N/S		
[Zhao et al., 2021]	CelebA-HQ <b>128x128</b>	BigGAN-ICR	N/A	N/A	15.43	N/A	N/A	3K		
[Costa et al., 2020]	CelebA <b>64x64</b>	COEGAN-NSGC	N/A	N/A	100	N/A	N/A	1024		
[Odena et al., 2017]	Hot Dog <b>128x128</b>	ACGAN	0.11, 0.05	N/A	N/A	200	N/A	N/A		

N/S: Not Specified; MS: MS-SSIM (Real, Synthetic); N/A: Not Applied; K: Thousand

### 3 Methodology

In this work, Chest X-ray and OCT images from the MedMNIST dataset [Yang et al., 2023] are used. The MedMNIST dataset contains several 28x28 resolution biomedical image datasets as a replica of the other benchmark datasets such as the F-MNIST dataset [Xiao et al., 2017], Pneumoniannist, a binary class dataset containing normal and Pneumonia images, and OCTmnist, a multiclass dataset containing normal and three disease conditions are used, see Table 3. In the non-biomedical imaging domain, F-MNIST, a dataset containing different wearable products as indicated in Table 4 was used. All images were used with the original 28x28 resolution.

Table 3: **Image distributions of Chest X-ray and Retinal OCT datasets.**

X-ray Images (Pneumoniamnist)				Retinal OCT Images (OCTmnist)				
	Total No. Img.	Normal	Pneumonia	Total No. Img.	Normal	Chor.	Diab.	Drusen
Train	4708	1214	3494	57884	27317	19867	6054	4646
Valid	524	135	389	10832	5114	3721	1135	862
Test	624	234	390	1000	250	250	250	250

Img: Images; Chor: Choroidal Neovascularization; Diab: Diabetic Macular Edema

Table 4: **Image distributions of Fashion MNIST dataset.**

	Total No. Img.	AB	Bg	Ct	Ds	Pr	St	Sl	Sr	T-st	Tr
Train	60,000	6000	6000	6000	6000	6000	6000	6000	6000	6000	6000
Test	10,000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

Total No. Img: Total No. of Images per Class; AB: Ankle Boot; Bg: Bag; Ct: Coat; Ds: Dress; Pr: Pullover  
St: Shirt; Sl: Sandal; Sn: Sneaker; T-st: T-shirt; Tr: Trouser

### 3.1 DCGAN Architecture

DCGAN is a baseline generative model, designed for generating realistic images. The DCGAN architecture implemented for the un-normalized X-ray images in [Saad et al., 2022] was adapted and modified for 28x28 images used in this work. DCGAN architecture comprises two main components: the generator and the discriminator as detailed in Fig. 1. The generator produces synthetic images using a latent input  $z$  of 100 and passes them to the discriminator. The discriminator distinguishes synthetic images from real images and provides gradient feedback to the generator. The generator updates its learning based on the discriminator's gradient feedback to improve the generation of realistic and diverse images. DCGAN is trained for 500 epochs to converge the training to a balanced state with a batch size of 128. The binary-cross-entropy loss was used to evaluate the generator and discriminator performances. DCGAN is used for minority classes such as normal Chest X-ray images from the Pneumoniamnist, Drusen OCT images from the OCTmnist, and all F-MNIST images. DCGAN was trained for each class separately and generated synthetic images accordingly.

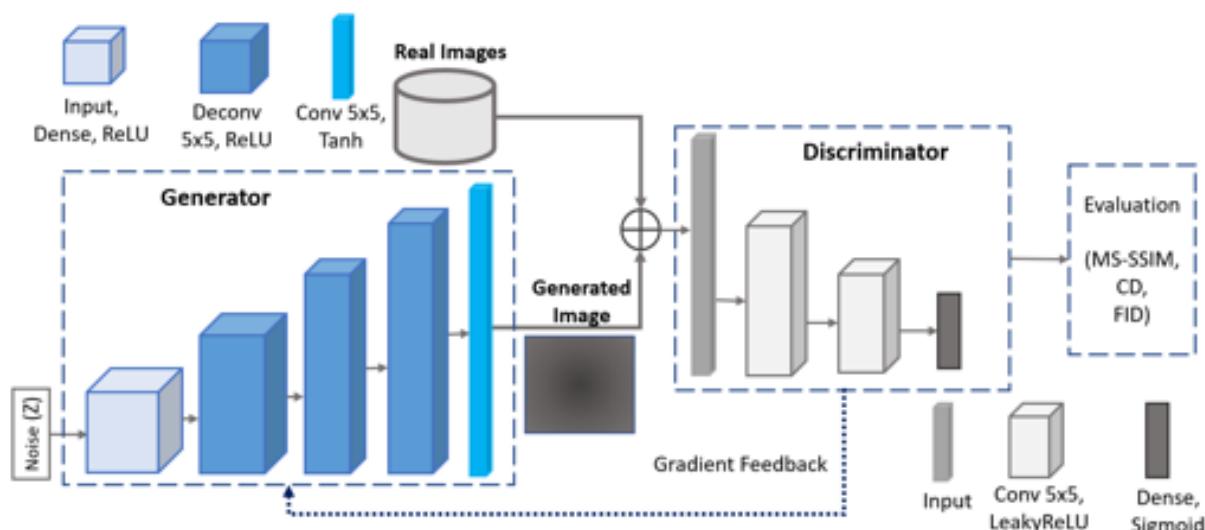


Figure 1: Architecture of DCGAN for synthesizing Chest X-ray, Retinal OCT, and Fashion MNIST images.

### 3.2 Diversity of Synthetic Images

MS-SSIM and CD are used to evaluate the intra-class diversity of synthetic images [Borji, 2019]. MS-SSIM computes the structural similarity at different levels, considering both local and global image features to evaluate the intra-class diversity of synthetic images. A higher MS-SSIM score of synthetic images than real images indicates that synthetic images lack diversity as compared to real images. Conversely, a lower MS-SSIM score of synthetic images indicates better diversity among the synthetic images as compared to real images. CD measures the distance between two feature vectors encoded by images to evaluate the intra-class diversity of synthetic images [Borji, 2019]. A higher CD among synthetic images indicates higher diversity compared to real images. Conversely, a lower CD among the synthetic images indicates limited diversity compared to real images.

### 3.3 Quality of Synthetic Images

FID measures the Wasserstein-2 distance between the two distributions (real and generated) using their mean vectors and covariance matrices to evaluate the quality of synthetic images compared to real images [Borji, 2022]. A lower FID score indicates that the synthetic images are closer in quality and distribution to the real images, indicating higher quality and better resemblance to the real images.

### 3.4 Selecting Sample Size to Assess Intra-class Diversity and Quality of Synthetic Images

In this work, DCGAN has generated synthetic images equal to the number of real images (a 1:1 ratio). The impact of sample size quantifying the intra-class diversity and quality of synthetic images using different sample sizes such as 25%, 50%, 75%, and 100% (all 1:1 ratios) is assessed.

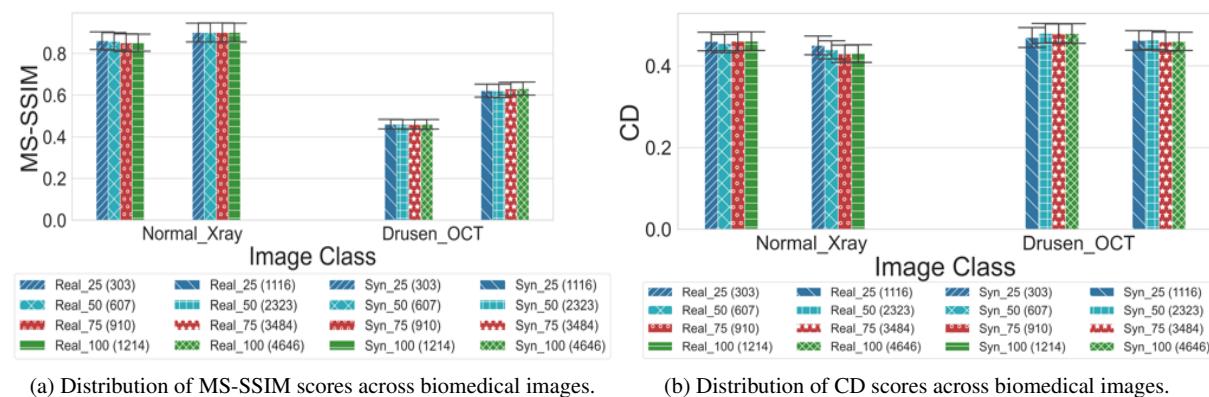


Figure 2: Intra-class diversity of biomedical images using different sample sizes.

## 4 Results and Discussion

### 4.1 Variance in Diversity of Synthetic Images

The intra-class diversity assessment of synthetic biomedical images using the MS-SSIM and CD scores is depicted in Fig. 2. The sample size has no significant impact on evaluating the intra-class diversity in X-ray and OCT images, as indicated by the uniform MS-SSIM and CD scores for different sample sizes. The MS-SSIM and CD scores vary across X-ray and OCT images because the distribution of features in images varies across X-ray and OCT modalities in real or synthetic images.

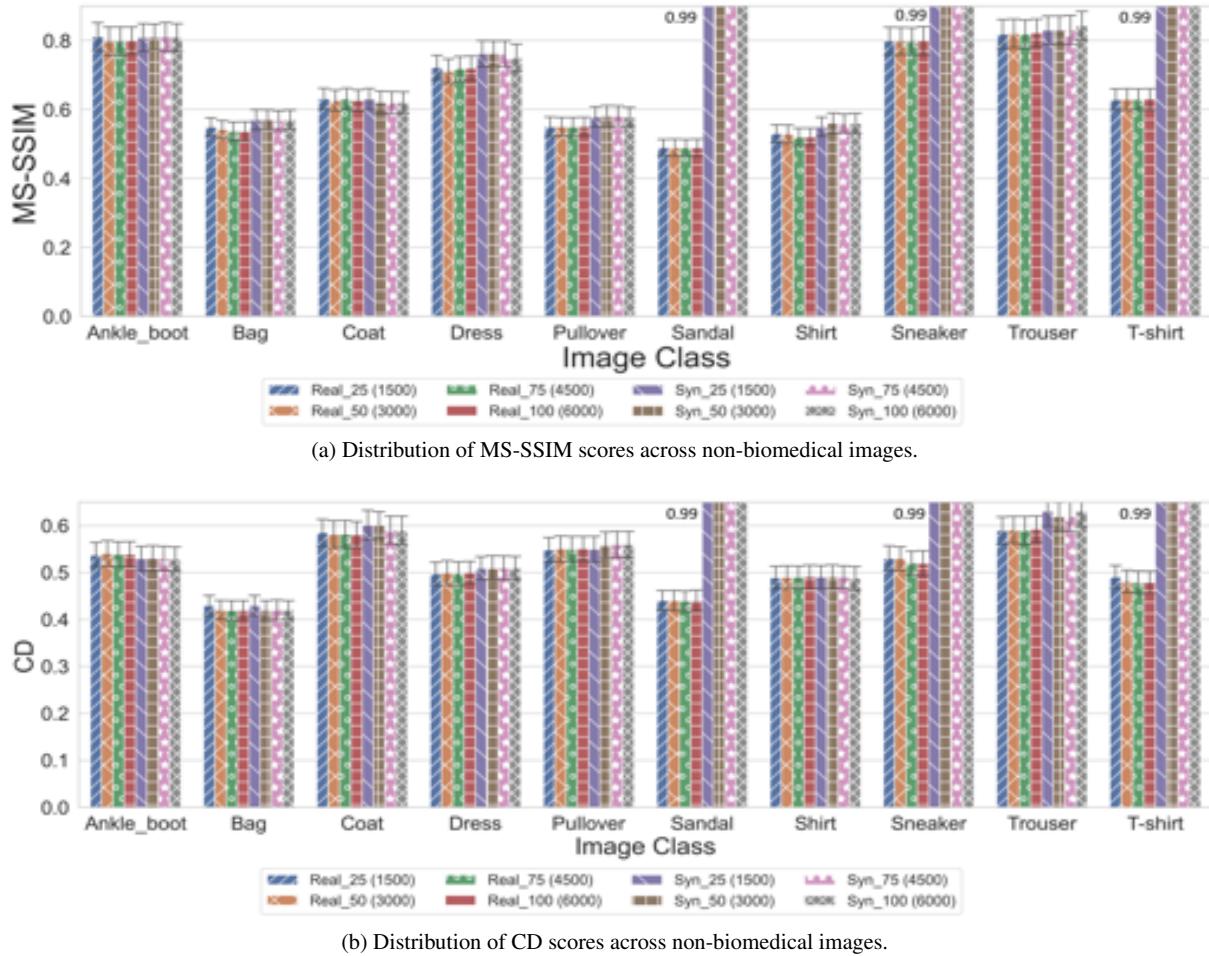


Figure 3: Intra-class diversity of non-biomedical images using different sample sizes.

The better MS-SSIM and CD scores of synthetic X-ray images over real images indicate better intra-class diversity of synthetic X-ray images. The poor MS-SSIM and CD scores of synthetic OCT images over real images indicate poor intra-class diversity of synthetic OCT images. MS-SSIM and CD scores of OCT images worsen over X-ray images because OCT images contain more diverse features than X-ray images, which are difficult to learn and train with the DCGAN. It also indicates that the architecture of GANs has a significant impact on the generation of synthetic images across different domains of biomedical imaging.

For non-biomedical images, the MS-SSIM and CD analyses of real and synthetic Fashion MNIST images indicate significant variation across different classes, as depicted in Fig. 3. Each class has distinct images with unique features that impact the learning and training of DCGAN architecture. The assessment of different sample sizes for evaluating the MS-SSIM and CD scores indicates that there is no significant impact of sample size in evaluating these metric scores for intra-class diversity assessment. The analysis of these metric scores also indicates that the distribution of MS-SSIM and CD scores also varies across biomedical and non-biomedical images.

#### 4.2 Variance in Quality of Synthetic Images

The assessment of the quality of synthetic images using the FID scores across biomedical images and non-biomedical images is indicated in Fig. 4. The sample size of real and synthetic images has no significant impact on the FID scores in evaluating the quality of synthetic images across both biomedical and non-biomedical

images as indicated by the uniform FID scores for different sample sizes. FID scores of normal X-ray, Drusen OCT, and Fashion MNIST images vary due to the distinct distribution of features in images of each biomedical and non-biomedical image class. FID has a different range of values across different biomedical and non-biomedical modalities. FID score can be higher or lower based on a specific biomedical image modality indicating the impact of salient features inherited in representative images.

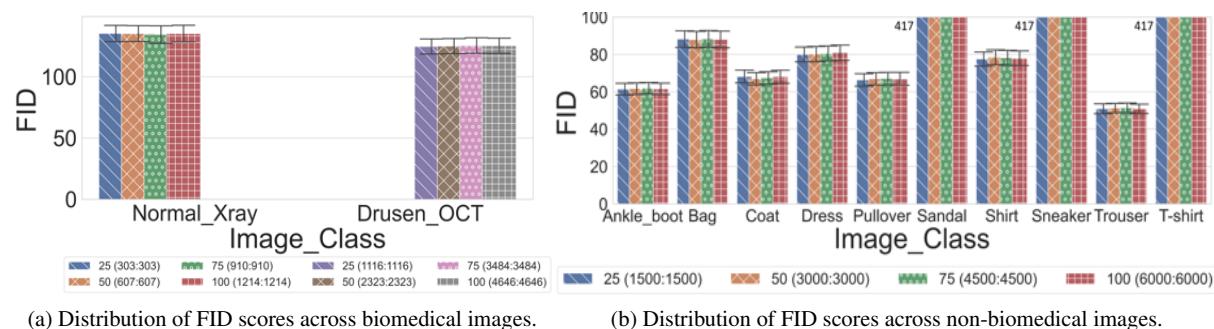


Figure 4: The quality of biomedical and non-biomedical images using different sample sizes.

## 5 Conclusion

This work concludes that the intra-class diversity using the MS-SSIM and CD scores while the quality of synthetic images using FID scores vary across different biomedical imaging modalities due to the distribution of diverse and unique image features across different imaging modalities.

Similarly, intra-class diversity and quality also vary across biomedical and non-biomedical imaging domains due to the distinct distribution of image features. Furthermore, the sample size has no significant impact on evaluating the intra-class diversity and quality of biomedical and non-biomedical images. One possible reason could be that synthetic images may exhibit limited diversity and quality as compared to real images. Therefore, varying the sample size does not impact the scores of evaluation metrics. Another reason could be a small image size, whereby varying the sample size does not impact the metric scores.

The minimum sample size should be a few hundred images of the real and synthetic images to measure the MS-SSIM, CD, and FID metrics. This research work also shows that the generation of synthetic images using DCGAN is dependent upon the nature of images to produce diversified and high-quality images across different imaging modalities in biomedical and non-biomedical domains.

## References

- [Allahyani et al., 2023] Allahyani, M., Alsulami, R., Alwafi, T., Alafif, T., Ammar, H., Sabban, S., and Chen, X. (2023). Divgan: A diversity enforcing generative adversarial network for mode collapse reduction. *Artificial Intelligence*, page 103863.
- [Borji, 2019] Borji, A. (2019). Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65.
- [Borji, 2022] Borji, A. (2022). Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329.
- [Costa et al., 2020] Costa, V., Lourenço, N., Correia, J., and Machado, P. (2020). Exploring the evolution of gans through quality diversity. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 297–305.

- [He et al., 2020] He, X., Fang, L., Rabbani, H., Chen, X., and Liu, Z. (2020). Retinal optical coherence tomography image classification with label smoothing generative adversarial network. *Neurocomputing*, 405:37–47.
- [Lee et al., 2022] Lee, G., Kim, H., Kim, J., Kim, S., Ha, J.-W., and Choi, Y. (2022). Generator knows what discriminator should learn in unconditional gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 406–422. Springer.
- [Odena et al., 2017] Odena, A., Olah, C., and Shlens, J. (2017). Conditional Image Synthesis with Auxiliary Classifier GANs. In *International conference on machine learning*, pages 2642–2651. PMLR.
- [Qin et al., 2022] Qin, X., Bui, F. M., Nguyen, H. H., and Han, Z. (2022). Learning from limited and imbalanced medical images with finer synthetic images from gans. *IEEE Access*, 10:91663–91677.
- [Saad et al., 2022] Saad, M. M., Rehmani, M. H., and O'Reilly, R. (2022). Addressing the intra-class mode collapse problem using adaptive input image normalization in gan-based x-ray images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2049–2052. IEEE.
- [Saad et al., 2023] Saad, M. M., Rehmani, M. H., and O'Reilly, R. (2023). A self-attention guided multi-scale gradient gan for diversified x-ray image synthesis. In *Artificial Intelligence and Cognitive Science: 30th Irish Conference, AICS 2022, Munster, Ireland, December 8–9, 2022, Revised Selected Papers*, pages 18–31. Springer.
- [Sánchez et al., 2023] Sánchez, P., Olmos, P. M., and Perez-Cruz, F. (2023). Enhancing diversity in gans via non-uniform sampling. *Information Sciences*, 637:118928.
- [Segal et al., 2021] Segal, B., Rubin, D. M., Rubin, G., and Pantanowitz, A. (2021). Evaluating the clinical realism of synthetic chest x-rays generated using progressively growing gans. *SN Computer Science*, 2(4):1–17.
- [Segato et al., 2020] Segato, A., Corbetta, V., Di Marzo, M., Pozzi, L., and De Momi, E. (2020). Data augmentation of 3D brain environment using Deep Convolutional Refined Auto-Encoding Alpha GAN. *IEEE Transactions on Medical Robotics and Bionics*.
- [Tajmirriahi et al., 2022] Tajmirriahi, M., Kafieh, R., Amini, Z., and Lakshminarayanan, V. (2022). A dual-discriminator fourier acquisitive gan for generating retinal optical coherence tomography images. *IEEE Transactions on Instrumentation and Measurement*, 71:1–8.
- [Xiao et al., 2017] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.
- [Yang et al., 2023] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.
- [Yu et al., 2022] Yu, S., Zhang, K., Xiao, C., Huang, J. Z., Li, M. J., and Onizuka, M. (2022). Hsgan: Reducing mode collapse in gans by the latent code distance of homogeneous samples. *Computer Vision and Image Understanding*, 214:103314.
- [Zhang et al., 2021] Zhang, Z., Li, M., Xie, H., Yu, J., Liu, T., and Chen, C. W. (2021). Twgan: Twin discriminator generative adversarial networks. *IEEE Transactions on Multimedia*, 24:677–688.
- [Zhao et al., 2021] Zhao, Z., Singh, S., Lee, H., Zhang, Z., Odena, A., and Zhang, H. (2021). Improved consistency regularization for gans. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11033–11041.

# Facial Camera-Based Heart Rate Estimation Using r-PPG Convolutional Neural Networks

Mohamed Moustafa<sup>1,2</sup>, Joseph Lemley<sup>2</sup>, and Peter Corcoran<sup>1,2</sup>

<sup>1</sup>*School of Engineering, University of Galway, Galway, Ireland*

<sup>2</sup>*Xperi Corporation, Galway, Ireland*

## Abstract

The human heart plays an essential role in maintaining an individual's well-being. Therefore, monitoring heart behaviour and condition is important as it provides insights into various physical and psychological conditions. As it is not always convenient to attach sensors to an individual, remote heart signal estimation has become a widely popular field of study over the past two decades. This is commonly achieved by monitoring and extracting the remote photoplethysmography signal from the subject's face, followed by signal filtering and heart rate calculation. Recently, interest in heart rate estimation using supervised deep networks has risen as they have demonstrated better results compared to unsupervised computer vision techniques. This paper aims to explore the limitations of the conventional method of using a loss function to assess the accuracy of models in predicting heart rate from face videos. We present the findings of our study, where we trained and tested three state-of-the-art deep neural networks using publicly available datasets. Our results reveal a significant divergence between model loss and heart rate accuracy.

**Keywords:** Imaging, Computer Vision, Deep Learning, Convolutional Neural Network, Heart Rate.

## 1 Introduction

The human body is composed of various systems that enable individuals to function optimally in various aspects of their lives. While some medical conditions necessitate specialized and invasive procedures with costly equipment for examination, research has shown that several vital health metrics can be measured outside traditional healthcare settings, offering a comprehensive overview of an individual's health and well-being [Chen and McDuff, 2018, Cardone et al., 2020]. One such critical metric is heart rate, which provides valuable insights into an individual's physiological and psychological state.

In light of this, the implementation of remote heart rate sensing has become increasingly significant wherever contact-free sensors are available, such as vehicle in-cabin systems. This involves continuous monitoring of a specific region of interest (ROI), typically the face, to extract a heart signal, which is then processed to estimate the heart rate. While initial attempts of remote heart rate estimation relied on unsupervised computer vision image processing and analysis, recent advancements in supervised deep learning methods, particularly Convolution Neural Networks (CNNs), have demonstrated significantly improved results in this field [Liu et al., 2022]. These models are trained to predict the heart signal using face video or frames as input features by continuously adjusting the model's internal parameters (weights) to minimize the distance between predicted and ground truth values of the heart signal.

Typically, evaluating model performance relies on using the loss, which is a measure of the distance between the model output and the ground truth. The two mostly used methods of selecting which weights to save or deploy is by either using the last set of weights after training, or by finding the weights which give the lowest loss on the validation set (which is a portion of the training set used to monitor for issues such as over fitting). However, for models trained to estimate heart signal, the loss is not an accurate method to find out the best set of weights for heart rate estimation.

In this paper we explain the rationale behind this claim. As our main contribution, we present experimental data from three supervised deep models trained on the Pulse Rate Detection (PURE) [Stricker et al., 2014] and tested on the University Bourgogne Franche-Comte (UBFC) [Bobbia et al., 2019] public datasets to support the proposed claim. We compare our results to those presented in [Liu et al., 2022], using similar workflow for fair comparison, and show the reduced correlation between model loss and heart rate error, as compared with the correlation between validation and test set losses.

The structure of this paper is as follows: in section 2, we present background literature review of deep learning and remote heart rate estimation, with special focus on the deep learning models we use in this study. Section 3 will discuss the methodology we followed starting from the data used until model training. Section 4 will display the results obtained, explain how the models were evaluated, then discuss what conclusions are to be drawn from the results reached. Finally, section 5 will summarise the contributions of this paper and how they were reached.

## 2 Background

**Deep Learning:** Heart rate estimation has seen advancements through the application of deep learning techniques. Deep learning, a subset of machine learning, has shown promise in training mathematical models to make predictions and perform tasks using data [Zhang, 2020]. Deep neural networks (DNNs), a key component of deep learning, have gained attention for their ability to extract features at different levels of abstraction, eliminating the need for manual feature engineering [LeCun et al., 2015]. This characteristic makes DNNs well-suited for heart rate estimation tasks, as they can directly process inputs without the requirement of hand-crafted features. Training supervised DNNs involves the use of a loss function to calculate the distance between predictions (i.e., outputs) and ground truth labels, the gradients of the calculated loss are then used to optimize the internal model parameters; this process is referred to as back propagation [LeCun et al., 2015].

Simple DNNs, however, are not well-suited to dealing with certain types of inputs such as images, represented in a computer as a 3D tensor of the shape: (height x width x channel number), channel number being the number of values each pixel in an image has. Classic techniques involved flattening images into a single row of values then inputting those to the DNN. A special class of DNNs, CNNs, were developed to process image inputs by replacing the weighted sum with element-wise multiplication with a kernel followed by summation then some form of sub-sampling. This approach allowed patches of images to be processed together and drastically lowered the number of weights needed, as they were no longer dependant on the input size, but had a fixed number instead i.e., the kernel of shared parameters. Additionally, it allows the model to leverage spatial locality as patches of the image are processed together [LeCun et al., 2015].

**Unsupervised heart signal estimation:** Initial proof of concept for remote heart rate estimation has demonstrated that the photoplethysmography (PPG) signals, which measures blood oxygenation by monitoring how the light reflective properties of blood vary over time due to fluctuation in blood oxygen content, can be remotely measured on the human face using simple digital consumer cameras [Verkruyse et al., 2008]. This was done by extracting the heart signal from the video's green channel after spatially averaging the RGB video. This concept was further developed in [Poh et al., 2010a], where the facial region was detected from the RGB video then the three channels were separated. Following that, each channel is spatially averaged, and the resulting signal is de-trended and normalised. Independent component analysis (ICA) was applied on the three signals, each obtained from one channel, to separate three independent sources. The PPG signal was most visible in the second source signal. In [Monkaresi et al., 2013], the authors use a similar method to that in [Poh et al., 2010a], but, after ICA, a power spectrum analysis is carried out followed by one of two possible machine learning techniques: k-nearest neighbour and linear regression. The kNN approach proved to be much better at drastically reducing the error.

**Supervised deep learning heart signal estimation:** Deep learning methods have seen a sharp rise in utilization for the task of heart rate estimation. One of the two main approaches used is to train the model to estimate the PPG signal from facial images then calculate heart rate during the output post-processing phase of the pipeline.

DeepPhys, an end-to-end model, is used in [Chen and McDuff, 2018] for measurement of heart and breathing rate from video using a CNN. The model architecture presented consists of two branches, motion and appearance branches, each made up of several convolutional layers. The motion branch takes in normalized frame differences. Its purpose, in a manner similar to Imaging Ballistocardiography [Balakrishnan et al., 2013], is to extract motion information as minute body motions, resulting from the mechanical flow of blood, provide complementary cardiac information to the PPG signal. The appearance branch takes in frame averages and its main purpose is learning spatial masks which are then sent to the motion branch. This is done to reduce the effect of illumination and other external factors, acting as a form of attention mechanism.

Another end-to-end model of interest is the PhysNet architecture [Yu et al., 2019]. The authors present two possible implementations of this network, the first one consisting of 3D convolutional layers, where several frames are processed simultaneously, and the second uses traditional 2D convolutions followed by a recurrent neural network (RNN), where the output obtained from the previous frame is used in calculating the output based on the next frame. Overall the 3D convolution implementation displayed significantly better results than the 2D RNN implementation. This model is of interest as the authors not only use the model output for heart rate estimation, but they also evaluate the prediction accuracy for other metrics such as heart rate variability and atrial fibrillation detection.

The work proposed in [Liu et al., 2020] builds further upon the concepts presented in [Chen and McDuff, 2018] by introducing several convolutional attention networks (CANs) of a similar structure to the DeepPhys model. 3D-CAN uses 3-dimensional convolutions in both branches, hybrid-CAN uses them only in the motion branch, and the temporal-shift-CAN (TS-CAN) uses regular convolutions but applies a temporal-shift (TS) before each convolutional layer in the motion branch. The TS function allows for temporal information to be exchanged between frames [Lin et al., 2019] thus allowing classical convolutions access to temporal information. Additionally, this approach allows for inference time to drop to 25% of that of 3D-CAN while only slightly reducing the performance.

### 3 Methodology

During the training process, it is common practice to save the weights that result in the lowest loss on a validation set, which is a subset of the training data. However, in the case of estimating heart signals, the loss metric used may not accurately reflect the loss in heart rate estimation, as these are distinct tasks. Our objective is to question the conventional method of selecting the best weights for deployment. To achieve this, we train multiple models on publicly available data, saving the weights after each training epoch. We then calculate the heart signal loss on the validation and test sets as well as the heart rate root loss for each set of weights. Next, we compare the performance of models selected based on heart signal loss with those selected based on heart rate loss. In the subsequent section, we present the findings of our study, which indicate that a low overall signal loss does not necessarily correspond to low heart rate estimation error. Additionally, we demonstrate that there exists a weaker correlation between the signal losses and the heart rate loss, further emphasizing the need to consider alternative evaluation metrics when selecting the best-performing models for heart rate estimation.

Our experiments utilise the rppg-toolbox repository developed by [Liu et al., 2022]. For a fair comparison we follow the same training steps while saving the model weights after each epoch, then we use our own approach to find the best set of weights. With regards to specific hyperparameters and feature shapes, we follow the default configurations provided by [Liu et al., 2022] in their implementation.

### 3.1 Dataset

As we wanted to compare our results with the state-of-the-art, following the work presented in [Liu et al., 2022], we use the Pulse Rate Detection (PURE) [Stricker et al., 2014] and the University Bourgogne Franche-Comte (UBFC) [Bobbia et al., 2019] public heart rate datasets as separate train and test datasets.

**PURE:** The PURE dataset contains acquisitions of 10 subjects (8 male and 2 female) performing different, controlled head motions. Six one-minute acquisitions were carried out per subject, the videos were captured at a frame rate of 30 Hz with a cropped resolution of 640x480 pixels. The heart signal ground truth data was simultaneously collected using a finger clip pulse oximeter at a sampling rate of 60 Hz. The setups for each acquisition were: Steady, talking, slow translation, fast translation, small rotation, medium rotation.

**UBFC:** The UBFC dataset contains data of 42 subjects sitting in front of the camera playing a time sensitive mathematical game (aimed at augmenting heart rate) while simultaneously emulating a normal human-computer interaction scenario. The video acquisition is carried out at 30fps with a resolution of 640x480. A pulse oximeter was used to obtain the ground truth data.

### 3.2 Data Pre-processing

As the PURE dataset frame rate and sensor sampling rate are mismatched, we downsample the heart signal to 30Hz using interpolation. Following that, a Haar Cascade face detector is used on all the frames to locate the subject's face and a square crop is extracted. The crop is resized to 72x72 for DeepPhys and TSCAN and 128x128 for PhysNet. Following that, the input feature (i.e. frames) representation is generated by calculating the normalized frame differences (differences between consecutive frames) and concatenating that to the standardized raw frames along the channel dimension to create a NumPy array of the shape [N, W, H, C] where N is the length of the sequence, W is the width of the frames, H is the height of the frames, and C is the channels (with a value of 6). For faster data loading, each video in the datasets is typically broken up into several “chunks” of non-overlapping 180 frame sequences (for PhysNet the chunk size is 128 instead as the model only supports factors of 512; this will be further explained below). The PPG waveforms (labels) are stored as numpy arrays in a [1, N] format [Liu et al., 2022].

### 3.3 Deep Neural Networks

**DeepPhys:** a 2D CNN that consists of two branches, motion and appearance branches, each made up of several convolutional layers. The motion branch takes in normalized frame differences. while the appearance branch takes in frame averages. However, the implementation provided by [Liu et al., 2022] utilizes raw frames instead of frame averages. The appearance branch generates an attention mask which is then fed to the motion branch to emphasize regions of the face with high physiological information.

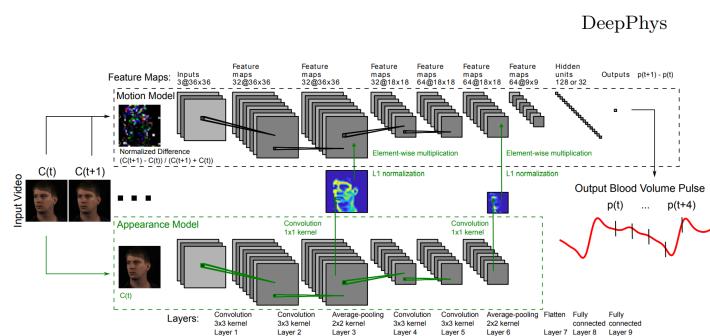


Figure 1: DeepPhys Architecture

**TSCAN:** An advanced implementation of DeepPhys where each of the main convolutional layers in the motion branch is preceded by a temporal shift function which performs a pixel shift across a certain number of frames (referred to as frame depth). As each frame represents a slice of time, this operation allows regular 2D convolutional to access temporal information without needing to use computationally-heavy 3D convolutions thus allowing for real-time applications.

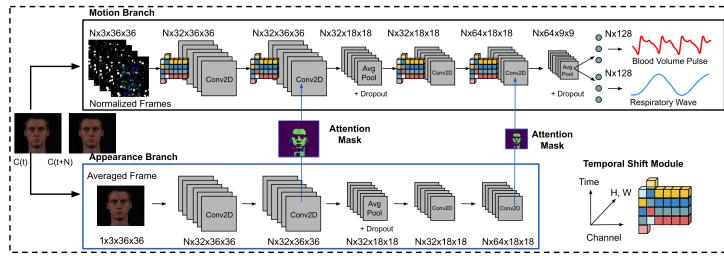


Figure 2: TSCAN Architecture

**PhysNet:** This model leverages 3D convolutions as part of a traditional feed-forward CNN. Unlike with DeepPhys, where the architecture complexity would mean that using 3D convolutions would prevent real-time applications, PhysNet's 3D implementation is deemed by the authors as real-time compatible. And while they present a 2D convolutional RNN alternative implementation, the significant performance improvement demonstrated by the first implementation is why it is the implementation used in the rppg-toolbox as well as the presented work.

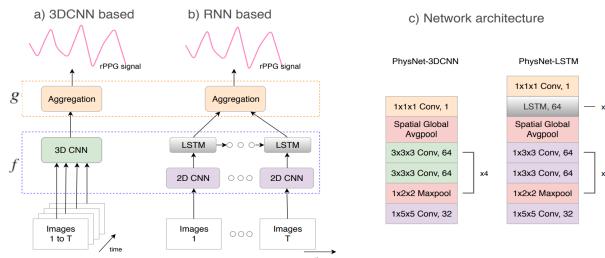


Figure 3: PhysNet Architecture

### 3.4 Model Training

While the authors in [Liu et al., 2022] present the results when training on both PURE and UBFC, the results shown in the second scenario were all beyond the accepted error range for remote heart rate estimation (< 5 bpm RMSE [Poh et al., 2010b]). Therefore, we choose to focus on the first case, training on PURE and testing on UBFC.

We trained all three architectures on 80% of the PURE dataset and used 20% as a validation set. We used a batch size of 4 for all models, however, in this implementation, each chunk is considered a single "item", meaning that batch of size 4 contains 720 frames (4 x 180); or in the case of PhysNet 512 frames (4 x 128). For the DeepPhys and TSCAN models we use a frame depth of 10.

The models are all trained for 30 epochs using the mean square error loss function as well as the AdamW optimizer [Loshchilov and Hutter, 2017]. The one cycle learning rate policy [Smith and Topin, 2019], which anneals the learning rate from an initial learning rate to some maximum learning rate and then from that maximum learning rate to some minimum learning rate much lower than the initial learning rate, was used. The maximum learning rate chosen was the initial learning rate itself, 9e-3, which means that the scheduler essentially acts as learning rate decay policy. After each epoch, the model weights are saved, losses and accuracy are then calculated after training is complete.

## 4 Experimental Results

After training is finished, we calculate the validation set signal loss, test set signal loss, and test set heart rate loss. Following that, we calculate the test set heart rate loss for the model weights that gave the lowest validation set and test set signal losses and compare the heart rate accuracy for all three set of weight as well as with the results reported in [Liu et al., 2022] for all three architectures.

### 4.1 Heart Rate Error

To calculate heart rate error, the model outputs were concatenated then de-trended. The output signal is then filtered using a Butterworth bandpass filter with cutoff frequencies of [0.75, 2.5] Hz. The average ground truth and predicted beat per minute (bpm) heart rates are calculated for each subject using the Fast-Fourier Transform (FFT) power spectrum analysis. These bpms are then used to calculate the overall model RMSE.

### 4.2 Results

We compare the results obtained using the best set of weights as defined by our approach, the validation loss, and the test loss, as well as the results reported by the author [Liu et al., 2022] in table 1.

For both DeepPhys and TSCAN models, we are only able to obtain results similar to those reported in [Liu et al., 2022] using our approach. In fact, following the two evaluation methods mentioned in the provided implementation (validation loss and last epoch), the results obtained are quite worse than the reported results.

PhysNet presented the most interesting results, using losses to find the "best" weights, the results we got were fairly close to the reported results (albeit beyond the acceptable range). However, using our approach, we found a set of weights that reduced the RMSE by a factor of 4, bringing it to near-medical accuracy.

Figure 4 compares the normalized model signal losses and heart rate loss per epoch for each of the models. At first glance, it is visible that validation and test signal losses have a strong correlation. This, however, is not applicable to the heart rate test error. For DeepPhys, the heart rate loss continuously fluctuates regardless of the downward trends exhibited by both signal losses. For the TSCAN model, accuracy heavily fluctuates within the first 5 epochs then remains fairly stable except at epoch 12. While it might be assumed this is correlated to the increase in both losses, we don't observe this behaviour at the other spikes in loss such as epochs 14 and 10. For PhysNet, the accuracy fluctuates quite rapidly, less so between epochs 15 to 20, while showing more sensitivity to changes in test signal loss.

This is further supported by the correlation coefficients presented in table 2 below. For both CAN models, the correlation coefficient of the losses is 20-70% higher than the correlation between the signal losses and heart rate RMSE. For PhysNet, the signal loss correlation is a bit lower, but the signal loss and heart rate RMSE correlation show proportional drop.

	Our Approach	Using Validation Loss	Using Test Loss	Last Epoch	Reported
DeepPhys	2.49	2.98	5.72	2.98	2.53
TSCAN	2.40	3.04	2.96	2.97	2.41
PhysNet	1.04	5.52	5.10	6.99	4.49

Table 1: Beat per minute RMSE comparison of our results and results reported in the rppg-toolbox paper

	Valid Loss and Test Loss	Valid Loss and HR RMSE	Test Loss and HR RMSE
DeepPhys	0.90	0.58	0.53
TSCAN	0.94	0.62	0.79
PhysNet	0.69	0.38	0.39

Table 2: Pearson product-moment correlation coefficients of losses and HR error for the three models

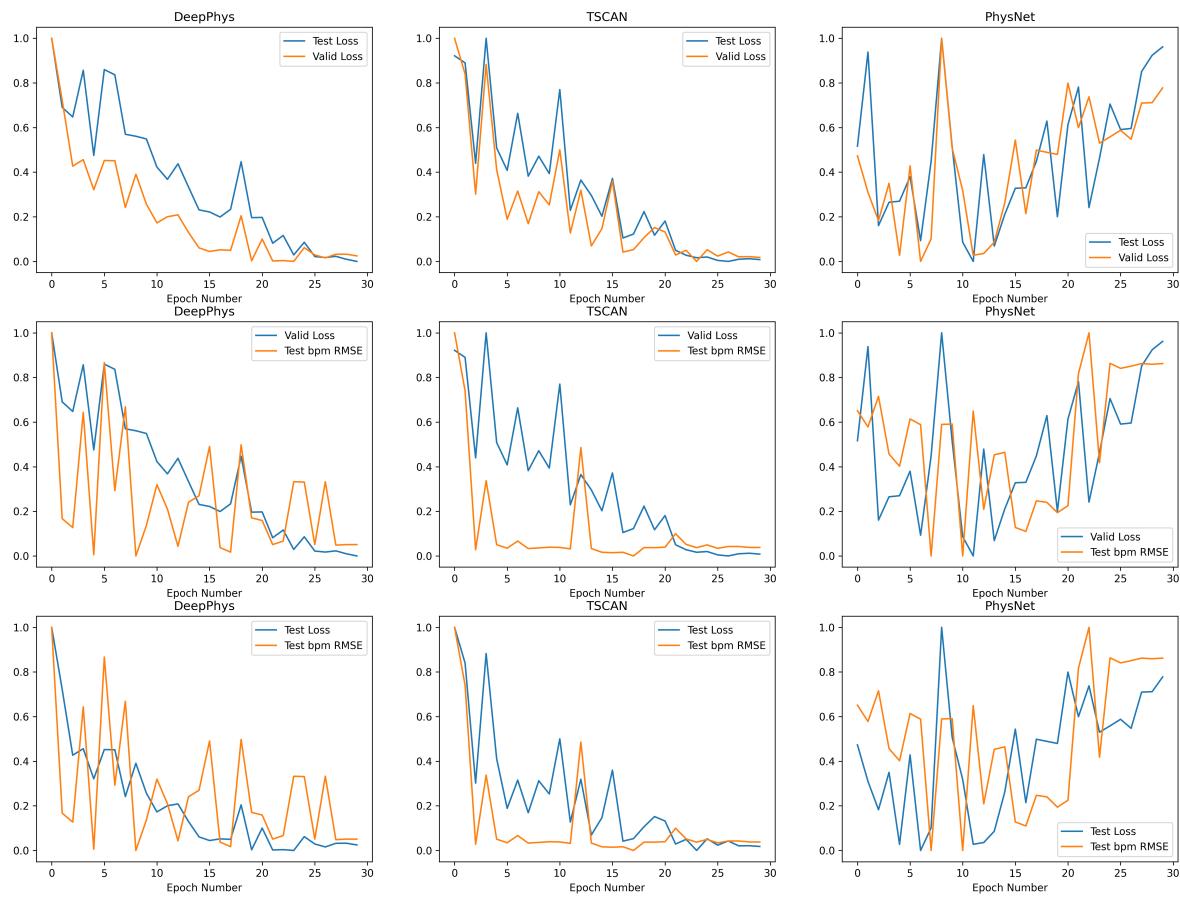


Figure 4: Comparison of losses and accuracy for all models

## 5 Conclusion

In this paper we present the hypothesis that the loss calculated directly from an video-based r-PPG supervised deep model, typically used for model evaluation, does not necessarily measure how well that model estimates the heart rate, given that those two are distinct tasks. Instead, the beat per minute root mean square error (or some other error metric) should be calculated for each set of weights saved during training to find the best model. We demonstrated how using this approach can allow researchers to find the best set of weights for three different state-of-the-art supervised deep learning networks trained and tested on publicly available data. Additionally, we showed that, across all training epochs, heart signal and heart rate losses do not have a strong correlation.

The results shown highlight the need to use error metrics crafted for the task the deployed model is meant to solve when selecting the model weights to utilize. It, however, does not discourage the use of a different error metric for training, as all the models presented in this study were trained using signal loss yet manage to demonstrate impressive performance once the right set of weights is found.

## 6 Acknowledgment

The research conducted in this publication was funded by the Irish Research Council under project ID EBPPG/2021/92 as a part of the Employment-Based Programme Postgraduate Scholarship in partnership with the Xperi Corporation.

## References

- [Balakrishnan et al., 2013] Balakrishnan, G., Durand, F., and Guttag, J. (2013). Detecting pulse from head motions in video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3430–3437.
- [Bobbia et al., 2019] Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., and Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90.
- [Cardone et al., 2020] Cardone, D., Perpetuini, D., Filippini, C., Spadolini, E., Mancini, L., Chiarelli, A. M., and Merla, A. (2020). Driver stress state evaluation by means of thermal imaging: A supervised machine learning approach based on ecg signal. *Applied Sciences*, 10(16):5673.
- [Chen and McDuff, 2018] Chen, W. and McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365.
- [LeCun et al., 2015] LeCun, Y., Bengio, Y., Hinton, G., et al. (2015). Deep learning. *nature*, 521 (7553), 436-444. [Google Scholar](#) [Cross Ref](#)
- [Lin et al., 2019] Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093.
- [Liu et al., 2020] Liu, X., Fromm, J., Patel, S., and McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411.
- [Liu et al., 2022] Liu, X., Zhang, X., Narayanswamy, G., Zhang, Y., Wang, Y., Patel, S., and McDuff, D. (2022). Deep physiological sensing toolbox. *arXiv preprint arXiv:2210.00716*.
- [Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- [Monkaresi et al., 2013] Monkaresi, H., Calvo, R. A., and Yan, H. (2013). A machine learning approach to improve contactless heart rate monitoring using a webcam. *IEEE journal of biomedical and health informatics*, 18(4):1153–1160.
- [Poh et al., 2010a] Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2010a). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11.
- [Poh et al., 2010b] Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2010b). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774.
- [Smith and Topin, 2019] Smith, L. N. and Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.
- [Stricker et al., 2014] Stricker, R., Müller, S., and Gross, H.-M. (2014). Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE.
- [Verkruyse et al., 2008] Verkruyse, W., Svaasand, L. O., and Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445.

[Yu et al., 2019] Yu, Z., Li, X., and Zhao, G. (2019). Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*.

[Zhang, 2020] Zhang, X.-D. (2020). A matrix algebra approach to artificial intelligence.

# A QoE and Visual Attention Evaluation on the Influence of Spatial Audio in 360° videos

Amit Hirway, Yuansong Qiao, Niall Murray

*Technological University of the Shannon, Athlone, Ireland*

*{a.hirway, ysqiao, nmurray}@research.ait.ie*

## Abstract

This paper presents a summary of an ongoing doctoral research focused on evaluating the quality of experience (QoE) and visual attention aspects of spatial audio in 360° videos. With the increasing popularity of virtual reality (VR) applications, it is crucial to understand the impact of spatial audio on user perception and engagement. The research investigates the influence of spatial audio on visual attention patterns within immersive 360° video environments in addition to subjective and objective measures of QoE. This paper provides an overview of the research objectives, experimental setup, methodology, overview of findings based on the analysis of subjective, objective and intrinsic data, and future directions.

**Keywords:** 360° videos, Ambisonics, Visual attention, Quality of Experience

## 1 Introduction

In recent years, the popularity of VR applications, particularly 360° videos, has significantly increased. These videos are experienced through Head Mounted Display (HMD) technology, offering users a more immersive and interactive viewing experience. However, streaming 360° videos to HMDs presents numerous challenges, including limited bandwidth, storage, and computational resources. Additionally, factors such as low motion-to-photon delay expectations, complex view adaptation, understanding of 360° video Quality of Experience (QoE) [Shojib et al., 2021], and the influence of visual attention [Huang and Wang, 2020] on user perception need to be understood.

Visual attention refers to the selective process through which our brains focus on specific information from our environment. It enables efficient extraction of relevant information and understanding of our surroundings. Various techniques have been developed for visual attention modeling, incorporating saliency feature vectors, object-based attention, and salient maps. Understanding visual attention is crucial for optimizing video content and delivery, ultimately enhancing the user experience.

While visual attention has received considerable research attention, the impact of audio, particularly spatial audio, on visual attention in immersive media has been relatively overlooked [Kim et al., 2020]. Current research predominantly focuses on analyzing attention in conventional non-spatial sound videos, neglecting the immersive multimedia context. Spatial audio, which adds depth and directionality to sound, has gained attention for its ability to create a more realistic VR experience. Investigating the role of spatial audio in visual attention is essential for enhancing the overall quality of VR experiences [Xylakis et al., 2020].

Ambisonics, a technique for capturing and reproducing 360° audio, has gained popularity due to its ability to provide an immersive audio experience in VR [Innes and Shellard, 2005]. This technique involves recording sound from all directions and encoding it into a multi-channel format, enhancing the sense of presence and realism. Higher-order Ambisonics formats offer greater spatial resolution and immersion, enabling precise localization of sound sources [Politis et al., 2018].

With the increasing adoption of VR and 360° videos, understanding the Quality of Experience (QoE) has become paramount. QoE encompasses the overall subjective user experience while interacting with multimedia applications and is influenced by factors such as content quality [Raake, 2014], network conditions [Xue et al., 2011], device performance [Singh and Kapoor, 2018], and user expectations [Li et al., 2016]. In the context of VR

and 360° videos, QoE plays a crucial role in determining the level of immersion and presence experienced by users [Wang and Li, 2021]. Investigating QoE provides valuable insights into user preferences [Shahriar and Shiratori, 2015], behavior [Ahmed et al., 2019], and perception [Wang et al., 2016], allowing the design of more effective and engaging applications.

This doctoral study explores the impact of different sound conditions, including no sound, first-order ambisonics(FOA) sound, high-order ambisonics(HOA) sound, and stereo sound, on visual attention and QoE of participants watching 360° videos wearing a VR headset. The study involves analyzing objective (head pose, eye gaze fixations), intrinsic (pupil diameter, heart rate) and subjective data (user questionnaire) to provide insights into the impact of spatial audio on visual attention and QoE in 360° videos.

The remainder of this paper is organized as follows: Section 2 provides an overview of state-of-the-art, Section 3 describes the experimental setup, Section 4 outlines the research methodology, and Section 5 presents details about the dataset. Section 6 presents the statistical analysis of the subjective questionnaire, and finally, Section 7 concludes the paper and discusses future work.

## 2 State of the Art

Numerous studies have examined visual attention in 360° videos, shedding light on the subject over the past decade. [Jean et al., 2017] conducted research on viewer attention in 360° videos and created a dataset containing sensor data and content information. The dataset consisted of ten YouTube videos categorized into three groups: Computer Generated (fast-paced), Natural Image (fast-paced), and Natural Image (slow-paced). Viewer orientations were captured using OpenTrack, an open-source head tracking tool integrated with HMD sensors.

[David et al., 2018] presented a publicly available dataset of head and eye movements recorded during a free-view trial where participants wore VR headsets equipped with eye trackers. The dataset encompassed 360° videos played without audio and included saliency maps and scan paths in addition to the videos. The impact of sound on visual attention in 360° videos was explored by [Chunjia et al., 2014]. They conducted eye-tracking tests with 60 videos under various conditions: videos with audio-video (AV) and videos without audio (V). Their findings suggested that the effect of sound depended on the consistency between visual and audio signals. Interestingly, they discovered that audio had minimal or no influence on visual attention when the sound sources precisely matched the salient objects in the video. However, when the sound sources differed from the salient objects, they tended to attract participants' attention.

[Marighetto et al., 2017] investigated the influence of audio on visual attention in non-360° videos. Their eye-tracking dataset comprised eye positions recorded during four eye-tracking studies. Participants watched videos under different audio conditions (with or without sound) and visual groups (moving objects, landscapes, and faces). The study revealed that audio-visual content consistently resulted in less dispersion of eye gaze compared to visual-only content. The presence or absence of sound influenced the spatial distribution of eye gaze, particularly in the face category. [Wu et al., 2017] introduced a head tracking dataset containing user behaviour patterns in VR applications. The dataset included data from 48 users watching 18 spherical videos across five categories. Users' head movements, directions of focus, and remembered content after each session were recorded. However, the videos were accompanied by non-spatial sound.

Additionally, [Almquist and Almquist, 2018] investigated the viewing behaviours of subjects experiencing various 360° videos from different categories using HMDs. Data related to head orientation and rotation speed was collected through the HMD sensors as the subjects watched the videos. The study found that the distribution of viewing angles heavily depended on the content of the video, with viewers spending considerable time looking at the front in the Static Focus and Rides categories. The Exploration and Moving Focus categories exhibited a nearly linear distribution, with yaw rotation being the most common compared to pitch and roll rotations.

These studies have contributed valuable insights into visual attention and the role of audio in immersive media. However, the novelty of this research lies in its approach to evaluating visual attention and QoE in 360° videos. Specifically, the research utilizes head pose and eye gaze data to assess visual attention. This approach goes beyond analysing only head movements or eye-tracking data to gain insights into the specific elements of the video

that viewers are focusing on. By incorporating eye gaze data alongside head pose data, a deeper understanding of viewer engagement with the content in 360° videos can be achieved. Moreover, the research evaluates QoE through pupil dilation measurements and subjective questionnaires. Pupil dilation is a physiological response that has been linked to cognitive and emotional processing. By analysing pupil dilation, researchers can gain insights into viewers' cognitive and emotional responses to the 360° videos. This comprehensive approach provides a more holistic understanding of the user experience beyond relying solely on traditional subjective questionnaires. Additionally, the research explores the influence of different audio types on visual attention and QoE in 360° videos. This aspect is of significant importance as audio can have a profound impact on the viewer's overall experience of the video.

### **3 Experimental set up and presentation systems**

#### **3.1 Experimental Setup**

This section provides an overview of the experimental setup, including details on the laboratory design, the presentation system used for immersive media, and the techniques employed to capture user head pose, gaze, and physiological data.

##### **3.1.1 Laboratory Design**

The laboratory design followed the recommendations outlined in [ISO 8589, 2007], which provides guidelines for designing test rooms for sensory analysis. Fig. 1 illustrates a test subject participating in the experiment within our laboratory. Table 1 outlines the hardware and software components utilized for the experiment.

##### **3.1.2 Presentation System**

###### **3.1.2.1 Selection of 360-degree videos**

For the experiment, a set of ten 360° videos with FOA and HOA sound were chosen from a pool of recordings available at [Farina, 2020]. The selection criteria considered factors such as video length, content, resolution, and Ambisonics sound order. The videos were divided into two categories: indoor and outdoor, with five videos in each category. Within these categories, the videos were further subcategorized as Opera, Instrument, Riding, and Exploration. To ensure impartial presentation to participants, the videos were randomly stitched together to form two 300-second (60-sec \* 5) segments—one for the indoor category and one for the outdoor category—using the ffmpeg [Ffmpeg, 2021] tool. The selected videos were processed to have a duration of 60 seconds each, stitched together, and the Ambisonics sound was converted to stereo for non-spatial audio experience or removed entirely. The videos did not include any narrative or subtitles.

The five videos in the Indoor category depicted an Opera performance with actors on an elevated platform and an orchestra performing below the platform. The camera remained stationary, positioned between the platform and the orchestra. Each video began with the participant facing the stage where the actors were performing. The other five videos belonged to the Outdoor category and were exploratory in nature, lacking a specific focal point for immediate visual interest. These videos contained sound-emitting objects, some of which were stationary, such as a clock tower or a person playing a musical instrument while seated, while others were in motion, such as people talking while walking or ducks quacking while wading in water. Fig.2 shows some of the representative frames for videos in the Indoor and Outdoor categories.

###### **3.1.2.2 Video presentation**

The GoPro VR player [GoPro, 2020], a free 360° video player, was employed to present the videos on the HTC Vive headset with an integrated Tobii Pro eye tracker [Pro, 2018]. The VR player transmitted 360° video playback

information, including camera orientation, video URL, playback status, and playback position, to a port on the system running the experiment. This allowed viewers to watch the 360° videos from any orientation, with their movements recorded as yaw, pitch, and roll angles (refer to Fig. 3a, b, and c).

### 3.1.2.3 Audio presentation

Although the HTC Vive headset came with an audio strap providing integrated earphones, the Beyerdynamic DT 990 Pro headphones [Beyerdynamic, 2020] were utilized to present the various forms of audio. These headphones feature larger open-back ear cups, which facilitate more natural and clear hearing, as recommended by the manufacturer.

### 3.1.3 Head Pose, Gaze, Pupil Diameter, and Heart-rate Data Acquisition

Python scripts were developed to capture pose, gaze, and pupil diameter data for each frame of the 360° video sequence. The pose data included the yaw, pitch, and roll angles measured in radians. Gaze data recorded the X and Y coordinates of the participant's point of focus on the screen, while pupil diameter data measured the size of the participant's pupils. The collected data was saved in CSV files for each participant, which were subsequently used for data analysis purposes. The integration of the Tobii Pro eye tracker with the HTC Vive headset allowed for simultaneous recording of pose and gaze data, enabling the analysis of the correlation between head movement and visual attention.

To collect heart rate data, we utilized the E4 wristband [Empatica, 2023], a medical-grade wearable device that offers real-time data acquisition and software for in-depth analysis and visualization. Following the official installation guide, we set up the hardware, and the E4 Connect application was used to capture physiological



Figure 1: Participant experiencing the VR environment in our lab



Figure 2: Representative frames for Indoor and Outdoor scenes (Top – Opera Stage, Orchestra; Bottom – Town square with clock tower, Man riding a motorbike with two dogs following)

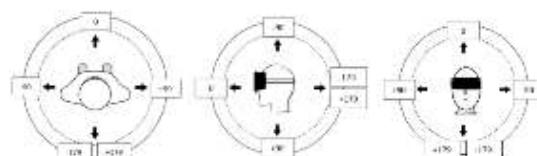


Figure 3a, b, c: Yaw, Pitch and Roll angles adapted from [28]

Component	Vendor/Specifications	Used For
PC	Intel Core™ i5 – 4590 CPU @ 3.30GHz, 10.0 GB RAM, 16GB nVidia GTX 970 Graphics Card, Windows 10	Running the hardware and software for the immersive environment
HMD	HTC Vive with Tobii Pro VR Integration	Watching 360° videos
Headphones	Beyerdynamic DT 990 Pro	Listening to non-spatial/spatial audio
360° Player	GoPro VR Player	Playing 360° videos to the HMD, obtaining head orientation as yaw, pitch and roll
360° videos		Audio-visual presentation to participants
Wristband	E4	Collecting physiological data

Table 1: Experiment setup components

data during the experiment.

Overall, the experimental setup included a controlled laboratory environment designed according to ISO standards. The presentation system employed the GoPro VR player for video playback on the HTC Vive headset, with audio presented through Beyerdynamic headphones. Head pose, gaze, and pupil diameter data were captured using Python scripts and the Tobii Pro eye tracker integrated with the headset. Additionally, heart rate data was recorded using the E4 wristband. These data acquisition techniques allowed for comprehensive analysis of user responses and behaviours during the viewing of 360° videos.

## 4 Methodology

The research methodology employed in this study follows an experimental evaluation approach, drawing inspiration from various sources such as [Keighrey et al., 2017], [Hynes et al., 2019], [Egan et al., 2016], and ITU-T recommendation P.913 [ITU, 2016].

### 4.1 Participants

A convenience sampling approach was used to recruit a total of 73 participants, consisting of 45 men and 28 women, with an average age of 29 years. Among the participants, 26 had prior experience with VR, while 26 had no prior experience. VR experience was not recorded for the remaining 19 participants, who were the initial participants when the study began.

### 4.2 Assessment Protocol

The assessment protocol consisted of five main phases: the information phase (10 minutes), screening phase (10 minutes), training phase (5 minutes), testing phase (10 minutes), and a subjective questionnaire (5-10 minutes).

During the information phase, participants were provided with detailed information about the experiment and had the opportunity to ask questions before signing a consent form. The screening phase involved evaluating participants' visual and auditory acuity, as well as screening for colour perception using tests such as Snellen [Provisu, 2021] and Ishihara [Colblindor, 2021], and an auditory test [Pigeon, 2020]. In the training phase, participants watched a 60-second 360° video to familiarize themselves with the VR environment and underwent a calibration process. The testing phase consisted of participants watching two 300-second, 360° video segments, one recorded indoors and the other outdoors in one of the four sound conditions (no sound, stereo, FOA, or HOA sound). Finally, each participant answered a subjective questionnaire that assessed their perception of presence, immersion, and spatiality of sound after watching the stimuli. The entire assessment took approximately 40-50 minutes on average.

### 4.3 Questionnaire and Rating Scale

A questionnaire comprising twenty questions [Hirway, 2023] was developed to evaluate participants' perception of presence, immersion, and spatiality of sound. Inputs from [Rigby et al., 2019] and [UC Lab, 2004] were considered during the questionnaire development process. The questionnaire utilized the absolute category rating (ACR) system described in [ITU, 2016]. Participants rated each question using a five-point Likert scale, indicating their level of agreement or disagreement with each statement.

## 5 Dataset description

The dataset used in this study consists of ten videos, five in the indoor category and five in the outdoor category. Each video has a duration of 60 seconds, resulting in a total of 600 seconds of video playback for each participant. The videos are numbered and categorized accordingly. Participants watched the videos in a randomized order within

each category, with the indoor sequence played first, followed by the outdoor sequence. Each participant experienced only one sound condition across all the videos, which was randomly selected from options such as no sound, stereo, FOA, or HOA sound.

The dataset is organized into eight folders: foin, fout, hoin, hout, nsin, nsout, stin, and stout. Each folder corresponds to a sound condition in both indoor and outdoor environments. Within each folder, there are three subfolders: gaze data (\_gazedata), heart rate data (\_HR), and pose data (\_posedata). The gaze data includes information about gaze and pupil diameter for both the left and right eyes. The pose data contains yaw and pitch information, representing head movement. The heart rate data captures participants' heart rate during the experiment. The data is stored in CSV files, with file names indicating the type of data, timestamp, and sequence of videos watched by the participant.

The gaze data consists of 146 files, with approximately 36,000 samples in each file, capturing the participants' gaze information throughout the 43,800 seconds of video playback. The heart rate data comprises 73 files, providing information on heart rate in different conditions. The pose data includes 146 files, each containing around 36,000 samples, representing participants' head pose during the video sequences.

We have included a subset of files from the complete dataset to provide readers with representative examples showcasing the data's characteristics and structure. [Hirway, 2023].

## 6 Statistical Analysis of Subjective Questionnaire

An ANOVA was conducted with a 95% confidence level to compare the mean responses between the sound condition groups. The results indicate that sound conditions influenced participants' ability to retain attention, conscious awareness of the real world, perception of experiencing the content, enjoyment of the experience, engagement of senses, and naturalness of interaction. HOA sound consistently received the highest mean responses for these measures, followed by FOA sound and stereo.

The findings from the analysis of the subjective questionnaire [Hirway, 2023] suggest that the presence and nature of sound significantly impact users' immersive experience and quality of interaction while watching the 360° videos. Participants reported higher levels of attention retention, enjoyment, engagement of senses, and perceived naturalness of interaction when exposed to HOA sound compared to other sound conditions. The use of spatially accurate sound, such as HOA, appeared to enhance the overall immersion and presence within the virtual environment. Furthermore, the absence of sound seemed to increase participants' conscious awareness of the real world, indicating that sound can serve as a means of distraction from external surroundings. The perception of experiencing the content, rather than just watching it, was also enhanced by the presence of sound, with participants reporting a greater sense of immersion and involvement in the virtual environment.

These findings have implications for content creators and developers of immersive media. They highlight the importance of carefully considering sound design and its spatial characteristics when aiming to create captivating and engaging experiences for users. The use of HOA sound, in particular, can offer a more realistic and immersive auditory environment, providing users with a heightened sense of presence and a natural interaction with the content. Overall, the statistical analysis of the subjective questionnaire underscores the significance of sound conditions in shaping users' perception, immersion, and quality of experience in virtual environments.

## 7 Conclusions and Future Work

In conclusion, the research conducted thus far has made significant contributions to understanding visual attention in the context of 360° videos. By analysing a comprehensive dataset that captures viewers' responses to different sound conditions, valuable insights have been gained regarding their visual attention, physiological reactions, and subjective perceptions. The findings shed light on the impact of sound on immersive experiences and offer important considerations for optimizing content delivery in VR and 360° videos.

One of the key findings of this research is the influence of HOA sound on viewers' physiological arousal.

The data revealed that when participants watched videos with HOA sound, their heart rate increased, indicating heightened physiological arousal compared to other sound conditions. This suggests that the complexity and richness of spatial audio can induce a stronger physiological response and potentially enhance the overall engagement and immersion of viewers. Moreover, the analysis of gaze fixations demonstrated that videos with spatial audio resulted in more evenly distributed gaze patterns. Participants exhibited a more comprehensive exploration of the visual content when immersed in environments with HOA sound. This suggests that spatial audio has the potential to guide and shape viewers' visual attention, creating a more captivating and interactive viewing experience.

Subjective ratings provided by the participants further supported the benefits of HOA sound. The HOA sound condition consistently received the highest ratings for realism and clarity, indicating its superiority over stereo and FOA sound. This finding emphasizes the importance of incorporating spatial audio techniques, specifically HOA, to enhance the perceived quality and authenticity of virtual environments.

Looking ahead, future work will focus on conducting joint correlation analysis to delve deeper into the relationships between participants' head pose, pupil diameter, eye gaze, and heart rate data. By examining these variables in conjunction, researchers can uncover intricate patterns and uncover hidden connections that may not be apparent when analysing each factor individually. This integrated approach will provide a more comprehensive understanding of viewers' behaviour and physiological responses during immersive experiences, ultimately contributing to the development of more effective and engaging VR and 360° video content.

Additionally, the dataset collected for this research will continue to serve as a valuable resource for researchers in the field. It can be utilized for further investigations and advancements in VR and 360° video technologies, enabling other scholars to explore different aspects of visual attention and immersive experiences. By sharing this dataset, the research community can collaboratively work towards enhancing the understanding of human perception in virtual environments and contribute to the development of more immersive and captivating virtual experiences.

## Acknowledgements

This research is supported by Science Foundation Ireland and the ADAPT Centre under Grant Number 12/RC/2106.

## References

- [Shojib et al., 2021] Shojib, M. I. M., Kaiser, M. S., Islam, S. M. R., & Al Mahmud, A. (2021). 360-degree video streaming: State-of-the-art and future directions. *IEEE Access*, 9, 30685–30708.
- [Huang and Wang, 2020] Huang, Y., & Wang, R. (2020). Visual attention modelling for virtual reality videos. *IEEE Transactions on Multimedia*, 22(7), 1711–1725.
- [Kim et al., 2020] Kim, J., Han, J., & Lee, S. (2020). Effects of 3D audio on presence, emotional responses, and behavioral intentions in virtual reality. *Multimodal Technologies and Interaction*, 4(4), 80.
- [Xylakis et al., 2020] Xylakis, A., Mellado, S., & Nixon, M. S. (2020). Immersive audio rendering techniques for virtual reality: A survey. *ACM Transactions on Multimedia Computing Communications and Applications*, 16(1s), 1–23.
- [Innes and Shellard, 2005] Innes, G. P., & Shellard, J. H. W. (November 2005). Ambisonic reproduction of 3D soundfields. *Journal of the Audio Engineering Society*, 53(11), 1022–1046.
- [Politis et al., 2018] Politis, A., Siltanen, S., & Pulkki, V. (2018). Higher-order ambisonics. In *Immersive* (pp. 51–102). Springer.
- [Raake, 2014] Raake, A. (2014). Quality of experience: What it is and why it matters. In *Understanding and improving quality of experience in multimedia communications* (pp. 1–15). Springer.
- [Xue et al., 2011] Xue, H. et al. (2011). Subjective quality assessment of video: A tutorial review. *IEEE Signal Processing Magazine*, 28(6), 29–48.
- [Singh and Kapoor, 2018] Singh, K., & Kapoor, K. (2018). A study on quality of experience and quality of service of YouTube videos on mobile devices. *Wireless Personal Communications*, 102(3), 2253–2273.
- [Li et al., 2016] Li, S. Z. et al. (2016). User expectation-aware mobile video streaming. *IEEE Journal on Selected Areas in Communications*, 34(4), 982–992.

- [Wang and Li, 2021] Wang, R., & Li, W. (2021). Objective and subjective quality assessment of 360-degree video streaming: A survey. *IEEE Communications Surveys and Tutorials*, 23(1), 416–445.
- [Shahriar and Shiratori, 2015] Shahriar, A. T. M., & Shiratori, N. (2015). Assessing users' quality of experience with smartphone applications: An empirical study. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK (pp. 52–59).
- [Ahmed et al., 2019] Ahmed, N. et al. (2019). Quality of experience-aware dynamic resource allocation for cloud gaming. *IEEE Transactions on Cloud Computing*, 7(1), 239–252.
- [Wang et al., 2016] Wang, Y. et al. (2016). Quality of experience (QoE) for video streaming: A review. In Proceedings of the 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Chengdu, China (pp. 248–255).
- [Jean et al., 2017] Jean, Huang, Chun-Ying, Chen, Kuan-Ta, Hsu, & Cheng-Hsin. (2017). Lo, Wen-Chih and fan, Ching-Ling and Lee, 360° Video Viewing Dataset in Head-Mounted Virtual Reality, 211–216.
- [David et al., 2018] David, Erwan and Gutiérrez, Jesús and Coutrot, Antoine, Perreira Da Silva, M., & Le Callet, P. (2018). A dataset of head and eye movements for 360° videos. 432–437.
- [Chunjia et al., 2014] Chunjia, Yang, & Xiaokang. (2014). Min, Xiongkuo and Zhai, Guangtao and Gao, Zhongpai and Hu. Sound Influences Visual Attention Discriminately in Videos 6th International Workshop on Quality of Multimedia Experience, QoMEX 2014, 2014 (pp. 153–158).
- [Marighetto et al., 2017] Marighetto, Pierre and Coutrot, Antoine and riche, Nicolas, Guyader, Nathalie, Mancas, Matei, Gosselin, Bernard, Laganiere, & Robert. (2017). Audio-visual attention: Eye-tracking dataset and analysis toolbox.
- [Wu et al., 2017] Wu, Chenglei and tan, Zhihao and Wang, Zhi, Yang, & Shiqiang. (2017). A dataset for exploring user behaviors in VR spherical video streaming. 193–198.
- [Almquist and Almquist, 2018] Almquist, M., & Almquist, V. (2018). Analysis of 360° video viewing behaviours.
- [ISO 8589, 2007] ISO 8589:2007. Sensory analysis—General guidance for the design of test rooms International Standards Organization [Online]. <https://www.iso.org/obp/ui/#iso:std:iso:8589:ed-2:v1:en>
- [Pro, 2018] Pro, T. (2018). Tobii Pro VR Integration – Based on HTC Vive Development Kit Description [Online]. <https://www.tobiipro.com/siteassets/tobiipro/product-descriptions/tobiipro-vr-integration-product-description.pdf?v=1.7>. Retrieved March 27, 2023
- [Beyerdynamic, 2020] Beyerdynamic. (2020). Beyerdynamic DT990 Pro [Online]. <https://europe.beyerdynamic.com/dt-990-pro.html>. Retrieved March 27, 2023
- [GoPro, 2020] GoPro. (2020). Gopro VR Player for Desktop FAQ. [ONLINE]. <https://gopro.com/help/articles/block/gopro-vr-player-for-desktop-faq>. Retrieved March 27, 2023
- [Farina, 2020] Farina, A. (2020). Index of/public [Online]. <http://www.angelofarina.it/Public/>. Retrieved March 7, 2021
- [Empatica, 2023] Empatica. E4 wristband support page. <https://support.empatica.com/hc/en-us/categories/200023126-E4-wristband>. Retrieved March 27, 2023
- [FFmpeg, 2021] FFmpeg.org. (2021). FFmpeg [Online]. <https://ffmpeg.org/>. Retrieved March 27, 2023
- [Keighrey et al., 2017] Keighrey, C., Flynn, R., Murray, S., & Murray, N. (2017). <https://doi.org/10.1109/QoMEX.2017.7965656>
- [Hynes et al., 2019] Hynes, Eoghan and Flynn, Ronan and Lee. (2019). A Quality of Experience Evaluation Comparing Augmented Reality and Paper Based Instruction for Complex Task Assistance (pp. 1–6). Niall.
- [Egan et al., 2016] Egan, Darragh and Brennan, Sean and Barrett, John, Qiao, Yuansong, Timmerer, Christian, Murray, & Niall. (2016). An evaluation of heart rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments. 1–6.
- [ITU, 2016] International Telecommunications Union. (2016). Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment [Online]. <https://www.itu.int/rec/T-REC-P.913/en>. Retrieved March 27, 2023. p.913.
- [Provisu, 2021] Provisu.ch. (2021) [Online]. [https://www.provisu.ch/images/PDF/Snellenchart\\_en.pdf](https://www.provisu.ch/images/PDF/Snellenchart_en.pdf) Retrieved March 27, 2023
- [Colblindor, 2021] Colblindor. (2021) [Online]. <https://www.color-blindness.com/ishiharas-test-for-colour-efficiency-38-plates-edition/> Retrieved March 27, 2023
- [Pigeon, 2020] Pigeon, D. (2020) [Online]. Online hearing test and audiogram printout. Hearingtest.online. <https://hearingtest.online/>. Retrieved March 27, 2023
- [Rigby et al., 2019] Rigby, J. M., Gould, S. J. J., Brumby, D. P., & Cox, A. L. (2019). Development of a questionnaire to measure immersion in video media: The Film IEQ, TVX. Proceedings of the 2019 ACM International Conference InterAct. Exp. TV Online Video, 2019, 35–46.
- [Hirway, A, 2023]. Hirwaam/qoe\_visual-attention\_spatial-Audio\_360-videos.GitHub.Retrieved July 5,2023, from [https://github.com/hirwaam/QoE\\_Visual-Attention\\_Spatial-Audio\\_360-videos.git](https://github.com/hirwaam/QoE_Visual-Attention_Spatial-Audio_360-videos.git)

# An Expert Evaluation of a VR Intervention for Children with ASD

Yujing Zhang, Conor Keighrey, Niall Murray

*Technological University of the Shannon: Midlands Midwest*

## Abstract

The potential of technology-based interventions for children with autism spectrum disorder (ASD) has received significant attention in recent years. In particular, immersive and interactive technologies such as virtual reality (VR) are being compared with traditional approaches. VR interventions claim to bring a richer experience to children with ASD. However, with it comes the potential risk of more sensory stimulation. The existing evaluations are focused on user evaluation, and then there is a gap in the expert evaluation of the autism intervention system to appropriately inform the design of such interventions. In this paper, the author proposes an expert evaluation-based method to conduct a comprehensive test of autism intervention systems.

**Keywords:** ASD, Children, Intervention, Expert Evaluation, VR.

## 1 Introduction

Autism spectrum disorder (ASD) is a complex developmental disorder. Individuals with ASD may be known to display limited facial expressions, atypical attention patterns, and delayed communication and cognitive abilities [Hyman et al., 2020]. Especially, children with ASD are reported to have challenges in social situations compared to typically developing (TD) peers [Broekhof et al., 2015]. These symptoms in individuals with ASD are associated with impaired theory of mind (ToM) [Astington and Jenkins, 1995], which makes people affected usually experience difficulty in understanding and interpreting the feelings and desires of others. Intervention can help to develop strategies for ToM in children with ASD in the early stage [Tager-Flusberg, 2007] and acquire basic social skills. Therefore, interventions targeting children are important. The combination of virtual reality (VR) technologies with traditional interventions in recent years has given rise to multimedia interventions. Meanwhile, the evaluation of VR intervention requires more attention.

Currently, the autism intervention systems are tested by people with ASD. The research on the evaluation of interventions for adults with ASD includes recording and analysing subjective, objective, and physiological data [Adjorlu et al., 2019, Zhang et al., 2017]. There are fewer studies on interventions for children with ASD, where accurate feedback cannot be provided due to difficulty with language comprehension in children with ASD [Kjellmer et al., 2012]. These assessments are therefore either subjective by parents and therapists [Miller et al., 2020], or objectively through their performance data as required by the intervention [Scattone, 2008, Uzuegbunam et al., 2017].

Children with ASD may be overreactive to sensory input and special environments [Corbett et al., 2009]. So, there are many potential challenges to evaluating multimedia-based interventions with them. Also, due to the fact that the user group is children with ASD, they provide very limited subjective feedback in terms of system measurement. There is a lack of comprehensive and professional evaluation of the intervention for children with

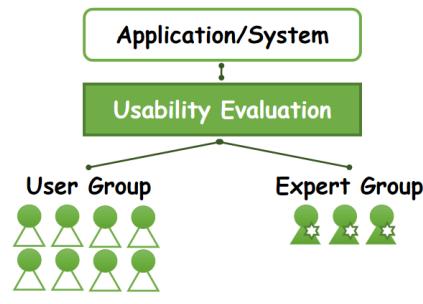


Figure 1: Usability evaluation method.

ASD, even though the systems are designed based on therapist advice [Scattone, 2008, Miller et al., 2020] or heuristic evaluation principles [Moreno et al., 2020]. The usability evaluation can be done in two methods, user evaluation and expert evaluation [Nielsen, 1994, Molich and Dumas, 2008], as shown in Figure 1. The former is primarily a test for target users, while the latter is a test of the system by a small number of experts in a particular field.

The expert evaluation method has been used to measure VR applications [Kloster and Babic, 2019, Kangas et al., 2022, Sutcliffe and Gault, 2004]. Experts can imagine the problems that users may encounter, also, they can identify problems in the system that users may not notice [Molich and Dumas, 2008]. Compared with evaluating the intervention system by testing children with ASD, the expert evaluation will be more detailed and comprehensive to help to design the system more accurately. Moreover, they can identify potential threats in the system and keep the children out of harm's way.

Hence, this paper presents an expert evaluation method to measure the VR intervention for children with ASD. Different from previous studies, it brings a new perspective to testing and evaluating the system. The remainder of this paper is structured as follows, Section 2 introduces existing VR evaluation, and Section 3 presents the design of the expert evaluation.

## 2 Related Work

This section introduces related works to this research.

### 2.1 The Usability of User Evaluation for VR Interventions

Current VR intervention studies focus on testing with targeting users. Interventions for adults and adolescents can be tested and evaluated by users, but the interventions for children are tested on children and evaluated by non-users (their parents/therapists).

#### 2.1.1 User Testing and Non-Users Evaluation

A VR training evaluation was presented by [Miller et al., 2020]. It focused on teaching children with ASD to gain air travel skills and measure their experience through parents' feedback and clinic observations. Five children (aged 4-8 years old) attended and used iPhone X with a Google Cardboard device to experience the VR intervention. Their parents quantified their travel skills on a 5-Likert Scale air travel questionnaire by pre-and post-intervention. A 4-Point clinical activity checkpoint (check-in, security, waiting at the gate, and boarding) was used during the intervention. A speech-language pathologist captured the clinical observations after each session and completed the responses to clinical communication observations. The results of the parents' questionnaires and clinical observations showed that the children's air travel skills improved after the intervention. The advantage of parental feedback is that they know their children very well, but their feedback is more subjective and focuses on their children's performance and lacks professional evaluation of the system.

A communication screen-based intervention [Scattone, 2008] collected the objective data of a nine-year-old boy through audio and video recorders, including eye contact, smile, and initiation as they occurred during the experiment. The experiment consisted of a baseline recording, a video story intervention, and 5 minutes of social interactions. The boy repeated the intervention 1-2 times per week. After three months, the boy's eye contact improved, but the other two targeted skills did not significantly increase. The study showed the video story has the potential possible to help children with ASD to improve their conversation skills. Another study [Uzuegbunam et al., 2017] invited three children (aged 7-11 years old) to practice social etiquette skills based on a computer screen. It recorded the users' body movements data, greeting responses time, frequency of greetings, and eye gaze data. Also, this experiment contained three phases, baseline, intervention, and maintenance. The children had the intervention for 6-14 days and then their maintenance results showed that only one child improved in all targeted skills, and the other two only improved in eye contact. The above studies can be summarised that the evaluation of children's performance takes longer because it requires documentation of baseline, intervention, and maintenance

for data comparison. At the same time, it increases the risk of children being exposed to experiments.

### 2.1.2 User Testing and Evaluation

A head-mounted display (HMD)-based VR music intervention [Adjorlu et al., 2019] to reduce social anxiety in adolescents with ASD, collected their subjective feedback through pre-and post-questionnaires. Three male participants (aged 18-20 years old) completed the experiment. They were asked to finish the Liebowitz Social Anxiety Scale (LSA) questionnaire with 5 Smiley-Likert Scale for measuring their anxiety level and frequency before the experiment. After the VR intervention, to explore the participants' immersive experience, they were asked to answer four questions from the Witmer Singer Presence Questionnaire (PQ) and four questions from the Immersive Tendency Questionnaire (ITQ), all using 7 Smiley-Likert Scale. One participant failed because he could not understand the questions, and the rest of them reported their feedback through the questionnaires. From this it can be seen that the data of the questionnaire (user subjective data) may be unreliable, the questionnaire requires users to have language comprehension skills, but users with ASD may not understand the questions fully.

Another HMD-based VR driving intervention [Zhang et al., 2017] was designed to improve the driving skills of adolescents with ASD and measure cognitive load through driving performance, questionnaires, and physiological signals. During the experiment, twenty participants (aged 13-18 years old) were asked to complete the questionnaires to rate their affective and cognitive states. Meanwhile, various sensors collected the data, including electrocardiogram (ECG), electromyography (EMG), respiration (RSP), skin temperature (SKT), photoplethysmography (PPG), and galvanic skin response (GSR). The performance data included the brake accelerator, failure times, and driving score. Comparing the different types of data, the results showed that multimodal fusion schemes could provide a more accurate measure of users' cognitive load. The design of the questionnaire plays a vital role, which determines the quality of the evaluation, but using subjective data and physiological signals to measure users' experiences can enhance evaluation accuracy.

## 2.2 The Usability of Expert Evaluation for VR Systems

A Mobile VR application [Kloster and Babic, 2019] for neck exercise was evaluated by one physiotherapist through an open-ended interview. He positively evaluated the prototype and gave advice that could make the neck exercises more entertaining. Because according to his work experience, the patients are sometimes not self-motivated enough to maintain the efficiency of their treatment. Also, six users participated in the test, but they only reported feedback on the 5-Likert Scale, with no more subjective feedback compared to expert feedback. The advantage of expert evaluation is experts can provide clear professional advice for the system based on their experience.

An HMD-based VR medical landmarking system [Kangas et al., 2022] for identifying and marking cephalometric landmarks was evaluated by a group of medical experts. The system can be used in three different conditions: "VR+Controller", "VR+Pen", and "Mouse+Display". Four medical experts participated in the experiment, and each completed the three condition tests. The evaluation method included objective measures (task completion times and marking accuracy) and subjective questions (evaluation of each interaction method and ranking of all interaction methods). The objective results showed the "VR+Pen" cost more time but had better marking accuracy than other conditions. In the subjective feedback, "VR+Controller" had the highest praise, and "VR+Pen" was the last, but it was rated as having the most potential for development by experts. The other advantage of the expert evaluation is that the experts with a small number can finish the comparison study.

Therefore, considering the advantages of expert evaluation (discussed above), the VR intervention for children with ASD can be tested and evaluated by experts, especially in pilot tests. It not only solves the problem of limited feedback for children with ASD but eliminates the risks of experimentation on them. Based on the related works, the evaluation can be included implicit and explicit measurements to improve the evaluation accuracy. The author proposes an expert evaluation design, including subjective questionnaires, objective measures, and physiological monitoring in Section 3.

### 3 Proposed Expert Evaluation Design

The author proposed the expert evaluation design, this section consists of a system introduction, participant screening, and various measurements.

#### 3.1 VR Intervention System Set Up

The VR intervention application in this research is based on the Social Story™ book designed to improve social skills in children with ASD through experiencing immersive video stories [Zhang et al., 2023]. The application was developed in Unity Version 2021.3.6f1 and can be used in two conditions, screen-based and HMD-based, as shown in Table 1. Participants need to wear sensors to experience the intervention in two conditions and finish the post-experience questionnaires. The whole process is around 10 minutes, including 5 minutes of baseline recording, 2 minutes of intervention, and 3 minutes of completing questionnaires.

Condition	Experience	Device
Screen-based	2D Story	24-inch computer screen
VR-based	Immersive 3D Story	HTC-Vive Pro Eye

**Table 1:** The VR intervention application in two conditions.

#### 3.2 Participants

According to ITU-T P.910's instructions on the number of experts required [P.910, 2022], the author plans to invite four experts working in the field of ASD. At the same time, it is necessary to avoid the homogeneity of the usability evaluator structure, resulting in insufficient evaluation objectively [Rosenbaum, 1989]. Considered bringing together experts in different areas of ASD in the proposed plan, including a clinical psychologist, a behavioural psychologist, a behaviour analyst, and a speech and language pathologist. The experts will complete the two conditions experiments. The process aims to gain feedback from them on the application design and evaluate the 2D and immersive 3D experience.

#### 3.3 Subjective Questionnaires

The subjective questionnaires have consisted of three different types of questionnaires.

##### 3.3.1 7-Likert Scale Questions

The 8 questions are selected from the 24-question Film Immersive Experience Questionnaire [Rigby et al., 2019] based on the VR story intervention content, as shown in Table 2. The questionnaire is consisted of three factors: immersion (Q1, Q3, Q6), enjoyment (Q2, Q5, Q8), and comprehension (Q4, Q7). The 8 questions need to be filled after each experiment by experts, and the 7-Likert Scales aim to get the quantitative data from them.

Q1	To what extent did the video hold your attention?
Q2	To what extent did you enjoy the imagery?
Q3	To what extent did you notice events taking place around you?
Q4	To what extent did you find the concepts and themes easy to understand?
Q5	To what extent did you feel motivated to keep on watching?
Q6	To what extent did you feel you were focused on the video?
Q7	How well do you think you understood what happened?
Q8	Would you like to watch more of this, or similar content, in the future?

**Table 2:** The questions with 7-Likert Scales to evaluate the experience.

### 3.3.2 Comparison Questions

The 3 questions shown in Table 3 are modified based on Kangas's questionnaire [Kangas et al., 2022], which aims to evaluate the screen-based and HMD-based and explore which experience is more acceptable for children with ASD. The experts need to complete this comparison questionnaire by combining the VR story intervention experience and imagining the needs of children with ASD to provide the indicative data.

Q1	Which experience do you think the children with ASD would enjoy most?
Q2	Which experience would be more appropriate for children with ASD?
Q3	Which experience has more potential for development?

Table 3: The questions to compare the two different experiences.

### 3.3.3 Open-ended Questions

The open-ended questions as shown in Table 4 are served for upgrading the application through experts' advice. Also, experts need to explore the potential risk of the application and provide solutions.

Q1	Where does the story content need to be improved?
Q2	How about this scene (background music, light, colour, voice, time, character)?
Q3	What are the potential risks for children with ASD in this application and how to avoid them?
Q4	Which age group of children with ASD is this application suitable for?

Table 4: The Open-ended questions based on the application.

### 3.4 Objective Measures

In objective measures, the screen-based application and HMD-based application are different, as shown in Figure 2. The user interface of screen-based intervention is 2D and fixed by moving four virtual cameras. The HMD-based intervention depends on the user's head movement, users will freely look around in the immersive environment. Unity Recorder is designed to record the user's entire usage process as mp4 files. Microsoft Azure Kinect DK records the body movements and gestures, which captures depth and colour images at a frame rate of 30 fps.

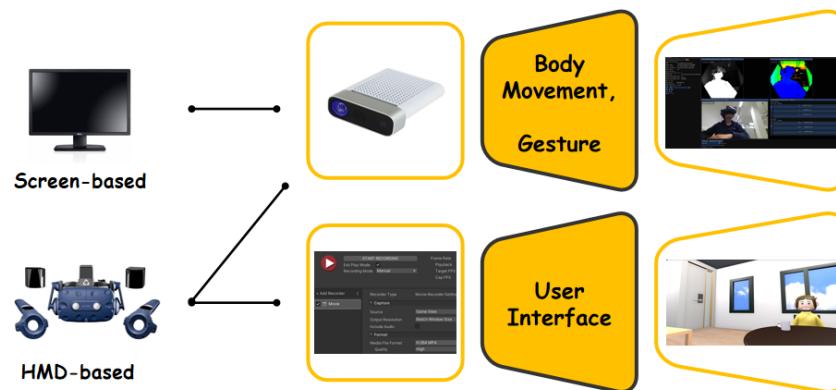
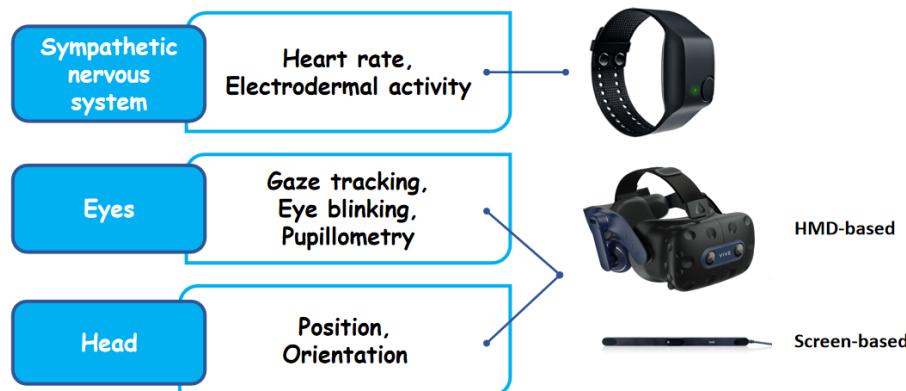


Figure 2: The objective measures recording of screen-based and HMD-based.

### 3.5 Physiological Monitoring

According to the ITU-T P.1320 standard, in addition to subjective questionnaires and objective measures, physiological signals are also important in the system evaluation [P.1320, 2022]. It is associated with covert processes that may cause emotional arousal. The heart rate (HR) and electrodermal activity (EDA) can reflect people's real inner activities, including stress levels and emotional reactions [Trotman et al., 2019, Prokasy, 2012,

Bradley et al., 2001]. Eye data measurement consists of gaze tracking, eye blinking, and pupillometry for measuring fatigue, cybersickness, and attention [Stern et al., 1994, Lopes et al., 2020]. Moreover, the head movement data should be recorded due to the head movement triggers the eye muscles [Somisetty et al., 2019]. The HR and EDA data will be monitored by the E4 wristband, and the eyes and head data will be recorded by HTC Vive Pro Eye and Tobii Pro Spark under two conditions separately, as shown in Figure 3.



**Figure 3: The various sensors in physiological monitoring.**

## 4 Conclusions and Future Work

This paper introduces the design of an expert evaluation method to measure VR intervention for children with ASD. The aim is to obtain more professional and comprehensive feedback. The author summarised the existing intervention evaluations and find out a gap in the evaluation of VR intervention for children with ASD, and then proposed that the expert evaluation method could be used to fill this gap. Based on the content of the VR intervention system, the author designed the evaluation includes post-experience questionnaires, objective measures, and physiological monitoring. The author plans to test the expert evaluation method on the VR intervention in future work.

## Acknowledgements

This research is funded by the President's Doctoral Scholarship of the Technological University of the Shannon.

## References

- [Adjorlu et al., 2019] Adjorlu, A., Barriga, N. B. B., and Serafin, S. (2019). Virtual reality music intervention to reduce social anxiety in adolescents diagnosed with autism spectrum disorder. In *16th Sound and music computing conference*, pages 261–268. Sound and Music Computing Network.
- [Astington and Jenkins, 1995] Astington, J. W. and Jenkins, J. M. (1995). Theory of mind development and social understanding. *Cognition & Emotion*, 9(2-3):151–165.
- [Bradley et al., 2001] Bradley, M. M., Codispoti, M., Cuthbert, B. N., and Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion*, 1(3):276.
- [Broekhof et al., 2015] Broekhof, E., Ketelaar, L., Stockmann, L., van Zijp, A., Bos, M. G., and Rieffe, C. (2015). The understanding of intentions, desires and beliefs in young children with autism spectrum disorder. *Journal*

- of autism and developmental disorders*, 45:2035–2045.
- [Corbett et al., 2009] Corbett, B. A., Schupp, C. W., Levine, S., and Mendoza, S. (2009). Comparing cortisol, stress, and sensory sensitivity in children with autism. *Autism Research*, 2(1):39–49.
- [Hyman et al., 2020] Hyman, S. L., Levy, S. E., Myers, S. M., Kuo, D. Z., Apkon, S., Davidson, L. F., Eller-Beck, K. A., Foster, J. E., Noritz, G. H., Leppert, M. O., et al. (2020). Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics*, 145(1).
- [Kangas et al., 2022] Kangas, J., Li, Z., and Raisamo, R. (2022). Expert evaluation of haptic virtual reality user interfaces for medical landmarking. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7.
- [Kjellmer et al., 2012] Kjellmer, L., Hedvall, Å., Holm, A., Fernell, E., Gillberg, C., and Norrelgen, F. (2012). Language comprehension in preschoolers with autism spectrum disorders without intellectual disability: Use of the Reynell developmental language scales. *Research in Autism Spectrum Disorders*, 6(3):1119–1125.
- [Kloster and Babic, 2019] Kloster, M. and Babic, A. (2019). Mobile VR application for neck exercises. In *ICIMTH*, pages 206–209.
- [Lopes et al., 2020] Lopes, P., Tian, N., and Boulic, R. (2020). Exploring blink-rate behaviors for cybersickness detection in VR. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pages 794–795. IEEE.
- [Miller et al., 2020] Miller, I. T., Wiederhold, B. K., Miller, C. S., and Wiederhold, M. D. (2020). Virtual reality air travel training with children on the autism spectrum: A preliminary report. *Cyberpsychology, Behavior, and Social Networking*, 23(1):10–15.
- [Molich and Dumas, 2008] Molich, R. and Dumas, J. S. (2008). Comparative usability evaluation (cue-4). *Behaviour & Information Technology*, 27(3):263–281.
- [Moreno et al., 2020] Moreno, G., Moreira, F., Collazos, C. A., and Fardoun, H. M. (2020). Recommendations for the design of inclusive apps for the treatment of autism: An approach to design focused on inclusive users. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.
- [Nielsen, 1994] Nielsen, J. (1994). Heuristic evaluation. *Usability Inspection Methods*.
- [P.1320, 2022] P.1320, I.-T. (2022). P.1320: Quality of experience assessment of extended reality meetings. The last version was published in 07/22.
- [P.910, 2022] P.910, I.-T. (2022). P.910: Subjective video quality assessment methods for multimedia applications. The last version was published in 07/22.
- [Prokasy, 2012] Prokasy, W. (2012). Electrodermal activity in psychological research. *Elsevier*.
- [Rigby et al., 2019] Rigby, J. M., Brumby, D. P., Gould, S. J., and Cox, A. L. (2019). Development of a questionnaire to measure immersion in video media: The film IEQ. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, pages 35–46.
- [Rosenbaum, 1989] Rosenbaum, S. (1989). Usability evaluations versus usability testing: When and why? *IEEE transactions on professional communication*, 32(4):210–216
- [Scattone, 2008] Scattone, D. (2008). Enhancing the conversation skills of a boy with Asperger's disorder through social stories™ and video modeling. *Journal of Autism and Developmental Disorders*, 38:395–400.
- [Somisetty et al., 2019] Somisetty, S. et al. (2019). Neuroanatomy, vestibuloocular reflex.
- [Stern et al., 1994] Stern, J. A., Boyer, D., and Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Human factors*, 36(2):285–297.

- [Sutcliffe and Gault, 2004] Sutcliffe, A. and Gault, B. (2004). Heuristic evaluation of virtual reality applications. *Interacting with computers*, 16(4):831–849.
- [Tager-Flusberg, 2007] Tager-Flusberg, H. (2007). Evaluating the theory-of-mind hypothesis of autism. *Current Directions in psychological science*, 16(6):311–315.
- [Trotman et al., 2019] Trotman, G. P., Veldhuijzen van Zanten, J. J., Davies, J., Möller, C., Ginty, A. T., and Williams, S. E. (2019). Associations between heart rate, perceived heart rate, and anxiety during acute psychological stress. *Anxiety, Stress, & Coping*, 32(6):711–727.
- [Uzuegbunam et al., 2017] Uzuegbunam, N., Wong, W. H., Cheung, S.-C. S., and Ruble, L. (2017). Mebook: multimedia social greetings intervention for children with autism spectrum disorders. *IEEE Transactions on Learning Technologies*, 11(4):520–535.
- [Zhang et al., 2017] Zhang, L., Wade, J., Bian, D., Fan, J., Swanson, A., Weitlauf, A., Warren, Z., and Sarkar, N. (2017). Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE Transactions on affective computing*, 8(2):176–189.
- [Zhang et al., 2023] Zhang, Y., Keighrey, C., and Murray, N. (2023). A VR intervention based on social story<sup>TM</sup> to develop social skills in children with ASD. In *ACM International Conference on Interactive Media Experiences*.

# A Quality of Experience Evaluation of an Interactive Multisensory 2.5D Virtual Reality Art Exhibit

Chen Chen, Niall Murray, Conor Keighrey

*Technological University of the Shannon: Midlands Midwest, Athlone, Co. Westmeath, Ireland*

## Abstract

In recent years, museums have become more interactive and immersive through the adaptation of technology within large scale art exhibitions. Due to these changes, these new types of cultural experiences are more appealing to younger audiences. Despite these positive changes, some museum experiences are still primarily focused on visual art experiences. These remain out of reach to those with visual impairments. Such unimodal and visual dominated experiences restrict these users who depend on sensory feedback to experience the world around them. In this paper, the authors propose a novel VR experience which incorporates multisensory technologies. It allows individuals to engage and interact with a visual artwork museum experience presented as a fully immersive VR environment. Users can interact with virtual paintings and trigger sensory zones which deliver multisensory feedback to the user. These sensory zones are unique to each painting, presenting thematic audio and smells, custom haptic feedback to feel the artwork, and lastly air, light and thermal changes in an effort to engage those with visual impairments.

**Keywords:** Virtual Reality, Visual Impairment, Multisensory, Accessibility, Immersive Experience

## 1 Introduction

There are many etiquettes which must be considered when attending a modern-day museum. Some examples include minimizing noise and avoiding flash photography, but most important of all to not touch the items which are on display. This etiquette is essential to the preservation of paintings and artifacts. However, this also limits the ability of those with visual impairments to experience visual art. In recent years, museums have begun to explore new and emerging technologies to present visual art which can be touched [1] [2] [3], with a prime objective of making the experiences more accessible to those with visual impairments.

Visual impairments are disorders which can affect an individual's ability to see or interpret visual information. There are numerous forms of visual impairments which can affect how people perceive the world. Some common problems range from challenges with blurred vision, sensitivity to light, and difficulties seeing at night. According to [4], cataracts is the leading cause of blindness and the second leading cause of moderate and severe visual impairment. Cataracts are primarily age related; however, they can occur as a result of injury to the tissue which makes up the eyes lens. Individuals who experience symptoms of cataracts are limited in how they engage with leisure activities such as visiting cinemas, theatres, and museums. Cataracts and other visual impairments affect not only how people see but can also play into other aspects of general life.

Throughout history, there have been many artists who have thrived whilst experiencing a visual impairment, examples of such include Claude Monet (cataracts) [5], Leonardo da Vinci (intermittent exotropia) [6], and Francis Bacon (dysmorphopsia) [7]. Despite these limitations, these influential artists continued their journey within the creative community and inspired generations. Although artists can develop methods to assist with the creative aspect of art, the consumption of art is still somewhat effected. As a visual impairment evolves, individuals often become more dependent on their sensory system to engage with the world around them [8]. The additional senses become more important as people depend more on touch, smell, and auditory experiences as a primary method of

understanding the world around them.

As technology has evolved, collections of art have transformed from the presentation of 2D canvas within museums to fully interactive media experiences utilizing projection-based media [9], or immersive technologies. Newer technologies such as touch-screen tablets and mobile phones include accessibility features to assist those with visual impairments. This includes a level of customization to increase font size alter the appearance of colours, provide access to magnification tools, and in some cases provide text-to-speech systems which guide users around an interface. Although this form of accessibility is sufficient for 2D media experiences, next generation and fully immersive technologies (Virtual, Mixed, and Augmented Reality) provide new challenges for inclusivity.

Traditionally, game development has focused on the creation of two dimensional (2D) or three dimensional (3D) graphical worlds to immersive users. 2D games are generated using collections of image planes which have been layered in the X and Y axis, the players view perspective is then limited to one axis. Meanwhile 3D games utilize advanced graphical processing and allow the player to perceive a world in three dimensions (X, Y, and Z), players are free to roam the world without view restrictions. Although traditionally consumed using monitors, advances in technology now allow 3D games to be presented in virtual reality (VR) further immersing the user into a fully simulated virtual world. Despite these advances in recent years, developers have begun to experiment with a new style of graphical rendering known as 2.5D (2.5 dimensional).

In 2.5D, flat 2D image planes and 3D elements are mixed and layered to allow players to gain more depth and immersion throughout gameplay. In this context, this paper presents a work in progress design and development of an accessible multimodal VR museum exhibit which aims to present 2.5D virtual artworks. Each layer within the 2.5D composition will coincide with sensory zones to allow those with a visual impairment to experience multimodal feedback from the works of art.

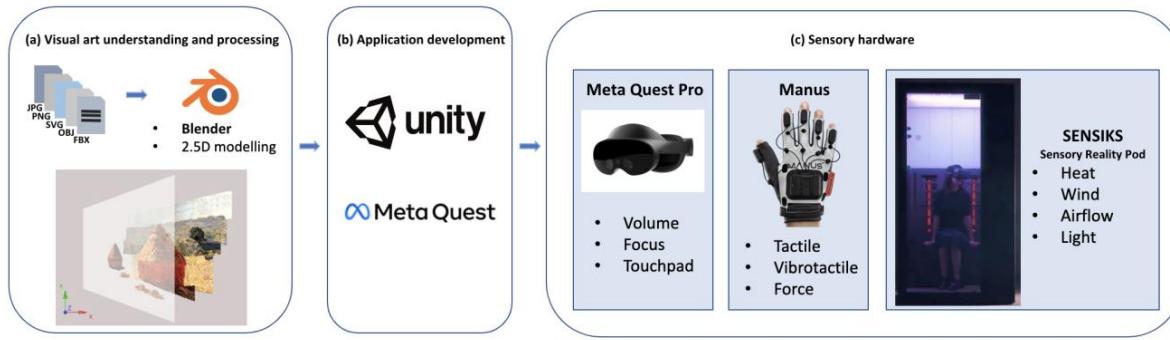
## 2 Related Work

Virtual reality has been explored as a tool for entertainment, education, retail, architecture, and virtual tourism. In recent years, virtual reality has also been used as a tool to assist those with visual impairments to experience the world around them. In [10], how virtual reality fostered embodied learning benefiting those with visual impairments in teaching scenarios was investigated. This is because it engages multiple senses, including proprioception, body movement, touch, and hearing, rather than focusing primarily on visual resources.

Exhibits which focus on accessibility are core to the development of new and inclusive experiences which allow those with a disability or impairment to engage with cultural galleries. In 2021, the “Blind Spot” exhibit was presented at the Central Museum in Utrecht [2]. It focused on stimulating all the senses. The exhibit presented a 4D art experience using real world objects which allowed visitors to touch, smell, and hear audio all of which were electively themed to the visual works. The primary objective of the exhibit was to make visual art more accessible for those with visual impairments. Similarly, Spain’s Prado Museum which has created 3D copies of some of the world’s most well-known visual art and made it part of the exhibit [1]. Utilizing a novel painting mixture, chemicals react to create a 3D depth to the paintings surface allowing those with a visual impairment to feel the colour and texture of the works.

Other examples of this can be seen in the work of tech startup NeuroDigital who are exploring the use of VR and haptic feedback as a mechanism to interact with sculptures on display in museums. Their exhibit titled “Touching Masterpieces”, appeared in the National Gallery of Prague. It aimed to transform physical objects into virtual objects which people could “see” using a haptic glove technology [3].

The work of [2] [1] [10] and [3] highlights the importance of stimulating the senses for individuals with visual impairments. As the ocular system degrades, a greater dependency is placed on the other 4 senses. Existing work has emphasized the importance of using audio to assist with localization [11] and present on-demand audio descriptions to improve independent access to and understanding of visual artworks [12]. Other sensory works have explored the utility of using thermal changes to convey depth and colour [13] [14]. In [15], authors presented a



**Figure 1: Proposed Multisensory VR System Design**

multimodal system to assist those with visual impairments to experience visual art. The process required the high-resolution scanning of images, the images were then processed to extract a heightmap which accentuated key details within the visual art. The height maps were then 3D printed utilizing special means to create tactile graphics which users could feel, this was combined with localised touch sensitive audio to trigger descriptive overviews.

### 3 Proposed System

The following section presents an overview of the proposed system, first the system design is presented outlining the hardware setup. Following this, the user experience section provides an overview of how the user will interact with the virtual world. Lastly, the proposed methodology for the system evaluation is presented.

#### 3.1 System Design

The system design (**Error! Reference source not found.**) consists of three core blocks: (a) Visual art understanding and processing, (b) application development, and (c) sensory hardware.

##### 3.1.1 Visual Art Understanding and Processing

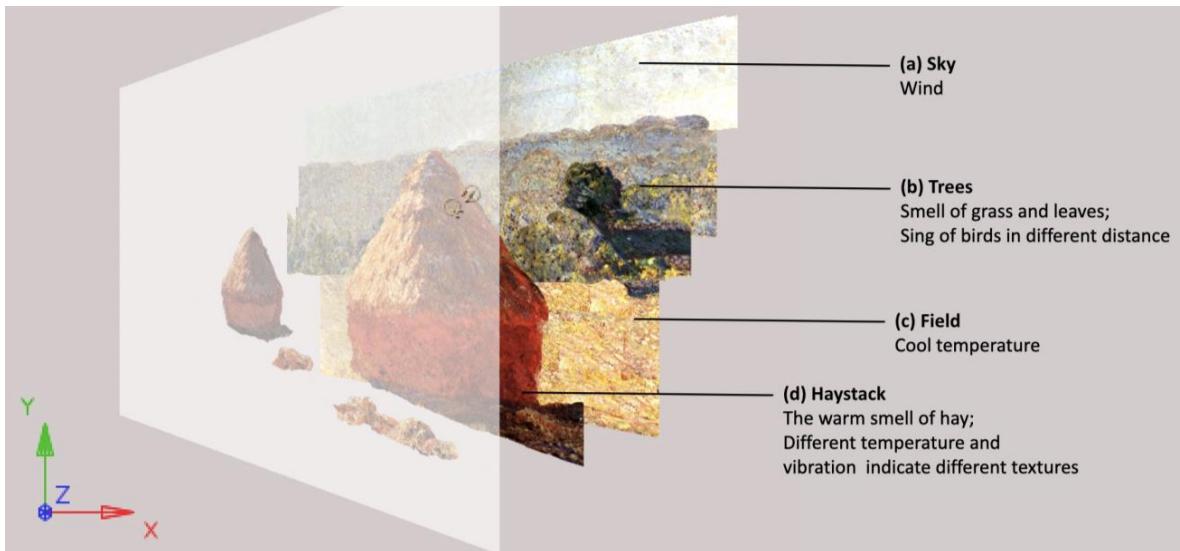
The initial development phase will focus on the careful selection and aggregation of visual works from the famous artist Claude Monet. Paintings will be processed using image editing tools such as Adobe Photoshop to separate layers of visual information.

These layers will be processed within a 3D modelling tool such as Blender to generate the 2.5D representations. An example of this can be seen in **Error! Reference source not found.** [16] which portrays Claude Monet's - Haystacks, End of Summer. In an effort to gain insight into the objective of each of the paintings, research will be carried out on a per painting basis in an effort to identify what the artist had visualized as part of the work. This process will be used to identify thematic audio for each of the paintings.

##### 3.1.2 Application Development and Sensory Hardware

The Unity3D game engine will be utilized to develop the experience. The core of VR interaction will be implemented using the Unity XR Interaction Toolkit [17]. However, additional integration will be required on a per device basis (as outlined later in this section). The proposed application will stimulate the audio and visual senses using the Meta Quest Pro. Although this system is capable of a wireless experience, the headset will be connected to a PC in a tethered setup. This approach is required due to the number of VR accessories which are needed for the proposed system. Due to this, a PC with 32GB of RAM, an Intel i9 (K-Series), and an Nvidia RTX 3080 will be used to run the application.

A haptic glove system from MANUS has been proposed to allow users to gain a tactile sensation whilst interacting with the immersive visual art [18]. Lastly, the SENSIKS Sensory Reality Pod [19] will be utilized to



**Figure 2: Claude Monet's- "Haystacks, End of Summer" visualized in 2.5D.  
Sensory zones are highlighted (a)- (d)**

generate thermal, light, and wind changes for the end user. The sensory reality pod has been selected due to the enclosed nature of the system and seated position which will serve as a measure of safety for users of the experience.

### 3.2 User Experience

The VR application aims to present users with a small collection of paintings from the world renown artist Claude Monet. Each painting will present a unique sensory response from the system. The primary objective of this is to convey the feeling and emotion inspired from the artist through an associated multisensory experience. Due to the variability of each work, the following provides an overview of the systems potential sensory feedback.

As users move their hands through the visual works, they will engage with sensory zones. An example of these sensory zones has been visualised in Figure 2 (a) – (d). These sensory zones will trigger multimodal feedback to the end user in the form of tactile, audio, olfactory, light, and airflow changes. Thematic audio will be presented, as users engage with different layers. Volume will be altered between louder and softer tones to convey depth and distance to the end user. Users will feel the depth of different artworks with haptic gloves as each layer conveys feedback alterations. Thermal changes will occur within the SENSIKS Sensory Reality Pod as users engage with different layers within the virtual painting. The correlations that exist between warm and near, cold and far will be used to convey depth and distance through temperature range. Considering already established approaches to multimedia accessibility [20], the VR experience will place an emphasis on the integration of user navigation and interaction paradigms which are cognisant of those with visual impairments. Examples of this include the use of audible cue's presentation of haptic feedback to the user.

### 3.3 Assessment Methodology

As part of the evaluation process, the system will be evaluated utilizing two independent groups. The initial group will focus on those without any visual impairments in an effort to gain insight into the usability of the system. Once this stage is complete, a smaller group of participants who experience a visual impairment will be invited to evaluate the system. On a larger scale, the objective is to have the system presented within an exhibit space, welcoming members of the public to the experience in an effort to raise awareness on accessibility.

In an effort to quantify and understand user perceived quality of experience (QoE), explicit, implicit, and objective measures will be captured throughout the experience. With an emphasis on accessibility, a post experience

questionnaire will be presented to capture an explicit measure of user perceived QoE. Although explicit measures of QoE have remained a cornerstone within experimental evaluations, they can suffer from user bias and memory effects. In this proposal, physiological measures [21] in the form of heart rate, electrodermal activity, blood and volume pulse will be captured using the Empatica E4 [22]. These implicit measures provide opportunity to gain insight into the experience, alleviating challenges such as difficulties with participants articulating the subjective experience accurately. Lastly, as a supportive measure, user interaction within the immersive experience will be captured, these measures aim to support the understanding of physiological measures of QoE [23].

## 4 Conclusion and Future Work

This paper proposes a multisensory 2.5D VR art exhibition which aims to enhance museum experiences for individuals who are living with a visual impairment. The VR environment allows users to experience visual art by engaging the users sense of touch, smell, and hearing. Users will experience the work of world renown painter Claude Monet within a virtual context. Each virtual painting will present unique sensory zones which will trigger thematic audio and smells, custom haptic feedback, thermal changes, and light and air changes which are unique to the artwork.

As part of future work, a study will be carried out which will capture and report measures of user perceived QoE. The experiment will evaluate two test groups, one with visual impairment and another test group which do not experience any visual impairment. Lastly, in an effort to raise awareness on the topic of accessibility for individuals with visual impairments, it's proposed that the system will be showcased to the public as part of an interactive and immersive arts exhibition.

## References

- [1] Museo Nacional del Prado, “Touching the Prado - Exhibition - Museo Nacional del Prado,” Museo Nacional del Prado, [Online]. Available: <https://www.museodelprado.es/en/whats-on/exhibition/touching-the-prado/0d94a9bf-07d7-491a-866a-37976169f929>. [Accessed 01 04 2023].
- [2] A. Murphy, “4D exhibition in Netherlands creates sensory experience for sight impaired visitors - MuseumNext,” MuseumNext, 20 August 2021. [Online]. Available: <https://www.museumnext.com/article/4d-exhibition-in-netherlands-creates-sensory-experience-for-sight-impaired-visitors/>. [Accessed 04 April 2023].
- [3] WPP plc, “Geometry Prague: NeuroDigital Touching Masterpieces | WPP,” WPP plc, 01 05 2019. [Online]. Available: <https://www.wpp.com/en/featured/work/2019/05/geometry---touching-masterpieces>. [Accessed 25 03 2023].
- [4] Wei Wang; William Yan; Kathy Fotis; Noela M. Prasad; Van Charles Lansingh; Hugh R. Taylor; Robert P. Finger; Damian Facciolo; Mingguang He, “Cataract surgical rate and socioeconomic: A global study,” *Clinical and epidemiologic research*, vol. 57, no. 14, pp. 5872- 5881, 2017.
- [5] R. Hajar, “Monet and Cataracts,” *Heart Views*, vol. 17, no. 1, pp. 40-41, 2016.
- [6] C. W. Tyler, “Evidence That Leonardo da Vinci Had Strabismus,” *JAMA Ophthalmology*, vol. 137, no. 1, pp. 82-86, 2019.
- [7] A. . B. Safran, N. Sanda and J.-A. Sahel, “A neurological disorder presumably underlies painter Francis Bacon distorted world depiction,” *Frontiers in Human Neuroscience*, vol. 8, no. 581, pp. 1-3, 2014.
- [8] A. Killen, M. J. Firbank, D. Collerton, M. Clarke, J. M. Jefferis, J.-P. Taylor, I. G. McKeith and U. P. Mosimann, “The assessment of cognition in visually impaired older adults,” *Age and Ageing*, vol. 42, no. 1, pp. 98-102, 2013.

- [9] Fever Labs Inc, "Van Gogh Exhibition: The Immersive Experience," Fever Labs Inc, [Online]. Available: <https://vangoghexpo.com/>. [Accessed 15 03 2023].
- [10] Nikoleta Yiannoutsou; Rose Johnson; Sara Price, "Non-visual virtual reality: considerations for the pedagogical design of embodied mathematical experiences for visually impaired children," *Educational Teachnology & Society*, vol. 24, no. 2, pp. 151-163, 2021.
- [11] A. N. Moraes, R. Flynn, A. Hines and N. Murray, "The role of physiological responses in a VR-based sound localization task," *IEEE Access*, vol. 9, pp. 122082-122091, 2021.
- [12] Jun Dong Cho; Jaeho Jeong; Ji Hye Kim; Hoonsuk Lee, "Sound coding colour to improve artwork appreciation by people with visual impairments," *Electronics*, vol. 9, no. 11, 2020.
- [13] Jorge Iranzo Bartolomé; un Dong Cho; Luis Cavazos Quero; Sunggi Jo; Gilsang Cho, "Thermal interactive for improving tactile artwork depth and colour-depth appreciation for visually impaired people," *Electronics*, vol. 9, no. 11, 2020.
- [14] Shaoyu Cai; Pingchuan Ke; Takuji Narumi; Kening Zhu, "ThermAirGlove: A pneumatic glove for thermal perception and material identification in virtual reality," in *Virtual Reality and 3D User Interfaces (VR)*, Atlanta, 2020.
- [15] Luis Cavazos Quero; Jorge Iranzo Bartolomé; Jundong Cho, "Accessible visual artworks for blind and visually impaired people: Cpmparing a multimodal approach with tactile graphics," *Electronics*, vol. 10, no. 3, 2021.
- [16] R. R. Brettell, "Monet's Haystacks Reconsidered," *Art Institute of Chicago Museum Studies*, vol. 11, no. 1, pp. 4-21, 1984.
- [17] E. Provencher, D. Ruddell, "Unity," Unity, 13 March 2023. [Online]. Available: <https://blog.unity.com/engine-platform/whats-new-in-xr-interaction-toolkit-2-3>. [Accessed 22 June 2023].
- [18] M. Meta, "'Prime X Haptic VR," MANUS Meta," [Online]. Available: <https://www.manus-meta.com/products/prime-x-haptic>. [Accessed 20 03 2023].
- [19] SENSIKS, "'Pod & Platform - SENSIKS,'" [Online]. Available: <https://www.sensiiks.com/pods-platform/>. [Accessed 01 04 2023].
- [20] M.T. Paratore, B. Leporini, "Exploiting the haptic and audio channels to improve orientation and mobility apps for the visually impaired," *Universal Access in the Information Society*, pp. 1- 11.
- [21] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chat, N. Ramzan and K. Brunnström, "Psychophysiology-Based QoE Assessment: A Survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6-21, 2017.
- [22] Empatica, "E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors," [Online]. Available: <https://www.empatica.com/research/e4/>. [Accessed 05 02 2023].
- [23] C. Keighrey, R. Flynn, S. Murray, S. Brennan and N. Murray, "Comparing User QoE via Physiological and Interaction Measurements of Immersive AR and VR Speech and Language Therapy Applications," in *in ACM Multimedia, Mountain View, CA, USA*, 2017.

# Virtual Rehabilitation for Patients with Osteoporosis: Translating Physiotherapy Exercises into Exergames

Eléa Thuilier<sup>1</sup>, John Carey<sup>1</sup>, Bryan Whelan<sup>1</sup>, John Dingliana<sup>2</sup>, Mary Dempsey<sup>1</sup>, Shane Biggins<sup>3</sup>, Kenzo Thuilier<sup>4</sup>, Attracta Brennan<sup>1</sup>

<sup>1</sup> University of Galway (Galway, Ireland), <sup>2</sup> Trinity College Dublin (Dublin, Ireland), <sup>3</sup> Hospital of Galway (Galway, Ireland), <sup>4</sup> Université de Rennes (Rennes, France).

## Abstract

Osteoporosis affects over 200 million people worldwide; osteoporotic fractures occur every three seconds. Physical therapy is a non-pharmaceutical approach to improving; muscle strength, balance and flexibility, thereby reducing the risk of falls and fractures. However, engagement in physical therapy can be challenging due to factors including accessibility, cost, fear and/or boredom. Virtual rehabilitation, incorporating extended reality, and exergames, has emerged as a promising solution to enhance and address engagement issues with traditional physical therapy. In this paper, the authors propose a set of five safe and clinically approved physical exercises targeting older adults with osteoporosis. Underpinned by guidelines derived from a systematic review on virtual rehabilitation and compliant with HSE care guidelines and the concept of ‘good’ movement, these exercises were reproduced in AR as exergames using the Microsoft Hololens 2 head-mounted display and the Azure Kinect camera. Through the integration of expert knowledge and technology, this research contributes to the development of adapted virtual rehabilitation interventions for patients with osteoporosis.

**Keywords:** Augmented reality, Rehabilitation, Osteoporosis, Exergames, Physical therapy.

## 1. Introduction

Today, more than 200 million people suffer from osteoporosis [Sözen, 2017]. One in three post-menopausal women and one in five men over the age of 50 years old are at risk of an osteoporotic fracture during their lifetime. These fractures can result in substantial morbidity, mortality and financial costs [McCabe, 2020; Sözen, 2017]. Given the significant societal and personal burden imposed by osteoporosis, novel approaches which improve quality of life and reduce healthcare costs for people with osteoporosis deserve further investigation [Sözen, 2017].

Physical therapy is an important non-pharmaceutical intervention which improves strength, balance and flexibility, helping to manage pain and reduce the risk of falls and fractures [Dionyssiotis, 2014]. However active engagement with physical therapy can be problematic due to environmental (accessibility, cost, time), physical (risk of falling, difficulty in movement, co-morbidities) and mental factors (fear, boredom or motivation) [Obuko, 2018]. Visual assessments of movement quality by physiotherapists can be accurate and reliable for rating various movements, such as gait [Whatman, 2012], but the quality of these assessments depends on the knowledge, skill and experience of the physiotherapist [Whatman, 2012]. This is one reason virtual rehabilitation is gaining popularity as an alternative or complementary solution to physical therapy [Amanda, 2014].

Virtual rehabilitation refers to any rehabilitation or recovery process which includes the use of a virtual technology (e.g., Virtual reality (VR), Augmented reality (AR), or other digital technologies). Virtual rehabilitation is used to enhance traditional rehabilitation and improve cognitive and/or physical skills [Peng, 2011]. Whilst VR immerses the user in a fully virtual environment, AR overlays digital information onto the real world [Bouchard, 2014]. Exergames are a type of video game or multimedia interaction which requires the player to physically move in order to play [Oh, 2010]. Studies show that exergames can improve physical (balance, strength, fall risk) and cognitive (fear, motivation, memory) function in older adults [Amanda, 2014]. Feedback during exergame participation is key to effective virtual training, similar to physical therapy where the therapist guides and supports the patient’s movement, posture and participation [Alnajjar, 2019].

Feedback is typically categorised as visual, auditory, and haptic [Lee, 2017], while provision of timely and useful feedback ensures patients are effectively supported [Doyle, 2011]. We didn't find exergames designed specifically for osteoporosis patients in a systematic literature review. In this paper, we present the integration of an extensive systematic literature review and communication with health and social care experts (i.e., Physiotherapists (PTs), Occupational Therapists (OTs) and Rheumatologists) to design of novel exergames for patients with osteoporosis, with the aim of reducing their risk of fall and fracture.

Note: This paper only presents the design decisions prior to exegame development and testing.

## 2. Method

In the process of translating approved physical therapy exercises into a set of five exergames for patients diagnosed with osteoporosis, we based our technology choices and game design considerations on:

- Guidelines resulting from a prior systematic review on virtual rehabilitation.
- Input from health and social care experts to ensure that the exergames are suitable and do not put patients at risk of fall or injury and are compliant with HSE care guidelines.
- Further collaboration with a chartered PT regarding 'good movement' and feedback.

Following the PRISMA guidelines our systematic literature review included 23 articles for final analysis from 130 at initial selection. The key results and advice of health and social care experts underpin:

- The choice of adapted, safe, and clinically approved training exercises.
- The definition of overall 'good' movement and its implication for people with osteoporosis,
- Instruction and feedback between patients and physiotherapists during training sessions.

In this paper solely we present considerations and justifications in the design process of our exergames.

## 3. Design Process and Decision Justification

### 3.1 Designing an adapted set of exergames for people with osteoporosis.

In our study a final set of five physical training exercises were agreed by a chartered PT, an OT and two Rheumatologists at Galway University Hospital. These exercises target older adults with osteoporosis and comply with HSE care guidelines [Brosna, 2020]. Their aim is to improve physical outcomes such as; muscle strength, flexibility, and balance, for those diagnosed with osteoporosis by Dual-energy X-ray Absorptiometry (DXA) criteria (T-score  $\leq -2.5$ ) per International Society for Clinical Densitometry (ISCD) modifications (postmenopausal women and men aged 50 years and older). These exercises are represented in Table 1, in addition to an associated instructional narrative as provided by the PT to his patients.

Exercise	Instructions from the PT
Sit to stand (with a chair)	Context: A chair is placed against a wall. The person sits in the chair. <i>"Sit on the chair. Interlace your fingers and reach forward with your arms. With your feet slightly apart and your hips at the edge of the chair seat, lift your hips up from the seat to stand. Slowly return to sitting."</i>
Squat (with a chair) comprises squat on a chair	Context for the squat on a chair: The person stands on front of the chair. <i>"Perform a squat by initiating the motion of pushing the hips back and then following this, by bending the knees. Keep your back straight and touch the wall lightly with your buttocks without resting on the wall. Come back up, keeping your back straight at all times."</i>

and partial squat	Context for a partial squat: A chair is placed in the middle of the room. The person stands behind the chair with their hands on it. <i>“Hold the chair in front of you with your hands. Perform a squat by initiating the motion. Push the hips back and then follow this by bending the knees. Keep your back straight and touch the wall lightly with your buttocks without resting on it. Come back up, keeping your back straight at all times.”</i>
Opposite arm and leg lift (with a chair)	Context: A chair is placed anywhere in the room. The person sits in the chair. <i>“Sit with your back in a neutral position (slightly arched); your chin must be tucked in. Slightly tighten your abdominals, lumbar muscles and pelvic floor muscles, then lift one arm and the opposite leg without allowing the trunk or pelvis to move or rotate.”</i>
Arm raises (with or without a chair)	Context for arms raises with a chair: A chair is placed against a wall. Ensure that there are no obstacles to either side of the chair. The person sits in the chair. <i>“Sit-up straight and lift both arms thumb up at the same time, keeping the thumbs up. Lift to shoulder's height.”</i>
	Context for arms raises without a chair: The person stands with more than an arm's length of space to their left and right. A chair is placed on front of them in the event that they may feel unsteady. <i>“Stand up and lift both arms thumb up at the same time. Lift to shoulder height. Ensure to maintain thumbs up.”</i>
Step up comprises Step up and Sidestep up	Context for step up: The person is standing. <i>“Stand up and take a step with one leg. Then bring the leg back to the starting position. Be careful of putting the foot entirely on the step.”</i>
	Context for sidestep up: The person is standing. <i>“Stand up and take a sidestep with one leg. Then bring the leg back to the starting position. Keep your pelvis levelled (making sure you do not hip hang).”</i>

**Table 1:** Set of exercises with associated instruction.

### 3.2 Technology choice and exergame design

Based on the results from our systematic literature review, we were able to determine guidelines and criteria for the technology choice and the exergame design. The comparison of four types of devices (Head-mounted display (HMD), body tracking camera, balance board and specific device) in this review and their evaluation through criteria such as: immersion, effectiveness in training, global body tracking, engagement, comfort, and interaction, guided us in selecting the Microsoft Hololens 2 HMD and the Azure Kinect body tracking camera. The exergame design took into account both the limitations arising from the target audience's vulnerability (related to osteoporosis) and the guidelines provided by existing literature and HSE. As such, design considerations included:

- The design of positive game features [Vanden, 2021; Blomqvist, 2021] i.e., positive, colourful, non-aggressive content (see Figure 1 for the prototype of each game) and immersive sounds.
- The possibility for content personalization [Andreikanich, 2019] through adapting the game to the needs of the player (e.g., type of virtual objects, level of difficulty etc.).
- Movement variability adaption based on the physical ability of the patient (see Section 3.3)
- The provision of positive and easy to understand instruction and feedback [Vanden, 2021].

Detail on the exergames is provided in Table 2, with the conceptual exergames' designs presented in Figure 1. During a training session, patients will participate in each of the five exergames. They will be able to choose whether they want to follow the pre-determined 'scenario' or select the exergames in any order. The 'Sit to stand' game is the only exergame presented in third person view to ensure a continuity of immersion, support interaction and mitigate against the risk of boredom during training.

Physical exercises	Conceptual exergames description	Camera position	Starting position
Sit to stand	Third person platform game. The participant moves a character from one platform to the next (up and down) by standing down and sitting.  Main goal: To progress on the path.	In front	Seated on a chair.  Position: Facing the virtual platform.

Squat	First person game. The participant is required to squat to avoid the cloud coming in front of his/her eyes.  Main goal: avoid the virtual object (e.g., cloud).	On the side	Seated on a chair.  Position: Facing the virtual object.
Opposite arm and leg lift	First person game. The participant is required to lift the correct body part (right/left arm/leg) to hit the ball coming in front of their eyes.  Main goal: Hit the virtual object (e.g., ball).	In front	Seated on a chair.  Position: Facing the virtual object.
Arm raises	First person game. The participant must raise his/her arms to help the (virtual) bird to fly in front of their eyes.  Main goal: Move the virtual object (e.g., bird).	In front	Seated on a chair.  Position: Facing the virtual object.
Step up	First person game. The participant must step over a virtual obstacle coming on the floor in front of them.  Main goal: Avoid the virtual object (e.g., rock).	In front	Standing.  Position: Facing the virtual object.

**Table 2: Conceptual exergame description.****Figure 1: Prototype of each exergame (from left to right: opposite arm leg lift, arm raises, squat, step up, sit to stand)**

### 3.3 ‘Good’ movement in physiotherapy and in games

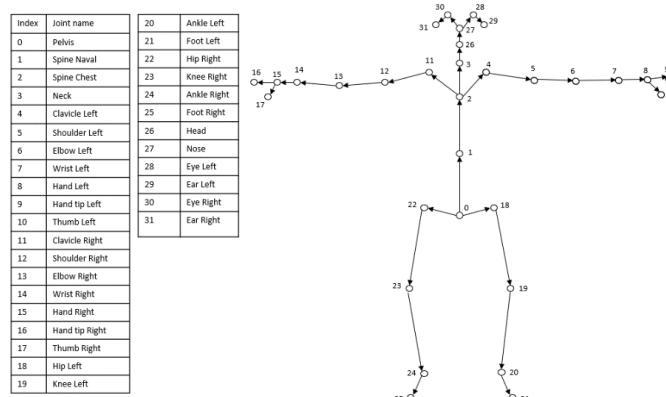
Notwithstanding the accruing positive cognitive and physical benefits of exercise, it is not atypical that participants can use improper techniques when performing exercises e.g., they are executed too fast, too intensely and/or with an incorrect position [ACSM, 2021]. A lack of adherence to correct posture, intensity and breathing can result in injuries [ACSM, 2021]. Consequently, whilst an ‘appropriate pace’ is critical, it might differ for each patient, as proper pace and alignment depend not only on the exercise but also on the ability of the patient. This is of particular importance in the context of frail patients. Through discussion with the chartered PT, ‘good’ movement is defined as exercise movement made with appropriate pace and alignment and which are suitable to the patient in question. Movement variability refers to the nonlinearity in movements which can occur over multiple repetitions [Harbourne, 2009]. In the context of physical therapy, it is important that a PT shouldn’t and won’t expect the patient in a training session, to repeat the movement in the exact same manner each time [Harbourne, 2009]. It is important to consider this concept of movement variability when designing both physical and virtual training as movement variability directly influences both the learning of the movement and the movement’s accessibility. To cater to movement variability, the human body is divided into biomechanical points which can be checked for alignment to ensure that the exercise is correctly executed. Biomechanics is the science of the movement of the human body [Raiola, 2020]. Table 3 presents the biomechanical points to be checked to ensure correct alignment to the five selected exercises used in our study (based on advice from the PT).

Exercise	Biomechanical point to check for alignment
Sit to stand	Spine - hip, Shoulder - hip (for osteoporotic patient)
Squat	Spine - hip, Shoulder - hip (for osteoporotic patient)
Arm raises	Elbow to wrist, No bent elbow

Opposite arm and leg lift	Shoulder-elbow-knee
Step up	Knee angle - hip, No need for a strictly exact movement

**Table 3: Biomechanical points to check for alignment for each exergame.**

The biomechanical points in Table 3 can be directly translated to the Azure Kinect, as the Azure Kinect body tracking SDK can track a participant's skeletal representation using 31 joints (Figure 2).

**Figure 2: Body joints with Azure Kinect body tracking**

In this research, correct alignment within each exergame will be measured by calculating the angle between the equivalent joint/biomechanical points. Care will be taken in leaving an appropriate threshold based on;

- Movement variability.
- Error and/or approximation (human and/or Azure Kinect).
- Ability of each participant - as not every patient might not be in the same physical condition.

Note: The PT will assess the patients in advance of using the exergames, in order to identify individual thresholds. The exergames will then be modified on the basis of this evaluation.

### 3.4. Implementation of PT's feedback in the Exergames

Feedback provides information and guidance to players, thereby impacting their learning, motivation and engagement [Boyer-Thurgood, 2017; Gautier, 2022; Johnson, 2022]. Feedback can be delivered through different modalities, including auditory, visual and haptic [Boyer-Thurgood, 2017; Gautier, 2022; Demain, 2012; Menelas, 2017]. The type of modality is dependent on the game's context and aim. As an example, auditory feedback may be more effective than visual feedback in highly visual games, as it will keep players' visual processing system from being overloaded [Gautier, 2022]. Feedback features (e.g., modality, duration, type etc.) are important as they respond to players' actions and provide information on their performance. Within the context of physical therapy, feedback from the PT is one of the key elements of effective physiotherapy. Indeed, [De Souza, 1990] highlighted the importance of the PT in providing a supportive and advisory role. Feedback from a PT to a patient during a training session can be divided into; direct feedback (i.e., verbal comments/advice on current movement, and physical re-positioning) and indirect feedback (i.e., instruction which is not directly related to the movement e.g., 'be careful that the chair is against the wall' and other safety cues). Focussing on the set of five exercises presented previously, Table 4 outlines direct and indirect in-person feedback given by a chartered PT during a training session. Table 4 also highlights how this in-person feedback can be translated to in-exergame feedback.

	General	Sit-to-Stand	Squat	Arm raises	Opposite Arm and Leg Lift	Step up
Example of in-person feedback	<b>Direct:</b> “Breathe”, <b>Indirect:</b> “Keep eye contact with the PT”, “Ensure a safe space around you during training”	<b>Direct:</b> “Align your forehead to your knee”, “Sit up straight” <b>Indirect:</b> “Use a chair against a wall to prevent it from slipping”	<b>Direct:</b> “Keep your back straight”	<b>Direct:</b> “Squeeze your shoulder and keep your tummy/stomach back”, “Extend your elbow” If seated, “Sit up straight”	<b>Direct:</b> “Sit up straight”	<b>Direct:</b> “Take your time”, <b>Indirect:</b> “Check your foot position while stepping”, “Do you have need of other supports?”
Example of in-game feedback	<b>Direct:</b> “Breathe”, “Don’t hesitate to take a break if necessary”, <b>Indirect:</b> “One more...”, “You are doing well”, etc.	<b>Direct:</b> “Remember to sit up straight”, <b>Indirect:</b> “Be careful that your chair is placed with the back against a wall”	<b>Direct:</b> “Remember to keep your back straight”	<b>Direct:</b> “Remember to sit up straight”, “Remember to stand straight”	<b>Direct:</b> “Remember to sit up straight”, <b>Indirect:</b> “Well done, now repeat it with the right/left leg/arm”	<b>Indirect:</b> “Take your time to step over the rock”

Table 4: Feedback from the PT to the patients.

Visual feedback in the exergames will be provided by:

- Displaying a score to indicate performance (e.g., number of repetitions performed) and success (e.g., number of correctly performed repetitions).
- Providing feedback on the current posture (e.g., ‘Maintain the pose for two more seconds’) and/or participation in the game (e.g., ‘Keep your arms raised to allow the bird to fly higher’).
- Providing real-time visual reaction to the action on the screen (e.g., in the arm raise exercise, the bird may fly in reaction to the player’s movement. In the sit to stand exercise, the virtual character moves from one block to the next).

Meanwhile, audio feedback using music, and sounds will enhance the player’s experience and provide additional cues and/or feedback related to the activity or game being played.

#### 4. Conclusions

Worldwide, osteoporosis is a significant challenge for more than 200 million people. Existing research shows that virtual rehabilitation and exergames present promising opportunities to enhance traditional physical therapy by increasing motivation and enjoyment. By translating physical therapy into exergames to be played using an Hololens 2 HMD and Azure Kinect, this research focuses on the design of an adapted physical training programme for older adults with osteoporosis. To ensure the suitability of the exergames, we integrated HSE care guidelines and the expertise of health and social care experts (i.e., PTs, OTs and Rheumatologists). Underpinned by their recommendations, a set of five safe exercises was selected to improve players’ balance, muscle strength and flexibility. These physical exercises were then mapped onto a series of conceptual exergames. A systematic literature review on virtual rehabilitation provided guidelines regarding the choice of the technology and the design of the exergames, within the context of usability, comfort, and personalization. Based on this review and technology evaluation via criteria such as; immersion, effectiveness in training, global body tracking, engagement, comfort and interaction, the Microsoft Hololens 2 HMD and Azure Kinect body tracking camera were selected. Meanwhile, ongoing interactions with a chartered PT supported the definition of ‘good’ movement i.e. movement made with a proper alignment at a proper pace. This definition will help us to define correct movement in the exergame

environment. Moreover, the calculation of the angle between the equivalent joint/biomechanical points for each exercise will help to ensure correct alignment.

In this paper, we also discussed the importance of feedback in supporting patients in their training. We propose to use both visual and audio feedback in the exergames. Written exergame feedback will be based on direct feedback given by a PT during a physical training session. In future work, we plan to investigate the impact of a training programme where patients with osteoporosis engage with the exergames, for twice weekly 20-minutes sessions over a period of six weeks.

## Acknowledgements

This work was conducted with the financial support of the SFI Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

## References

- [ACSM, 2021] American College of Sports Medicine (2021), ACSM's resource manual for Guidelines for exercise testing and prescription, 11<sup>th</sup> Edition. [accessed on the 26/06/2023]
- [Alnajjar, 2019] Alnajjar F, Jishtu N, Alsinglawi B, Al Mahmud A, Mubin, O. Exoskeletons With Virtual Reality, Augmented Reality, and Gamification for Stroke Patients' Rehabilitation: Systematic Review. MIREhabilitation and assistive technologies, 6(2), (2019). <https://doi.org/doi.org/10.2196/12010>
- [Amanda, 2014] Amanda E. Staiano and Rachel Flynn. Therapeutic Uses of Active Videogames: A Systematic Review. Games for Health Journal.351-365 (2014). <https://doi.org/10.1089/g4h.2013.0100>
- [Andreikanich, 2019] Andreikanich, A., Sousa Santos, B., Amorim, P., Sagalo, H., Marques, B., Margalho P., Laíns, J., Faim, F., Coelho, M., Cardoso, T., Dias, P. (2019) An Exploratory Study on the use of Virtual Reality in Balance Rehabilitation. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, pages 3416–3419. <https://doi.org/10.1109/EMBC.2019.8857469>
- [Bouchard, 2014] Bouchard S, Baus, O. Moving from virtual reality exposure-based therapy to augmented reality exposure-based therapy: a review. Frontiers in human neuroscience, 8, (2014). <https://doi.org/10.3389/fnhum.2014.00112>
- [Boyer-Thurgood, 2017] Boyer-Thurgood, J. M., The Anatomy of Virtual Manipulative Apps: Using Grounded Theory to Conceptualize and Evaluate Educational Apps that Contain Virtual Manipulatives. All Graduate Theses and Dissertations. 6178. (2017). <https://digitalcommons.usu.edu/etd/6178>
- [Blomqvist, 2021] Blomqvist, S., Seipel S, Engstrom M. (2021) Using augmented reality technology for balance training in the older adults: a feasibility pilot study. BMC geriatrics, 21(1):144. <https://doi.org/10.1186/s12877-021-02061-9>
- [Brosna, 2020] Brosna, A., Easy exercises: A chair based programme for older adults, <https://www2.hse.ie/living-well/exercise/indoor-exercises-older-people/sitting-exercises/> [accessed on the 22/06/2023]
- [Demain, 2012] Demain, S., Metcalf, C. D., Merrett, G. V., Zheng, D., & Cunningham, S. A narrative review on haptic devices: relating the physiology and psychophysical properties of the hand to devices for rehabilitation in central nervous system disorders. Disability and Rehabilitation: Assistive Technology, 8(3), 181–189 (2012). doi: <https://doi.org/10.3109/17483107.2012.697532>
- [De Souza, 1990] De Souza, L.H. (1990). Physiotherapy. In : Multiple Sclerosis. Therapy in Practice Series, vol 18. Springer, Boston, MA. [https://doi.org/10.1007/978-1-4899-3107-8\\_4](https://doi.org/10.1007/978-1-4899-3107-8_4)
- [Dionyssiotis, 2014] Dionyssiotis, Y., Skarantavos, G., Papagelopoulos, P. Modern rehabilitation in osteoporosis, falls, and fractures. Clinical medicine insights. Arthritis and musculoskeletal disorders, 7, 33–40 (2014). <https://doi.org/10.4137/CMAMD.S14077>

- [Doyle, 2011] Doyle, J., Kelly, D., Patterson, M., and Caulfield, B. (2011) The effects of visual feedback in therapeutic exergaming on motor task accuracy. 2011 International Conference on Virtual Rehabilitation. Zurich, Switzerland, page 1-5. <https://doi.org/10.1109/ICVR.2011.5971821>
- [Gauthier, 2022] Gauthier, A., Benton, L., Bunting, L., Herbert, E., Sumner, E., Mavrikis, M., Revesz, Vasalou, A. I Don't Usually Listen, I Read: How Different Learner Groups Process Game Feedback. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 88, 1–15 (2022). <https://doi.org/10.1145/3491102.3517480>
- [Harbourne, 2009] Harbourne, R. T., and Stergiou N. (2009) Movement Variability and the Use of Nonlinear Tools: Principle to Guide Physical Therapist Practice, Physical Therapy, vol 89, issue 3, page 267-282. <https://doi.org/10.2522/ptj.20080130>
- [Johnson, 2017] Johnson, C.I., Bailey, S.K.T., Van Buskirk, W.L. Designing Effective Feedback Messages in Serious Games and Simulations: A Research Review. In: Wouters, P., van Oostendorp, H. (eds) Instructional Techniques to Facilitate Learning and Motivation of Serious Games. Advances in Game-Based Learning. Springer, Cham. (2017) [https://doi.org/10.1007/978-3-319-39298-1\\_7](https://doi.org/10.1007/978-3-319-39298-1_7)
- [Menelas, 2017] Menelas B-AJ, Benaoudia RS. Use of Haptics to Promote Learning Outcomes in Serious Games. Multimodal Technologies and Interaction. 1(4), 31 (2017). <https://doi.org/10.3390/mti1040031>
- [McCabe, 2020] McCabe, E., Ibrahim, A., Singh, R. et al. A systematic review of the Irish osteoporotic vertebral fracture literature. Arch Osteoporos 15, 34 (2020). <https://doi.org/10.1007/s11657-020-0704-0>
- [Peng, 2011] Peng W. Lee K. M. Song, H. Promoting exercise self-efficacy with an exergame. Journal of health communication, 16(2):148–162, (2011). <https://doi.org/10.1080/10810730.2010.535107>
- [Obuko, 2018] Okubo Y. Woodbury A. Lord S. R. Delbaere K. Valenzuela, T. (2018). Adherence to Technology-Based Exercise Programs in Older Adults: A Systematic Review. Journal of geriatric physical therapy, 41(1):49–61 <https://doi.org/10.1519/JPT.0000000000000095>
- [Oh, 2010] Oh, Y., Yang, S. (2010) Defining exergames and exergaming. Proceedings of Meaningful Play, Easi Lansing MI. pages 1-17.
- [Lee, 2017] Lee, S., Kim, W., Park, T., and Peng, W. (2017) The psychological effects of playing exergame: A Systematic Review. Cyberpsychology, Behavior, and Social Networking, vol 20, issue 9, pages 213-532.
- [Raiola, 2020] Raiola, G., Domenico, F. D., Isanto, T. D., Altavilla, G., & Elia, F. D. (2020). Biomechanics core. Acta Medica Mediterranea, 36(5), 3079-3083.
- [Sözen, 2017] Sözen, T., Özışık, L., & Başaran, N. Ç. (2017). An overview and management of osteoporosis. European journal of rheumatology, 4(1), 46–56. <https://doi.org/10.5152/eurjrheum.2016.048>
- [Vanden, 2021] Vanden, A. V., Schraepen, B., Huygelier, H., Gillebert, C., Gerling, K., Van Ee, R. (2021). Immersive Virtual Reality for Older Adults: Empirically Grounded Design Guidelines. ACM Trans. Access. Comput. 14, 3, Article 14, 30 pages. <https://doi.org/10.1145/3470743>
- [Whatman, 2012] Whatman, C., Hing, W., and Hume, P. (2012) Physiotherapist agreement when visually rating movement quality during lower extremity functional screening tests, Physical Therapy in Sport, vol 13, issue 2, pages 87-93, ISSN 1466-853X. <https://doi.org/10.1016/j.ptsp.2011.07.001>.

# Designing the VR Probe: An Introductory Application to Virtual Reality for People Living with Dementia (PLWD)

Gearóid Reilly<sup>1</sup>, Gabriel-Miro Muntean<sup>2</sup>, Aisling Flynn<sup>3</sup>, Attracta Brennan<sup>1</sup>, Sam Redfern<sup>1</sup>

<sup>1</sup>*University of Galway, School of Computer Science, Galway, Ireland*

<sup>2</sup>*Dublin City University, School of Electronic Engineering, Dublin 9, Ireland*

<sup>3</sup>*University of Galway, School of Nursing and Midwifery, Galway, Ireland*

## Abstract

Dementia is a condition which involves a gradual degradation of cognitive functioning over time. The number of people living with dementia (PLWD) is increasing every year. The use of Virtual Reality (VR) in health related research has gained popularity, in part due to the availability of VR headsets such as the Oculus Quest and the HTC Vive and the immersive nature of VR. VR interventions are recognised as having great potential for PLWD. To ensure effective VR based experiments involving PLWD, it is critical to introduce this audience to VR. While there are commercially available introduction applications, none have been designed with PLWD in mind. This paper describes the design considerations made for a VR Probe application having a dual purpose: to introduce PLWD to VR and to help inform design decisions for future VR applications. This paper also describes the technological implementation decisions for a replicable VR environment such as task design, interaction system, multisensory design and virtual representation.

**Keywords:** Virtual Reality, Older Adults, Dementia, Immersion, Simulation

## 1 Introduction

Virtual Reality (VR) is a form of technology that uses a head mounted display (HMD) to display an artificial 3D environment, utilising a combination of visual, audio and in some cases haptic feedback to provide an immersive experience to the user [Gigante, 1993]. Research using VR has increased due to the greater accessibility of HMD technology, with one such area of research being the use of VR for people living with dementia (PLWD).

Dementia is a neurodegenerative condition which leads to a gradual decline in cognitive functioning over time [WHO, 2020]. Globally, over 55 million people have been diagnosed with dementia, with 10 million new people being diagnosed every year [WHO, 2020]. While currently there is no cure for dementia [WHO, 2020], research has found that non-pharmacological activities (e.g. reminiscence therapy and light exercise) can help to enrich the lives of PLWD [Shigihara et al., 2020]. Furthermore, studies show that technology (e.g. VR) provides opportunities to support PLWD from assessment, to monitoring to the support of cognitive and physical functioning [Astell et al., 2019]. Examples of VR interventions include: the use of adaptive music therapy to reduce negative emotions and increase memory performance in PLWD [Byrns et al., 2020] and the use of VR for ‘light’ physical activities (e.g. virtual bowling) to support PLWD in recalling motor tasks [Fenney and Lee, 2010]. The capability of VR interventions to provide old and new experiences for PLWD without the difficulties of travel [Möhler et al., 2020] has become increasingly important after identifying PLWD’s heightened feelings of isolation as a result of the Covid-19 restrictions in 2020 [Curelaru et al., 2021].

As 91% of PLWD are diagnosed after the age of 65 years [WHO, 2020], PLWD who are being involved in VR research likely have no prior experience using the technology [Flynn et al., 2022b]. As such, there is a need to introduce them to VR as part of the research process. While there are commercially available VR introduction applications such as ‘First Steps’ or ‘First Contact’ developed by Oculus, these applications are designed for a general audience and do not contain effective design considerations for older adults [Abeele et al., 2021].

This paper describes the design decisions of a VR probe: an application to introduce PLWD to VR and its basic functionality. This paper aims to describe the technical implementation and development considerations for the VR probe. Gaver et al (1999) defines a cultural probe as a design led approach to research where participants are shown design artefacts to promote ideas and discussion between participants and researchers to gain a better understanding of the research field and promote new design ideas [Gaver et al., 1999]. The aim of this probe is to ensure that such PLWD would be able to experience VR to better inform design decisions for future VR applications [Hutchinson et al., 2003].

## 2 System Architecture

The VR probe was created using the Unity3D game engine. This engine was chosen due to its popularity for creating cross platform 2D and 3D environments [González et al., 2017]. Unity also supports the creation of VR applications through its XR Interaction Toolkit [Unity3D, 2023]. Using the XRRig object provided by this toolkit allowed a quick implementation of the user's virtual representation as a part of a VR setting. The VR probe environment was created using a living room model developed in LiveHome2D which was then imported into Unity. Unity3D uses the toolkit's interaction system to manage interactions between virtual objects and the user. The probe environment was designed to function on the Oculus Quest 2 HMD as the device can run the application without being connected to a computer. This was important as the portability of the Oculus Quest 2 allows it to be easily used by PLWD regardless of where they are situated (e.g. sitting on a chair, in bed, etc [Saredakis et al., 2021]), and can be applied in home or healthcare settings.

## 3 Task Design

The Responsive, Engaging, Augmenting, Failure-Free (REAFF) framework defines a set of principles to guide the design of technology to support PLWD [Astell, 2009]. With this framework in mind, several tasks were developed for the VR probe to reflect common actions within VR. These tasks consisted of; (1) selection - where the participant is required to select one of the three cubes on the table (Figure 1) using the controller, (2) object placement - where the participant is required to use the controller to pick up the ball on the table and place it within the container on the table (Figure 2), (3) gaze - where the participant is required to look at a screen and watch a video until it has concluded (Figure 3), and (4) teleportation - where the participant is instructed to direct the controller to a specified point on the ground and press the trigger button to move to that point (Figure 5). To promote autonomy in PLWD and in line with the REAFF guidelines, the tasks were designed with simple functionality and flexibility in mind (i.e. each task was designed for easy completion: all tasks could be completed in no particular order [Hutchinson et al., 2003]). The tasks were developed by placing game objects within the environment and then attaching interactable components which allowed these objects to be picked up or interacted with by the player. Components were also attached to the objects to call an event when the task was completed.



**Figure 1: Selection Task**



**Figure 1: Object Placement Task**



**Figure 3: Gaze Task**

When a participant approaches a task, a user interface (UI) panel appears displaying task instructions. This is in line with Braley et al. (2018) who suggested that the use of prompting in technology can improve the functional independence of PLWD [Braley et al., 2018]. When the task completion event is called, the UI panel changes to an encouraging message (Figure 4) as studies show that PLWD respond faster to positive feedback [Maki et al., 2018]. The foreground and background colours were specifically selected to provide appropriate visual contrast [McIntyre et al., 2019].



**Figure 4: Encouraging Message on the UI panel**

## 4 Interaction and Locomotion

All Interaction and locomotion decisions relating to the VR probe were underpinned by the fact that a higher level of interaction fidelity can lead to a greater sense of immersion within the virtual environment [Rogers et al., 2019]. To avoid physical strain on the PLWD participants as they interacted with the virtual environment (e.g. picking up objects similar to how they would do so in the real world using the provided controllers), interaction rays were added to the hands to allow objects to be grabbed from a distance. This is in line with findings by Argelaguet and Andujar (2013) who identified raycasting selection as one of the most popular techniques for its effectiveness and ease of use [Argelaguet and Andujar, 2013]. The interaction rays point out from the hands and highlight when an object is interactable (Figure 5). To reduce the cognitive load of having to memorise different buttons and associated actions, the trigger button is assigned to grabbing, teleporting and interaction actions. An extra interaction ray, attached to the virtual head triggers gaze interactions when the participant looks at the video-screen virtual object (Figure 3).

As locomotion in VR allows participants to move their virtual representation in the VR space [Bozgeyikli et al., 2016], this enables the creation of larger VR environments that are not limited to the boundary set by the HMD. The most common forms of VR locomotion are joystick movement and teleportation [Buttussi and Chittaro, 2021], however other methods such as arm cycling have also been developed [Coomer et al., 2018]. As joystick movement can cause cybersickness and arm-cycling is tiring for participants, we integrated teleportation into the VR probe for its ease of use. [Coomer et al., 2018]. Teleportation anchor points were placed beside each task space, so that once participants pointed their hand rays at the red squares and pressed the trigger button, they would be instantaneously transported to that location. The teleportation anchors change colour once the interaction ray hovers over them, in order to inform the participant to which position they can teleport (Figure 5). The teleportation anchors contain components to detect interactions from a teleportation provider component attached to the XRRig.



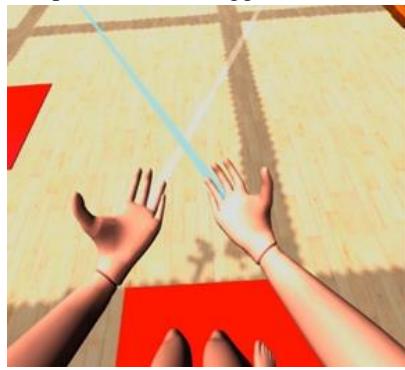
**Figure 5: Teleportation visual cues**

## 5 Multisensory Design

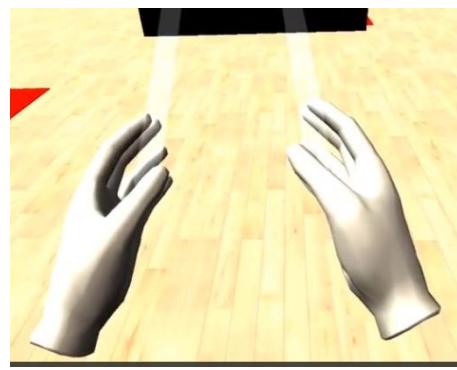
VR can be utilised to provide a multisensory, immersive experience for participants which can lead to a more meaningful time in the virtual world [Abbeel et al., 2021]. The combination of visual, audio and haptic stimuli can provide a greater sense of immersion [Muñoz et al., 2022]. Furthermore, the use of multisensory VR design supports participants with impaired senses (i.e. cognitive and/or physical) [Hodge et al., 2018]. Within the context of the VR probe, the video component of the gaze task (Figure 3) provides both visual and audio stimuli to the participants. Furthermore, as the participant hovers over interactable objects, an event is called which adds a distinctly coloured outline or changes the material colour (Figures 1 and 2). The interaction rays also change colour as they hover over the interactable objects. In addition to the visual changes, the controllers vibrate during the hovering event. Through the use of both visual and haptic feedback, the participant is informed of the interactable nature of the object through pressing the trigger button. This haptic feedback is alongside visual cues (Figure 5) which indicate to the participant the place to which they can teleport.

## 6 Virtual Representation

A virtual avatar is a visual representation of the participant while they are inside the VR environment [Hodge et al., 2018]. Avatars can have full bodies, half bodies or virtual hands. Studies showed that for younger adults, the inclusion of an avatar can improve performance in cognitive tasks regardless of the type of avatar used [Pan and Steed, 2019]. PLWD are able to accept the presence of virtual hands as their own while they are using VR [Matsangidou et al., 2020]. Mendez et al. (2015) demonstrated that PLWD can accept the presence of other virtual avatars within their virtual environment [Mendez et al., 2015]. In the VR probe, a full body avatar was initially used (Figure 6). However, prior to testing on PLWD, the use of a full body avatar was found to cause disembodyment due to the inconsistency in movement between the virtual and physical forearms. As a result, virtual hands were chosen as they move consistently with the controllers (Figure 7). Hand models were downloaded from the Oculus Developer website and attached to the hand interactors on the XRRig. Animator components were also added to add realism to object grabbing interactions, by changing the pose of the hands when the participants held the trigger button.



**Figure 6: Full Body Avatar**



**Figure 7: Virtual Hands**

## 7 Contributions to Research

Underpinned by an extensive literature review and in collaboration with an occupational therapist, a VR probe was designed, developed and tested by nine PLWD, along with their care-givers. While previous VR studies involving PLWD have mentioned the technology which was used, they rarely go into sufficient technical detail for other researchers to replicate the experimental environment for further research. This paper describes the design and implementation decisions which scaffolded the VR probe and which can be replicated by future researchers. Note: The VR probe was used in a study which introduced PLWD and their care-givers to VR [Flynn et al., 2022a]. The findings from this study were used to further refine the probe e.g. implementing graphical anti-aliasing to ensure that the graphics appeared smoother, and including background audio to improve the sense of realism within the virtual environment.

## 8 Future Work

This paper details the design and technical implementation decisions for a VR probe for PLWD. Feedback from the use of this probe as outlined in [Flynn et al., 2022a] can be used to inform the design of future VR applications for PLWD. As the design and technical implementation considerations of the VR probe were made with older PLWD in mind, this probe can also be used to introduce older adults without dementia to VR. It should be noted that gameplay based data was not gathered during the probe's testing period as its primary purpose was to introduce PLWD to VR. However, future uses of the VR probe will include both quantitative (i.e. game play data) and qualitative (e.g. interview) data in order to gain more specific insight into how PLWD both navigate VR environments and engage with task based VR activities.

## 9 Conclusion

With the number of dementia diagnoses increasing every year, the need for effective interventions to improve the quality of life for PLWD is becoming more necessary. While VR has proven to be an effective tool for non-pharmacological interventions, it is clear that PLWD need an appropriate introduction to VR. This paper described the design considerations underpinning the VR probe for introducing PLWD to VR within the context of interaction, locomotion and immersion. In addition to outlining the underpinning design considerations, this paper also presented the technology architecture and implementation decisions of the VR probe.

## Acknowledgements

This work was conducted with the financial support of the Science Foundation Ire-land Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

## References

- [Abeele et al., 2021] Abeele, V. V., Schraepen, B., Huygelier, H., Gillebert, C., Gerling, K. & EE, R. V. 2021. Immersive Virtual Reality for Older Adults: Empirically Grounded Design Guidelines. *ACM Trans. Access. Comput.*, 14, Article 14.
- [Argelaguet and Andujar, 2013] Argelaguet, F. & Andujar, C. 2013. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics*, 37, 121-136.
- [Astell, 2009] Astell, A. 2009. REAFF - A framework for developing technology to address the needs of people with dementia. *CEUR Workshop Proceedings*, 499, 5-10.
- [Astell et al., 2019] Astell, A. J., Bouranis, N., Hoey, J., Lindauer, A., Mihailidis, A., Nugent, C. & Robillard, J. M. 2019. Technology and Dementia: The Future is Now. *Dementia and Geriatric Cognitive Disorders*, 47, 131-139.
- [Bozgeyikli et al., 2016] Bozgeyikli, E., Raij, A., Katkoori, S. & Dubey, R. 2016. Point & Teleport Locomotion Technique for Virtual Reality. *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. Austin, Texas, USA: Association for Computing Machinery.
- [Braley et al., 2018] Braley, R., Fritz, R., Van Son, C. R. & Schmitter-Edgecombe, M. 2018. Prompting Technology and Persons With Dementia: The Significance of Context and Communication. *The Gerontologist*, 59, 101-111.
- [Buttussi and Chittaro, 2021] Buttussi, F. & Chittaro, L. 2021. Locomotion in Place in Virtual Reality: A Comparative Evaluation of Joystick, Teleport, and Leaning. *IEEE Transactions on Visualization and Computer Graphics*, 27, 125-136.
- [Byrns et al., 2020] Byrns, A., Ben Abdessalem, H., Cuesta, M., Bruneau, M.-A., Belleville, S. & Frasson, C. Adaptive Music Therapy for Alzheimer's Disease Using Virtual Reality. 2020 Cham. Springer International Publishing, 214-219.
- [Coomer et al., 2018] Coomer, N., Bullard, S., Clinton, W. & Williams-Sanders, B. Evaluating the effects of four VR locomotion methods. *Proceedings of the 15th ACM Symposium on Applied Perception*, 2018-08-10 2018. ACM.
- [Curelaru et al., 2021] Curelaru, A., Marzolf, S. J., Provost, J.-C. K. G. & Zeon, H. H. H. 2021. Social Isolation in Dementia: The Effects of COVID-19. *The Journal for Nurse Practitioners*, 17, 950-953.
- [Fenney and Lee, 2010] Fenney, A. & Lee, T. D. 2010. Exploring Spared Capacity in Persons With Dementia: What WiiTM Can Learn. *Activities, Adaptation & Aging*, 34, 303-313.
- [Flynn et al., 2022a] Flynn, A., Barry, M., Qi Koh, W., Reilly, G., Brennan, A., Redfern, S. & Casey, D. 2022a. Introducing and Familiarising Older Adults Living with Dementia and Their Caregivers to Virtual Reality. *International Journal of Environmental Research and Public Health*, 19, 16343.
- [Flynn et al., 2022b] Flynn, A., Healy, D., Barry, M., Brennan, A., Redfern, S., Houghton, C. & Casey, D. 2022b. Key Stakeholders' Experiences and Perceptions of Virtual Reality for Older Adults Living With Dementia: Systematic Review and Thematic Synthesis. *JMIR Serious Games*, 10, e37228.

- [Gaver et al., 1999] Gaver, B., Dunne, T. & Pacenti, E. 1999. Design: Cultural probes. *Interactions*, 6, 21-29.
- [Gigante, 1993] Gigante, M. A. 1993. 1 - Virtual Reality: Definitions, History and Applications. In: Earnshaw, R. A., Gigante, M. A. & Jones, H. (eds.) *Virtual Reality Systems*. Boston: Academic Press.
- [González et al., 2017] González, J. D., Escobar, J. H., Sánchez, H., Hoz, J. D. L. & Beltrán, J. R. 2017. 2D and 3D virtual interactive laboratories of physics on Unity platform. *Journal of Physics: Conference Series*, 935, 012069.
- [Hutchinson et al., 2003] Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., Evans, H., Hansen, H., Roussel, N. & Eiderbäck, B. 2003. Technology probes: inspiring design for and with families. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Ft. Lauderdale, Florida, USA: Association for Computing Machinery.
- [Maki et al., 2018] Maki, Y., Sakurai, T., Okochi, J., Yamaguchi, H. & Toba, K. 2018. Rehabilitation to live better with dementia. *Geriatrics & Gerontology International*, 18, 1529-1536.
- [Matsangidou et al., 2020] Matsangidou, M., Schiza, E., Hadjiaros, M., Neokleous, K. C., Avraamides, M., Papaianni, E., Frangoudes, F. & Pattichis, C. S. 2020. Dementia: I Am Physically Fading. Can Virtual Reality Help? Physical Training for People with Dementia in Confined Mental Health Units. *Lecture Notes in Computer Science*. Springer International Publishing.
- [McIntyre et al., 2019] McIntyre, A., Harding, E., Yong, K. X. X., Sullivan, M. P., Gilhooly, M., Gilhooly, K., Woodbridge, R. & Crutch, S. 2019. Health and social care practitioners' understanding of the problems of people with dementia-related visual processing impairment. *Health & Social Care in the Community*, 27, 982-990.
- [Mendez et al., 2015] Mendez, M. F., Joshi, A. & Jimenez, E. 2015. Virtual reality for the assessment of frontotemporal dementia, a feasibility study. *Disability and Rehabilitation: Assistive Technology*, 10, 160-164.
- [Möhler et al., 2020] Möhler, R., Renom, A., Renom, H. & Meyer, G. 2020. Personally tailored activities for improving psychosocial outcomes for people with dementia in community settings. *Cochrane Database of Systematic Reviews*.
- [Muñoz et al., 2022] Muñoz, J., Mehrabi, S., Li, Y., Basharat, A., Middleton, L. E., Cao, S., Barnett-Cowan, M. & BOGER, J. 2022. Immersive Virtual Reality Exergames for Persons Living With Dementia: User-Centered Design Study as a Multistakeholder Team During the COVID-19 Pandemic. *JMIR serious games*, 10, e29987-e29987.
- [Pan and Steed, 2019] Pan, Y. & Steed, A. 2019. Avatar Type Affects Performance of Cognitive Tasks in Virtual Reality. *25th ACM Symposium on Virtual Reality Software and Technology*. Parramatta, NSW, Australia: Association for Computing Machinery.
- [Rogers et al., 2019] Rogers, K., Funke, J., Frommel, J., Stamm, S. & Weber, M. 2019. Exploring Interaction Fidelity in Virtual Reality: Object Manipulation and Whole-Body Movements. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland UK: Association for Computing Machinery.
- [Saredakis et al., 2021] Saredakis, D., Keage, H. A., Corlis, M., Ghezzi, E. S., Loffler, H. & Loetscher, T. 2021. The effect of reminiscence therapy using virtual reality on apathy in residential aged care: Multisite nonrandomized controlled trial. *Journal of medical Internet research*, 23, e29210.
- [Shigihara et al., 2020] Shigihara, Y., Hoshi, H., Shinada, K., Okada, T. & Kamada, H. 2020. Non-pharmacological treatment changes brain activity in patients with dementia. *Scientific Reports*, 10, 6744.
- [Unity3D, 2023] Unity3D. 2023. *XR Interaction Toolkit* [Online]. Available: <https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@2.3/manual/index.html> [Accessed 07/05/2023].
- [WHO, 2020] WHO. 2020. *Dementia* [Online]. World Health Organization. Available: <https://www.who.int/news-room/fact-sheets/detail/dementia> [Accessed 30/12 2020].

# Evaluation of Explainable AI Localisation Performance Using Relevance F-Score

Gregory Balogh, Niall McLaughlin, Austen Rainer

*Queen's University Belfast*

## Abstract

Explainable AI (XAI) has gained significant attention in the AI research community. However, the absence of a standard method for measuring the quality of explanations, and the diversity of existing XAI methodologies, make it hard to compare the performance of different XAI methods. An essential property of image-based XAI methods is their localisation performance, which describes how well the XAI method identifies the relevant object of interest. Existing methods for measuring XAI localisation performance do not consider the full XAI output, including image regions with negative saliency. To address the limitations of existing XAI localisation metrics, we introduce the Relevance F-score (RFS), which considers image regions with both positive and negative saliency to better capture an XAI method's localisation performance. We evaluate our approach using a Visual Question Answering (VQA) network trained on CLEVR-XAI dataset.

**Keywords:** Explainable AI Evaluation, Computer Vision, Visual Question Answering

## 1 Introduction

Explainable AI (XAI) has gained substantial attention in research and industry. The effectiveness of XAI methodologies is becoming increasingly important with the emergence of new methods [Huang et al., 2020, Yuan et al., 2020, Wang et al., 2020]. However, evaluating these methods objectively presents a challenge [Krishna et al., 2022, Zhou et al., 2021]. This study focuses on computer vision tasks and uses VQA to compare several XAI methods. The attention maps generated using commonly applied state-of-the-art techniques are analysed, and the inefficiencies of current evaluation metrics are demonstrated. The proposed explanation methods enable the creation of heat maps that portray an image's positive and negative attributions, revealing insights into the neural network's attention. Saliency maps help identify the essential pixels for accurate machine learning predictions. The current metrics only consider the positive attributions of the heat maps while neglecting the negative attributions. This approach limits the applicability of these metrics, as the location of negative attributions can indicate sub-optimal model performance. To overcome this limitation, we apply recall and F-score to assess neural network attention maps comprehensively.

## 2 Related Work

There exist many ways to evaluate the localisation quality of an explanation based on the location of the feature attributions with respect to the ground truth. Pointing game [Zhang et al., 2018] is a top-down visual attention model which extracts the maximum value on the attention map and assesses its position relative to the ground truth mask. If the attribution and the object of interest overlap, it's a hit; otherwise, it is a miss. The overall accuracy of the XAI method is calculated as the total number of hits relative to all explanations. Since this metric utilises a singular highest-value pixel, it cannot provide enough depth regarding the location of other attributions. Attribution localisation [Kohlbrenner et al., 2020] measures the ratio of positive attributions on the

ground truth mask relative to the total number of positive attributions. Top-K intersection [Theiner et al., 2022] calculates the intersection between the ground truth mask and the location of the top-K (1,000 by default) feature attributions. AUC (Area under the ROC curve) [Fawcett, 2006] compares the ratio of the explanations' true and false positive rates. And Focus [Arias-Duart et al., 2021] quantifies the precision of the explanation by creating mosaics of data instances from different classes. Essentially, it estimates the reliability of explanations using positive salience.

The methods most closely related to our work are Relevance Mass Accuracy and Relevance Rank Accuracy [Arras et al., 2020]. Relevance Mass Accuracy [Arras et al., 2020] computes the ratio of the positively attributed pixels inside the ground truth mask towards the overall positive attributions. Similar to Attribution localisation [Kohlbrenner et al., 2020], Relevance Mass Accuracy is mathematically equivalent to precision. Relevance Rank Accuracy [Arras et al., 2020], sorts the positive attributions of the explanation by their magnitude in descending order and calculates the ratio of the most highly attributed pixels inside the bounding box. Unlike in the Top-K intersection [Theiner et al., 2022], Relevance Rank Accuracy has no static value for the number of attributions to consider. The number of pixels considered is determined by the size of the ground truth mask. Top-K intersection and Relevance Rank Accuracy only consider a subsection of pixels with positive attributions.

The importance of positive attributions is undeniable since these pixels contribute to the prediction. On the other hand, the current metrics do not offer insights based on the location of negative attributions. Negative pixel attributions are features that negatively impact the predictive model performance and can result in an incorrect prediction if they overlap with the object of interest. The negative attributions can be more meaningful if the aim is to understand causes of sub-optimal model performance. The above-mentioned metrics, such as Relevance Mass Accuracy, are mathematically equivalent to precision. However, precision does not consider negative attributions and is inadequate when we have imbalanced classes [Reio Jr, 2016]. In addition, high precision does not automatically translate to accurate predictions or good-quality explanations because the negative attributions, which are not considered, also affect the model's performance. This, in turn, prevents a deeper understanding of the models' attention with respect to the object of interest. In this work, we propose a new method to assess the localisation accuracy of an XAI methods that addresses these limitations.

### 3 Methodology

In this work, we propose a new way to compare the localisation performance of XAI methods. Our approach is based on comparing the saliency map produced by an XAI method with ground truth information indicating the most important object(s) in the scene. From our review of the literature (See Section 2), we can see it is challenging to compare different XAI methods.

Assume we are given a Visual Question Answering (VQA) model that takes an image as input and produces an answer. Also, assume we have ground truth information indicating the relevant parts of the image needed to answer the question. An XAI method will produce a saliency map showing where the classifier's attention is focused. The saliency map can have positive and negative values, depending on which areas contribute positively or negatively towards the VQA model's final decision. In the ideal case, the positive areas of the saliency map will exactly cover the object(s) of interest and have no overlap with the background or other objects. To evaluate the localisation performance of the XAI method, we measure how well the saliency map matches the relevant object(s) in the ground truth mask. In Fig. 1, we show how the problem of evaluating the localisation performance of an XAI method can be modelled as a classification problem. We assume that for each input, ground truth information is available that specifies the locations of the relevant object(s). Then, for a given saliency map, we can then define true-positive area, true-negative area, false-positive area and false-negative areas as follows:

- **True Positive Area (TPA)** - Regions with a positive saliency value located inside the ground truth mask of the object of interest. These pixels are expected to correspond to the relevant area(s) that the machine learning model should use to make correct predictions.

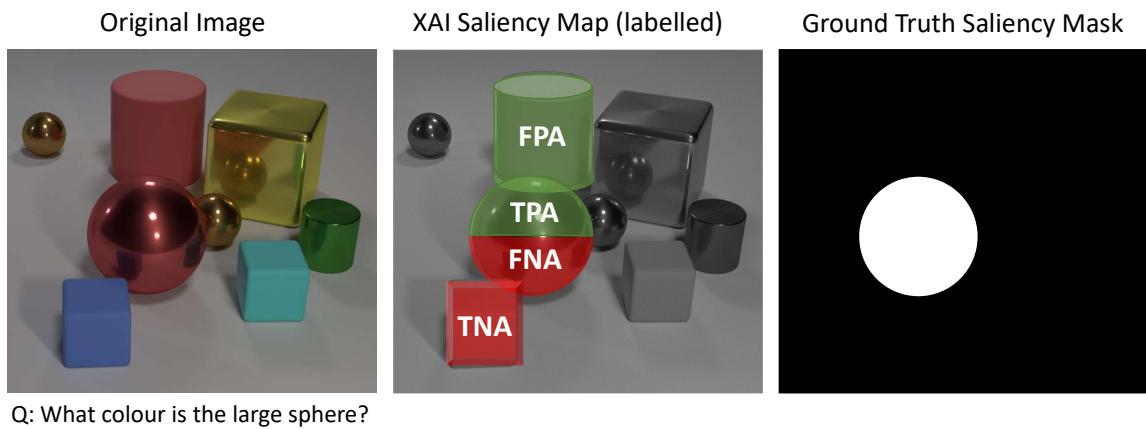


Figure 1: *Overview of our method. Areas of positive salience are green, while areas of negative saliency are red. We treat the problem of XAI evaluation as a classification problem. We define True Positive Area (TPA), True Negative Area (TNA), False Negative Area (FNA) and False Positive Area (FPA) based on the XAI saliency map and the ground truth. Can then evaluate XAI methods performance using Relevance F-score (RFS), which is equivalent to F1-score.*

- **True Negative Area (TNA)** - Regions with negative saliency located outside the ground truth mask of the object of interest. These pixels correspond to the irrelevant area(s) that the machine learning model should not use to make correct predictions.
- **False Negative Area (FNA)** - Regions with negative saliency located inside the ground truth mask of the object of interest. These regions are falsely identified as irrelevant.
- **False Positive Area (FPA)** - Regions with a positive saliency value located outside the ground truth mask of the object of interest. These regions are falsely identified as relevant.

In an ideal scenario, we expect all positive attributions to be located on the relevant object identified in the ground truth. And we expect that pixels located outside the ground truth are not required to answer the question, so should have negative or zero saliency. Current approaches to evaluating XAI saliency maps, such as Relevance Mass Accuracy (RMA) and Relevance Rank Accuracy (RRA), tend to ignore the negative saliency values [Arras et al., 2020]. This is because they have a pooling function, such as squaring all saliency values, to make all saliency values positive, which allows easy comparison with the ground-truth mask. RMA is defined as  $RMA = R_{\text{within}} / R_{\text{total}}$  i.e., the sum of saliency scores within the ground truth mask, divided by the sum of saliency values in the whole image. Using our definitions above, and ignoring negative attributions, we can see this is equivalent to true-positive pixels divided by the sum of true positives and false positives, i.e. the standard definition of precision. Precision alone cannot fully characterise the performance of a classifier. We address this limitation by introducing the Relevance F-score (RFS).

### 3.1 Relevance F-Score (RFS)

To address the limitations of existing XAI evaluation methods, we propose to evaluate the localisation performance of XAI methods using a measure based on the F1-score, which we call Relevance F-Score (RFS). A good classifier should have both high precision and high recall. Likewise, a good XAI method with a high RFS, has two desirable properties that correlate with what a human evaluator would expect from a good explanation. Firstly, the model makes a prediction based on the relevant pixels only, i.e., the positive attributions overlap with the object of interest or the ground truth mask. Secondly, negative attributions occur outside the ground

truth mask. We define Relevance F-Score (RFS) as follows:

$$\text{Relevance F-Score} = 2 \frac{P_a R_a}{P_a + R_a} \quad (1)$$

$$P_a = TPA / (TPA + FPA) \quad (2)$$

$$R_a = TPA / (TPA + FNA) \quad (3)$$

Where True Positive Area (TPA), False positive Area (FPA) and False Negative Area (FNA) are defined above (See Section 3 and Fig. 1). We can define  $P_a$  and  $R_a$  equivalently to precision and recall. Finally, the Relevance F-Score is defined in an equivalent way to F1-score. From this definition, we can see that for RFS to take a high value, both  $P_a$  and  $R_a$  must take high values. For this to happen the positive area of the saliency map must fall solely within the ground-truth mask of the object of interest, while any negative attributions must fall outside the object of interest.

## 4 Experiments

In this section, we compare Relevance F-Score (RFS) against the literature. To do this, we evaluate various XAI algorithms on a visual question-answering (VQA) network. We use a Relation Network [Santoro et al., 2017], trained and evaluated on the CLEVR-XAI dataset [Arras et al., 2020], as our VQA model. The final layer of the Relation Network is a softmax with 28 outputs, covering all the possible answers to the CLEVR questions. The model achieves 98% and 90% accuracy on the CLEVR simple and complex questions, respectively. We generate saliency maps for each XAI method using the Captum [Kokhlikyan et al., 2020] XAI framework.

The CLEVR XAI dataset [Arras et al., 2020] consists of 10,000 synthetic images of various geometric objects. It has two sets of queries: simple and complex. The dataset includes questions, answers, and ground truth masks for all questions. CLEVR XAI simple has 39,761 questions and focuses on object attributes related to a single object, such as material, size, shape and colour. It does not include inter-object relational questions, for example, a position of an object relative to another. CLEVR XAI complex includes 100,000 questions related to quantitative, relational, and existential properties of the scene.

### 4.1 Quantitative Comparison of RFS with the literature

We compare our proposed method, RFS with two existing XAI evaluation methods from the literature Relevance Mass Accuracy and Relevance Rank Accuracy [Arras et al., 2020]. Given a VQA network and the CLEVR dataset; for each prediction in the CLEVR simple and complex sets, we applied six different post-hoc XAI methods: Guided Backprop [Springenberg et al., 2014], Deconvnet [Zeiler and Fergus, 2014], GradientxInput [Shrikumar et al., 2016], IG [Sundararajan et al., 2017], Gradient [Simonyan et al., 2013]. The resulting explanations are represented as a  $128 \times 128 \times 3$  pixel matrix, due to the RGB channels of the original images. To allow comparison with the one-channel ground truth we sum across the RGB channels. This allows negative values to be present in the saliency maps, which are needed for RFS. For Relevance Mass Accuracy (RMA), which requires positive saliency values, we use min-max normalisation.

#### 4.1.1 Simple Protocol

The simple protocol is intended to measure a VQA model's ability to answer questions related to the attributes of a single object e.g., shape, colour, size or material. To answer these questions, the model only needs to focus on a single object in the scene. We perform several variations of this experiment to measure different aspects of performance. Experiment 1 calculates the localisation metrics with ground truth containing only a single object and when the answer is correctly predicted. This experiment serves as a baseline as we compare the subsequent experiments with the first one. In experiment 2, we report results only for cases where the underlying classifier has high confidence. In experiment 3, we measure the location of the attributions for cases where the object of interest has an area greater than 1,000 pixels. In experiment 4, we use the ground truth

Experiment	RFS mean	RMA (Ours) mean	RRA (Ours) mean	RMA (Original) [Arras et al., 2020] mean	RRA (Original) [Arras et al., 2020] mean
1) Single object, correct answers	0.112	0.070	0.095	0.478	0.384
2) High confidence ( $>0.99999$ )	0.129	0.082	0.108	0.502	0.400
3) GT mask $> 1,000$ pixels	0.112	0.070	0.094	0.714	0.544
4) All objects, correct answers	0.392	0.356	0.282	0.640	0.424
5) Colour only	0.075	0.105	0.120	0.498	0.412

Table 1: Mean RFS, RMA and RRA averaged over all six XAI methods for simple protocol experiments. (Ours) indicates our own implementation of RMA and RRA [Arras et al., 2020].

mask for all objects and include only correctly predicted answers. Finally, in experiment 5, we consider only correctly predicted questions and query an object’s colour. For each experiment, we apply the six XAI methods to explain the predictions of the VQA model. We then measure the similarity between the ground-truth mask and the XAI saliency map produced by each XAI method using RFS and two methods from the literature, RMA and RRA [Arras et al., 2020]. We then average the results across all the XAI methods. In Table 1, we show the results of all our experiments. We also include the results from the same experiments on the original CLEVR-XAI paper [Arras et al., 2020].

The results in Table 1 show that the RFS scores are generally lower in magnitude than the original RMA and RRA scores. This is because RFS considers negative attributions, leading to fewer positive attribution pixels within the ground-truth mask. (We will confirm this by qualitative examination in Section 4.2). Compared to the original RMA and RRA figures, our RMA and RRA figures are produced using min-max normalisation. So we observe that our average RMA and RRA scores are significantly lower than those in the original CLEVR XAI paper and lower than our RFS. We treat Experiment 1 as a baseline. The results of Experiment 2 show that when looking at only correct and high-confidence predictions, the RFS scores are generally higher than the baseline Experiment 1. This indicates that the underlying network emphasises the correct object of interest more. We note that the same is not observed when looking at RMA. With RMA, the average change compared to the baseline experiment is only 5%, whereas, for RFS, the difference is 15%. In Experiment 3, which looks at larger objects, we can see little change in the RFS from the baseline. However, we can see that both RMA and RRA are more sensitive to changes in object size. In the original CLEVR XAI paper RMA and RRA increase by over 40% on average (49% and 42% respectively) relative to the baseline, whereas our mean RMA and RRA scores change significantly less (0.1% and 1.5% respectively). This highlights that our approach is more stable. Looking at the results for Experiment 4, where all objects are used in the ground truth, RFS and our RMA and RRA scores, increase significantly compared to using a single object ground truth, which means that the Relation Network uses several objects to make predictions. On the other hand, the original RMA and RRA scores increase to a smaller extent. In Experiment 5, the mean Relevance F-score, RMA, and RRA all increase. We note that our results differ from those in the CLEVR XAI paper.

#### 4.1.2 Complex Protocol

The complex protocol measures how accurately the model can answer questions related to counting, or spatial relationships between multiple objects. The queries in this experiment are related to multiple objects, so a variety of ground truth masks are used:

- **Unique:** This mask shows the location of a single object that must be uniquely identified among the set of objects present in the scene to answer the question.
- **Unique first non-empty:** This mask identifies one or more objects that are necessary to answer the question correctly. It includes the objects in the question.

<b>XAI Method</b>	Unique		Unique First-n-e.		Union		All Objects	
	<b>RFS</b>	<b>RMA</b>	<b>RFS</b>	<b>RMA</b>	<b>RFS</b>	<b>RMA</b>	<b>RFS</b>	<b>RMA</b>
Guided BProp [Springenberg et al., 2014]	0.23	0.17	0.23	0.17	0.38	0.28	0.46	0.34
Deconvnet [Zeiler and Fergus, 2014]	0.23	0.17	0.23	0.17	0.38	0.28	0.46	0.34
Gradient×Input [Shrikumar et al., 2016]	0.20	0.15	0.16	0.11	0.34	0.25	0.42	0.30
IG [Sundararajan et al., 2017]	0.12	0.08	0.15	0.10	0.25	0.12	0.22	0.14
Gradient [Simonyan et al., 2013]	0.11	0.07	0.14	0.10	0.20	0.13	0.25	0.17

Table 2: Mean RFS and RMA measured across all XAI methods on the CLEVR-XAI-complex dataset. Averaged across all correctly predicted questions that do not involve counting. We discard questions about an object’s existence where the answer is no.

- **Union:** This mask shows the location of all objects that are related to the question.
- **All objects:** This mask includes all objects in the image, regardless of their relevance to the question.

In Table 2, we show the mean RFS score averaged across the correct predictions where the question is not quantitative and where we exclude questions where the answer is ‘no’. The results suggest that XAI methods that use back-propagation, for example, DeConvNet [Zeiler and Fergus, 2014] and Guided Back-propagation [Springenberg et al., 2014], achieve higher RMA, which implies their positive attributions are more concentrated inside the ground truth mask than in gradient-based methods. However, these methods also achieve higher RFS, because backpropagation-based methods have a tendency to generate less noise on the attribution map. In other words, these explanation methods can provide a more accurate representation of positive and negative attributions. On the other hand, simple gradient-based methods that calculate the gradients via a forward pass, including Gradient, Input × Gradient, and Integrated Gradients appear to produce more attributions (Figure 3). As a result, these explanations appear to include significant noise meaning the attributions are visible even if their value is negligible. As a result, they achieve lower RFS.

## 4.2 Qualitative Evaluation

We now perform a qualitative evaluation to compare RFS with the existing RMA method. This experiment examines explanation saliency maps with different combinations of recall and precision, hence showing qualitatively how RFS varies as the underlying saliency map varies. For each condition, we report the corresponding RFS and RMA values. The saliency maps were produced using Integrated Gradients.

The results in Fig. 2 demonstrate that our proposed method RFS is better able to indicate when the XAI method has correctly localised the relevant object(s). In contrast, we see that when the RMA is high, it does not necessarily indicate that the model is looking at the correct object or that the attributions are concentrated. In addition, the shortcoming of RMA in our example is that it turns negative attributions into positives. However, the RFS uses the location of both positive and negative attributions, providing a better understanding of the underlying functionality of different explainable AI methods. In Fig. 2 we find cases where the RMA is high, but recall is low. This illustrates that RMA is an insufficient evaluation metric whereas RFA requires both precision and recall to be high to achieve a high value. Hence RFA is better at identifying explanations that focus on the object of interest, and hence have a high localisation accuracy, as in the lower right of Fig. 2.

## 5 Conclusion

In this paper, we have proposed a new metric, Relevance F-Score, to measure the localisation quality of saliency-map-based XAI methods. Relevance F-Score considers both positive and negative saliency regions to give a more rounded understanding of the localisation performance of different XAI methods. We have evaluated our methods both quantitatively and qualitatively on the CLEVR-XAI dataset. We show that our method is better able to identify which XAI methods have good localisation performance compared to existing

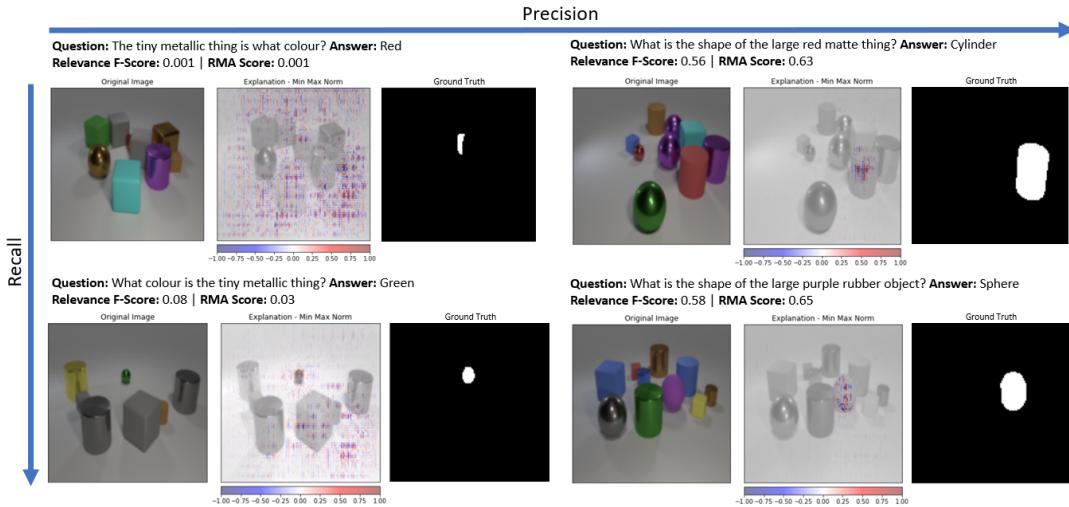


Figure 2: The above examples demonstrate that RFS with min-max normalisation can show the location of negative attributions, leading to a more robust explanation than RMA alone.

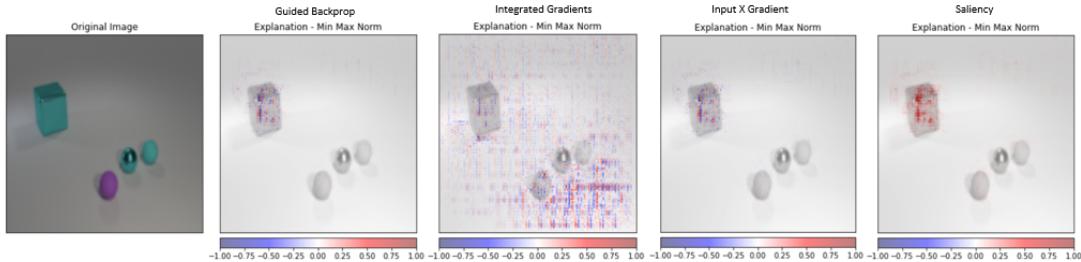


Figure 3: This figure illustrates the wide variety of outputs from different XAI methods. From left to right - Guided Backprop, Integrated Gradients, Input  $\times$  Gradient and Saliency.

XAI evaluation methods. Our evaluation is made possible by the ground-truth information on the CLEVR-XAI dataset. In future work, we will seek to extend the evaluation of XAI methods to more complex real-world datasets.

## References

- [Arias-Duart et al., 2021] Arias-Duart, A., Parés, F., and Garcia-Gasulla, D. (2021). Who explains the explanation? quantitatively assessing feature attribution methods. *CoRR*, abs/2109.15035.
- [Arras et al., 2020] Arras, L., Osman, A., and Samek, W. (2020). Ground truth evaluation of neural network explanations with clevr-xai. *arXiv preprint arXiv:2003.07258*.
- [Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.
- [Huang et al., 2020] Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., and Chang, Y. (2020). Graphlime: Local interpretable model explanations for graph neural networks. *CoRR*, abs/2001.06216.
- [Kohlbrenner et al., 2020] Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., and Lapuschkin, S. (2020). Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

- [Kokhlikyan et al., 2020] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al. (2020). Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.
- [Krishna et al., 2022] Krishna, S., Han, T., Gu, A., Pomba, J., Jabbari, S., Wu, S., and Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.
- [Reio Jr, 2016] Reio Jr, T. G. (2016). Nonexperimental research: Strengths, weaknesses and issues of precision. *European Journal of Training and Development*, 40(8/9):676–690.
- [Santoro et al., 2017] Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P. W., and Lillicrap, T. P. (2017). A simple neural network module for relational reasoning. *CoRR*, abs/1706.01427.
- [Shrikumar et al., 2016] Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.
- [Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- [Springenberg et al., 2014] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.
- [Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- [Theiner et al., 2022] Theiner, J., Müller-Budack, E., and Ewerth, R. (2022). Interpretable semantic photo geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 750–760.
- [Wang et al., 2020] Wang, J., Wiens, J., and Lundberg, S. M. (2020). Shapley flow: A graph-based approach to interpreting model predictions. *CoRR*, abs/2010.14592.
- [Yuan et al., 2020] Yuan, H., Tang, J., Hu, X., and Ji, S. (2020). Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 430–438, New York, NY, USA. Association for Computing Machinery.
- [Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.
- [Zhang et al., 2018] Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102.
- [Zhou et al., 2021] Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. (2021). Do feature attribution methods correctly attribute features? *CoRR*, abs/2104.14403.

# Sampling Matters in Explanations: Towards Trustworthy Attribution Analysis Building Block in Visual Models through Maximizing Explanation Certainty

Jiaolin Luo (Róisín)<sup>\*,†</sup>, James McDermott<sup>\*,†</sup>, and Colm O’Riordan<sup>\*,†</sup>

<sup>\*</sup>*University of Galway (Ireland)*

<sup>†</sup>*SFI Centre for Research Training in Artificial Intelligence (Ireland)*

{j.luo2,james.mcdermott,colm.oriordan}@universityofgalway.ie

## Abstract

Image attribution analysis seeks to highlight the feature representations learned by visual models such that the highlighted feature maps can reflect the pixel-wise importance of inputs. Gradient integration is a building block in the attribution analysis by integrating the gradients from multiple derived samples to highlight the semantic features relevant to inferences. Such a building block often combines with other information from visual models such as activation or attention maps to form ultimate explanations. Yet, our theoretical analysis demonstrates that the extent to the alignment of the sample distribution in gradient integration with respect to natural image distribution gives a lower bound of explanation certainty. Prior works add noise into images as samples and the noise distributions can lead to low explanation certainty. Counter-intuitively, our experiment shows that extra information can saturate neural networks. To this end, building trustworthy attribution analysis needs to settle the sample distribution misalignment problem. Instead of adding extra information into input images, we present a semi-optimal sampling approach by suppressing features from inputs. The sample distribution by suppressing features is approximately identical to the distribution of natural images. Our extensive quantitative evaluation on large scale dataset ImageNet affirms that our approach is effective and able to yield more satisfactory explanations against state-of-the-art baselines throughout all experimental models.

**Keywords:** Explainability and Interpretability, Trustworthy Computer Vision, Image Attribution Analysis

## 1 Introduction

Image semantic features are conveyed by the joint distribution of image pixels. Visual models learn how to extract the features in training process. Pixel-wise attribution analysis seeks to highlight the learned feature representations relevant to inferences as explanations (Linardatos et al., 2020; Adadi and Berrada, 2018). Gradient integration is a major building block in attribution analysis by integrating the gradients from multiple samples sampled from some distributions (Simonyan et al., 2013; Baehrens et al., 2010; Smilkov et al., 2017; Omeiza et al., 2019; Sundararajan et al., 2017; Jalwana et al., 2021).

The first-order gradients merely provide limited local information regarding the inferences of models at a given point. Moreover, the extra information from samples can suppress the activation of networks and causes saturation (See Figure 4(b)). Furthermore, our experiment suggests that some gradient integration based algorithms are susceptible to the training with data augmentation. For example, fine-tuning models with data augmentation can fail the explanations (See Figure 2). This is because the models training with data augmentation learn to ignore the samples from the sampling distributions not aligned with natural image distribution.

---

Paper reproducibility: [https://anonymous.4open.science/r/sampling\\_matters\\_reproducibility-BB60/](https://anonymous.4open.science/r/sampling_matters_reproducibility-BB60/).

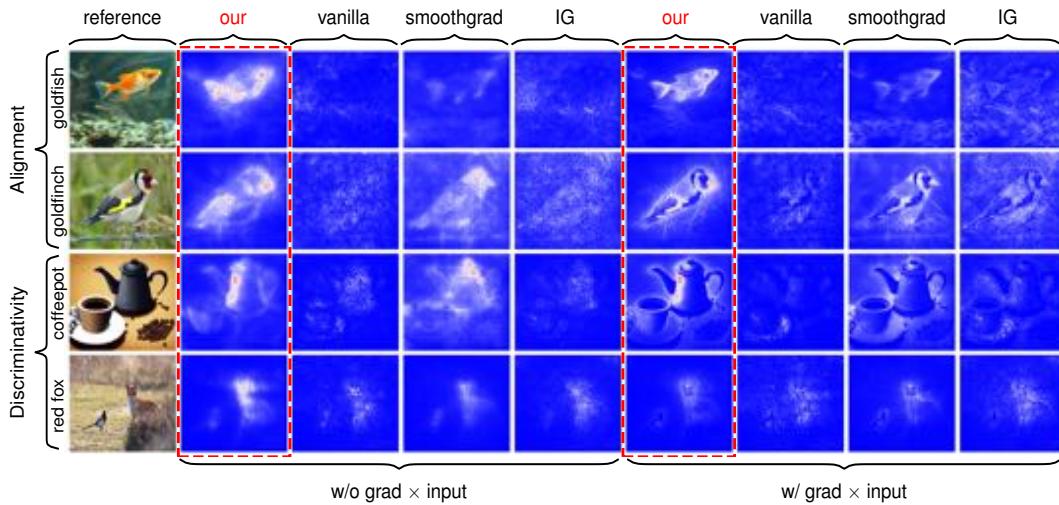


Figure 1: The figure is a performance showcase. The performances are qualitatively compared from two aspects: (1) The semantic alignment and (2) the object discriminativity with the presence of multiple objects. The evaluation model is a ResNet50 pre-trained on ImageNet. The iteration numbers are set to 50. The pixel dropping probability of our algorithm is set to 0.7. The noise level of Smoothgrad is set to 0.15 as suggested in the original paper. The explanations are normalized to [0, 1] by using min-max normalization and visualized with the quantitative color map “bwr” ([blue = 0, white = 0.5, red = 1]).

In recent works, e.g. Grad-CAM (Selvaraju et al., 2017) and CAMERAS (Jalwana et al., 2021), gradient integration often serves as a building block by combining with activation or attention maps to provide more robust explanations.

Yet, our theoretical analysis suggests that such a building block in attribution analysis often suffers from low explanation certainty due to the misalignment of the sample distribution with respect to natural image distribution. Building trustworthy visual models is an imperative call for crucial scenarios such as medical imaging and autonomous driving. To this end, we must settle the sampling misalignment problem in gradient integration and seek more trustworthy sampling approach towards trustworthy visual models.

We theoretically revisit the sampling problem from the perspective of explanation certainty and propose a semi-optimal sampling approach by suppressing features from input images. The intuition behind is that natural images densely encode information with considerable redundancy and a small fraction of pixels can still convey considerable feature information to make reasonable inferences. For example, humans can still recognize the objects in images by dropping 80% of pixels (See Figure 3).

## 2 Related work

We focus on the discussion of the sampling problem in the building block gradient integration. Therefore, we conduct a brief literature review to provide the context for this research in visual models from two aspects: (1) Local explainability and (2) gradient integration.

**Local explainability:** Local explainability in visual models provides the explanations for individual images in inferences. Recent methods fall into the follow categories:

- Local approximation based methods: e.g. LIME (Local Interpretable Model-Agnostic Explanations) (Mishra et al., 2017);
- Shapley value theory based methods: e.g. SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017);
- Gradient integration based methods: e.g. Smoothgrad (Smilkov et al., 2017) and IG (Sundararajan et al., 2017);

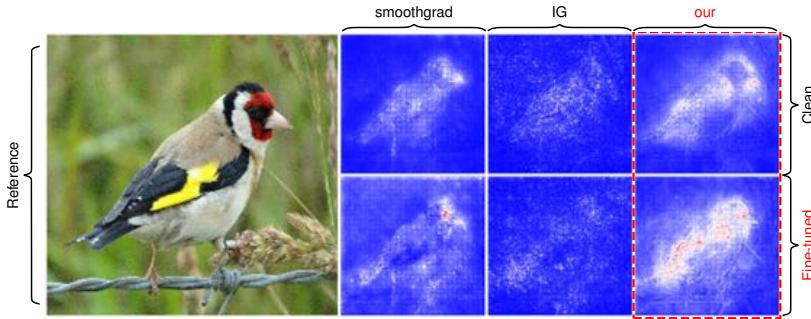


Figure 2: This experiment exhibits the failures and the sensitivity of the explanations by fine-tuning a pre-trained ResNet50 on ImageNet with data augmentation. The model is fine-tuned with two data augmentations: (1) Randomly adding Gaussian with  $\sigma$  in  $[0.1, 0.3]$  and (2) randomly adjusting the luminance within  $[0.1, 0.9]$ . We train 5000 batches with batch-size by 8, learning rate  $10^{-3}$  and SGD optimizer. The quality of explanations degrades due to the model learns to ignore the irrelevant perturbations in the samples by adding noise.

- Activation decomposition based methods: e.g. DeepLIFT (Deep Learning Important FeaTures) (Shrikumar et al., 2017), LRP (Layer-wise Relevance Propagation) (Montavon et al., 2019), and CAM (Class Activation Map) (Li et al., 2018);
- Combination of gradient and activation maps: e.g. Grad-CAM (Selvaraju et al., 2017) and CAMERAS (Jalwana et al., 2021).

**Gradient integration:** Gradient maps capture the local semantic features from images relevant to inferences. Integrating multiple gradients from multiple samples can improve explanation quality. Smilkov et al. add noise from some normal distributions into images to form samples (Smilkov et al., 2017). The gradients from the derived samples are then integrated to form explanations. Yet, the samples by adding random noise do not align with the distribution of natural images. Such a misalignment can yield the explanations with low explanation certainty (See the fourth and the eighth columns in Figure 1). Our experiment shows that their approach can fail if visual models are trained with data augmentation as models learn to ignore the added noise (See Figure 2). Inspired by the Aumann–Shapley theory (Shapley, 1953; Roth, 1988; Aumann and Shapley, 2015), Sundararajan et al. attempt to tackle the neuron saturating problem by globally scaling inputs to create multiple samples from inputs. The derived multiple gradient maps from the samples are then integrated to create ultimate explanations (Sundararajan et al., 2017). Such an approach can preserve the features aligned with natural images. However, their samples suffer from the lack of feature diversity. Thus, the explanations using IG remain unsatisfactory (See the fifth and the ninth columns in Figure 1). To overcome the limit that gradient maps can merely capture local semantic information, Grad-CAM (Selvaraju et al., 2017) and CAMERAS (Jalwana et al., 2021) combine gradient maps with activation maps to form the explanations with higher quality.

### 3 Sampling problem

We formulate the explanation certainty and theoretically analyze the distribution alignment problem in gradient integration.

#### 3.1 Explanation certainty

The explanation task for visual models asks a question for given model  $\mathcal{M}$ , given input image  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$  (where  $x_i$  denotes the  $i$ -th pixel and  $d$  denotes the input dimension), given ground-truth category  $y \in \mathbb{R}$  and given explanation  $\mathbf{z} \in \mathbb{R}^d$ : To what extent, by showing the explanation  $\mathbf{z}$ , we can infer that the explanation is more relevant to the input  $\mathbf{x}$ . If the explanation  $\mathbf{z}$  and the input  $\mathbf{x}$  are not relevant for given model  $\mathcal{M}$ , we can infer that the model  $\mathcal{M}$  does not align with human-specific intuitions and lack of explainability. An explanation



Figure 3: This example shows the information redundancy in nature images. The leftmost is the reference. The images from the rightmost to leftmost aside the reference are with random pixel sampling ratio by 0.7, 0.6, 0.5, 0.4, 0.3 and 0.2 respectively.

algorithm is map:  $g: (\mathbf{x}, \mathcal{M}, y) \mapsto \mathbf{z}$  which is expected to faithfully reflect the relevance between inputs and explanations. We define ‘explanation certainty’ as the conditional probability  $Pr(\mathbf{x}|\mathbf{z}; \mathcal{M}, y)$ : The probability of the input  $\mathbf{x}$  for given explanation  $\mathbf{z}$ , given model  $\mathcal{M}$  and given ground-truth category  $y$ . We simplify and denote the explanation certainty as  $Pr(\mathbf{x}|\mathbf{z})$  to keep depiction succinct thereafter.

### 3.2 A lower bound of explanation certainty

It is not difficult to show that the mutual information (Kullback, 1997; Cover, 1999) between inputs and its explanations is a lower bound of  $Pr(\mathbf{x}|\mathbf{z})$ . Let further assume natural images are pixel-wise i.i.d. to simplify the theoretical analysis. Let  $p(x|z)$  be the conditional p.d.f. for some pixel  $x \in \mathbf{x}$  at given some pixel  $z \in \mathbf{z}$ . Hence the  $Pr(\mathbf{x}|\mathbf{z})$  can be approximately rewritten as:

$$Pr(\mathbf{x}|\mathbf{z}) = \prod_x \prod_z Pr(x|z) = \exp\left(\sum_x \sum_z \log Pr(x|z)\right) \approx \exp\left(\iint_{z,x} \log p(x|z) dx dz\right). \quad (1)$$

Considering:

$$\iint_{z,x} p(x, z) dx dz = 1 \quad (2)$$

and the mutual information between  $\mathbf{x}$  and  $\mathbf{z}$ :

$$I(\mathbf{x}; \mathbf{z}) = \iint_{z,x} p(x, z) \log\left(\frac{p(x|z)}{p(x)}\right) dx dz \quad (3)$$

and the entropy of  $\mathbf{x}$ :

$$H(\mathbf{x}) = - \int_x p(x) \log p(x) dx. \quad (4)$$

Combining above results in equations (1, 2, 3 and 4) and applying Hölder’s inequality:

$$\begin{aligned} \iint_{z,x} \log p(x|z) dx dz &= 1 \cdot \iint_{z,x} \log p(x|z) dx dz = \iint_{z,x} p(x, z) dx dz \cdot \iint_{z,x} \log p(x|z) dx dz \\ &\geq \iint_{z,x} p(x, z) \log\left(\frac{p(x|z)}{p(x)}\right) p(x) dx dz \\ &= \iint_{z,x} p(x, z) \log\left(\frac{p(x|z)}{p(x)}\right) dx dz + \int_x (\int_z p(x, z) dz) \log p(x) dx \\ &= I(\mathbf{x}; \mathbf{z}) + \int_x p(x) \log p(x) dx \\ &= I(\mathbf{x}; \mathbf{z}) - H(\mathbf{x}) \equiv -H(\mathbf{x}| \mathbf{z}). \end{aligned} \quad (5)$$

Hence mutual information gives a lower bound of explanation certainty by:

$$Pr(\mathbf{x}|\mathbf{z}) \geq \frac{\exp(I(\mathbf{z}; \mathbf{x}))}{\exp(H(\mathbf{x}))} \equiv \exp(-H(\mathbf{x}| \mathbf{z})). \quad \square \quad (6)$$

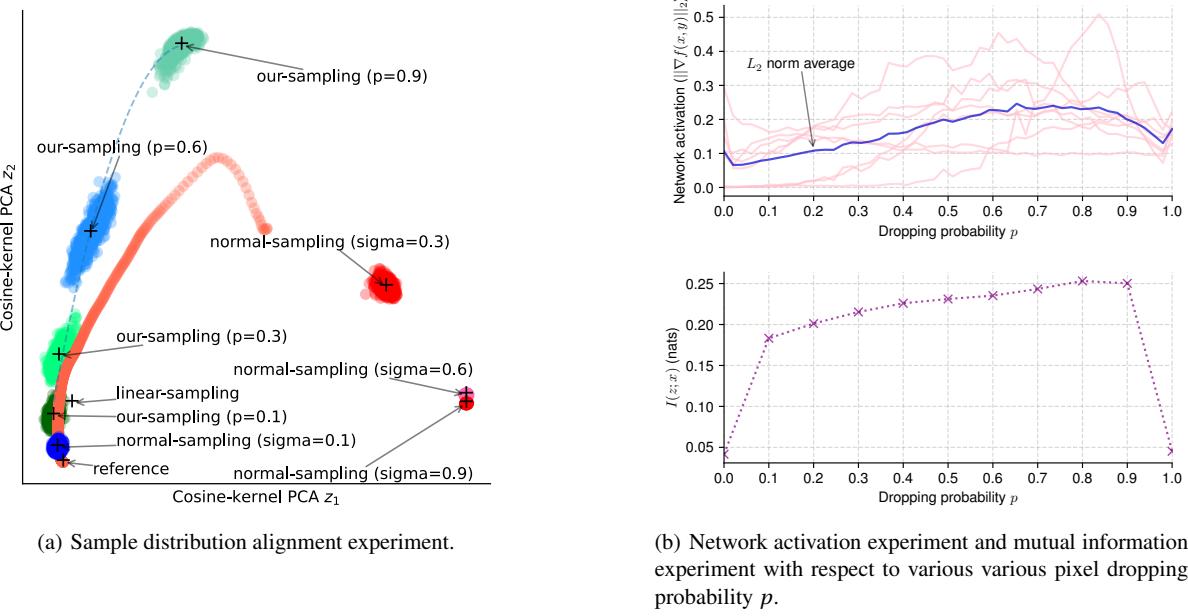


Figure 4: In the left figure we investigate how neural networks respond to various sampling distributions. We use a ResNet50 pre-trained on ImageNet and project the embeddings from the penultimate layer with cosine-kernel PCA. We collect 1000 samples for each sampling approach. We vary the noise level of normal-sampling (used in Smoothgrad) from 0.1 to 0.9. We vary the pixel dropping probability of our-sampling approach from 0.1 to 0.9. We use “+” to indicate the projection centers. In the right figure, we empirically investigate how neural networks activate by measuring the average gradient norms with respect to various pixel dropping probability  $p$  in our approach. We also conduct the experiment to measure how the dropping probability  $p$  affects explanation certainty (mutual information). The result shows that the optimal dropping probability  $p$  is from 0.6 to 0.8.

### 3.3 Gradient integration

The gradient integration without the technique of ‘inputs  $\times$  gradients’ (Shrikumar et al., 2017; Sundararajan et al., 2017) can be formulated as:

$$z := \mathbb{E}_{\hat{x} \sim q(x)} [|\nabla_{\hat{x}} f(\hat{x}, y)|] \quad (7)$$

where  $|\cdot|$  denote point-wise absolute value modulus,  $x$  denotes some input,  $z$  denotes some explanation,  $\hat{x}$  denotes some sample from some sampling distribution  $q(x)$ ,  $y$  denotes ground-truth label, and  $f : (x, y) \mapsto \mathbb{R}$  defines some image classifier.

### 3.4 Optimal sampling

Explanation certainty is lower bounded by the mutual information between inputs and explanations for given models. Maximizing the explanation certainty in gradient integration based approach can improve the trustworthiness and faithfulness of explanations. Let  $q_x$  be some sampling distribution for given input  $x$  in gradient integration. Let  $p$  be natural image distribution. Let  $q^*$  be the optimal sampling distribution. We have:

$$q^* = \arg \max_q \mathbb{E}_{x \sim p} [I(x, z)] = \arg \max_q \mathbb{E}_{x \sim p} \left[ I(x, \mathbb{E}_{\hat{x} \sim q_x} [|\nabla_{\hat{x}} f(\hat{x}, y)|]) \right] \quad (8)$$

Considering that mutual information function  $I(x, z)$  is convex for given  $x$  and applying Jensen’s inequality:

$$I(x, z) = I(x; \mathbb{E}_{\hat{x} \sim q_x} [|\nabla_{\hat{x}} f(\hat{x}, y)|]) \leq \mathbb{E}_{\hat{x} \sim q_x} [I(x; |\nabla_{\hat{x}} f(\hat{x}, y)|)] = \mathbb{E}_{\hat{x} \sim q_x} [I(x; \hat{z})]. \quad (9)$$

By using equation (9), we now rewrite the equation (8) into:

$$q^* = \arg \max_q \mathbb{E}_{\mathbf{x} \sim p_x} \left[ \mathbb{E}_{\hat{\mathbf{x}} \sim q_x} [I(\mathbf{x}; \hat{\mathbf{z}}(\hat{\mathbf{x}}))] \right] \quad (10)$$

which has the optimal solution  $q^* = p$ . This result implies that if samples come from the same distribution as natural images. The mutual information in gradient integration between inputs and explanations will be maximized.

## 4 Sampling with feature suppression

Natural images encode information densely with considerable redundancy (Figure 3). Humans can still give reasonable inferences for images when a majority of pixels is dropped. Unlike that prior works add extra noise into images to create samples in gradient integration, we instead remove information from images. Interestingly, in our experiment, we have also observed that less information by dropping a proper proportion of pixels can further increase the activation of networks and maximize the explanation certainty. Figure 4(b) provides the experiments to show such an observation.

### 4.1 Feature suppression

For each image  $\mathbf{x} = (x_i)_{i=1}^d$  with  $d$  pixels, we sample an image  $\mathbf{x}^*$  from  $\mathbf{x}$  by doing pixel-wise Bernoulli trial for the  $i$ -th pixel from set  $\{x_i, 0\}$  with dropping probability  $p$ . Thus  $x_i^* = \mathbb{B}(\{x_i, 0\}; 1 - p)$  where  $\mathbb{B}$  gives a Bernoulli trial with probability  $1 - p$ . Our approach is formulated as:

$$\mathbf{z} := \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbb{B}(\{\mathbf{x}, 0\}; 1 - p)} [|\nabla_{\hat{\mathbf{x}}} f(\hat{\mathbf{x}}, y)|] \quad (11)$$

where  $\mathbb{B}(\{\mathbf{x}, 0\}; 1 - p)$  denotes our sampling operation,  $\hat{\mathbf{x}}$  denotes a sample,  $\mathbf{z}$  denotes the explanation,  $y$  denotes some ground-truth label, and  $f : (\hat{\mathbf{x}}, y)$  denotes some model.

### 4.2 Sampling alignment experiment

The experiment in Figure 4(a) shows how neural networks respond to samples by taking the embeddings from the penultimate layer of a ResNet50 pre-trained on ImageNet and projecting the embeddings with cosine-kernel PCA. The ‘linear-sampling’ refers to simply scale the pixel intensities of images in global. The distribution of the ‘linear-sampling’ aligns with the distribution of natural images. The ‘normal-sampling’ refers to the samples derived by adding the noise from various normal distributions. The result shows that our samples (marked as ‘our-sampling’) align with the natural images. The samples from ‘normal-sampling’ do not align well with natural images. The projections show that models treat the samples from various distributions with different patterns.

#### 4.2.1 Optimal $p$

We empirically investigate the optimal dropping probability  $p$  in Figure 4(b) by: (1) Measuring how networks activate with respect to various  $p$  and (2) measuring how explanation certainty (mutual information) changes with respect to various  $p$ . We vary  $p$  from 0 to 1 with stride 0.1. All results are measured on a ResNet50 pre-trained on ImageNet. The result shows that when  $p$  falls in  $[0.6, 0.8]$  the network activation will be maximized and the mutual information will be maximized as well. We also repeat this experiment on other architectures. The pattern remains the same. There are two findings which are counter-intuitive: (1) When we suppress information from inputs, the activations of networks first increase and networks extract more relevant features, and, (2) as we continue to suppress more information, networks become more confusing and less activated and fail to extract relevant features. This experiment can further justify our approach by suppressing features from inputs. The behind mechanism still needs further investigation.

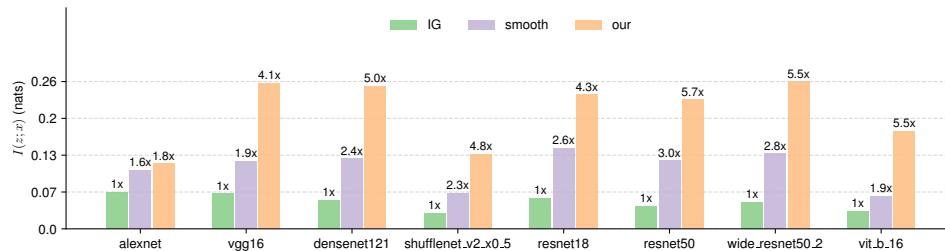


Figure 5: We extensively evaluate our algorithm against the gradient integration baselines over ImageNet dataset. All random seeds are set to 1860867 to guarantee the reproducibility. We choose the models pre-trained on ImageNet. We use 100 epochs and 100 images for all methods. For all chosen models, our approach outperforms state-of-the-art baselines.

## 5 Evaluation

As we focus on investigating the sampling problem in the building block for attribution analysis, we set three baselines: Smoothgrad, IG and vanilla due the research scope. We perform both qualitative and quantitative evaluations.

### 5.1 Qualitative showcase

In Figure 1, we showcase the results with and without the ‘Gradient  $\times$  Input’ technique. We choose a ResNet50 pre-trained on ImageNet. The experiment examines the performance from two aspects: (1) Semantic feature alignments and (2) the object discriminativity with the presence of multiple objects. The result shows that our approach can yield more satisfactory explanations intuitively.

### 5.2 Quantitative evaluation

We also conduct an extensive quantitative evaluation by measuring the lower bound (mutual information) of explanation certainty. We choose multiple models pre-trained on ImageNet. The mutual information unit is in *nats* and the results are also normalized with respect to the IG on the model basis. The images are randomly sampled from ImageNet by fixing random seed for reproducibility purpose. The result shows that our approach outperform all baselines for all models.

## 6 Conclusions

We theoretically show that mutual information between explanations and inputs gives a lower bound of explanation certainty in the the explanation building block for image models with gradient integration approach. Maximizing explanation certainty can achieve trustworthy and faithful explanations. Due to the discussion scope, we have not unfolded: (1) The analysis in terms of the sensitivity to the training with data augmentation and (2) the results when our approach is incorporated with other information such as class activation or attention maps. A further investigation is necessary to address such concerns.

## Acknowledgments

This publication has emanated from research [conducted with the financial support of/supported in part by a grant from] Science Foundation Ireland under Grant number 18/CRT/6223. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We also thank reviewers for their constructive comments which can significantly improve our research quality. We thank the support from the ICHEC (Irish Centre for High-End Computing). We also thank the contributor Jiarong Li for the help in the discussions and proofreading.

## References

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.
- Aumann, R. J. and Shapley, L. S. (2015). *Values of non-atomic games*. Princeton University Press.
- Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831.
- Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.
- Jalwana, M. A., Akhtar, N., Bennamoun, M., and Mian, A. (2021). Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16327–16336.
- Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.
- Li, K., Wu, Z., Peng, K.-C., Ernst, J., and Fu, Y. (2018). Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Mishra, S., Sturm, B. L., and Dixon, S. (2017). Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, volume 53, pages 537–543.
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209.
- Omeiza, D., Speakman, S., Cintas, C., and Weldermariam, K. (2019). Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*.
- Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.
- Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100.
- Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
- Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

# Learning from Exemplary Explanations

Misgina Tsighe Hagos<sup>1,2</sup>, Kathleen M. Curran<sup>1,3</sup>, and Brian Mac Namee<sup>1,2</sup>

<sup>1</sup>*Science Foundation Ireland Centre for Research Training in Machine Learning*

<sup>2</sup>*School of Computer Science, University College Dublin*

<sup>3</sup>*School of Medicine, University College Dublin*

## Abstract

eXplanation Based Learning (XBL) is a form of Interactive Machine Learning (IML) that provides a model refining approach via user feedback collected on model explanations. Although the interactivity of XBL promotes model transparency, XBL requires a huge amount of user interaction and can become expensive as feedback is in the form of detailed annotation rather than simple category labelling which is more common in IML. This expense is exacerbated in high stakes domains such as medical image classification. To reduce the effort and expense of XBL we introduce a new approach that uses two input instances and their corresponding Gradient Weighted Class Activation Mapping (GradCAM) model explanations as exemplary explanations to implement XBL. Using a medical image classification task, we demonstrate that, using minimal human input, our approach produces improved explanations (+0.02, +3%) and achieves reduced classification performance (-0.04, -4%) when compared against a model trained without interactions.

**Keywords:** Explanation based Learning, Interactive Learning, Medical Image Classification.

## 1 Introduction

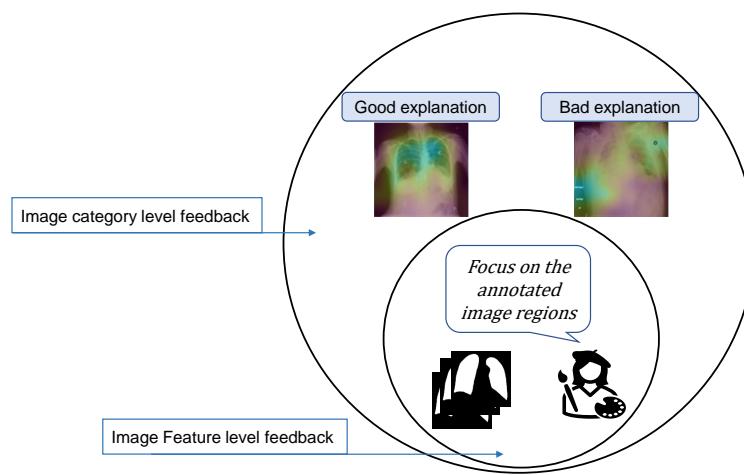


Figure 1: The inner circle shows the typical mode of feedback collection where users annotate image features. The outer circle shows how the Exemplary eXplanation Based Learning (eXBL) approach requires only identification of one good and one bad explanation.

Interactive Machine Learning (IML) is an approach that aims to provide a platform for user involvement in the model training or retraining process [Fails and Olsen Jr, 2003]. The literature on IML is dominated by active learning which reduces the manual effort associated with creating labelled training datasets by interactively selecting a sub-sample of an unlabelled dataset for manual labelling [Budd et al., 2021]. However, eXplanation Based Learning (XBL) has recently begun to gain traction as it allows deeper interaction with users by providing an opportunity to collect feedback on model explanations [Stumpf et al., 2009, Kulesza et al., 2015, Teso et al., 2023]. This form of interaction allows a more transparent form of model training than other IML approaches as users get a chance to refine a model by interacting-with and correcting its explanations.

XBL starts off with a learner model,  $f$ , that was initially trained using a simple classification loss, categorical cross entropy for example, which is calculated based on the error between the model's prediction and ground-truth label. Then, XBL typically refines  $f$  by augmenting its classification loss with an explanation loss,

$$L = L_{CE} + L_{expl} + \lambda \sum_{i=0} \theta_i \quad (1)$$

In Equation (1),  $L_{CE}$  is the traditional categorical cross entropy which is calculated based on the error between the model's predictions and ground-truth labels;  $L_{expl}$  is an explanation loss that is computed between the explanation produced from a model and a manual annotation of input instances,  $M$ ;  $\lambda$  is a regularisation term used to avoid overfitting that could be caused by the introduction of the new loss term,  $L_{expl}$ ; and  $\theta$  refers to network parameters.  $M$  can be a mask showing the important image regions that a learner should focus on or a mask of confounding or non-salient regions that a model should ignore. Saliency based feature attributions are usually used to generate model explanations. One example, from [Schramowski et al., 2020] formulates the explanation loss for training instances  $x \in X$  of size N and Gradient Weighted Class Activation Mapping (GradCAM) model explanations generated using a trained model  $f$  as shown in Equation (2). GradCAM is a saliency based local model explanation technique [Selvaraju et al., 2017].

$$L_{expl} = \sum_{i=0}^N M_i \text{GradCAM}(x_i) \quad (2)$$

As is seen in the inner circle of Figure 1, in XBL, the most common mode of user interaction is image feature annotation. This requires user engagement that is considerably much more demanding than the simple instance labelling that most IML techniques require [Zlateski et al., 2018] and increases the time and cost of feedback collection in XBL. As can be seen in the outer circle of Figure 1, we are interested in lifting this pressure from users (or feedback providers) and simplifying the interaction to ask for identification of two explanations as exemplary explanations and ranking them as good and bad explanations, and so make feedback collection cheaper and faster. This kind of user interaction where users are asked for a ranking instead of category labels has also been found to increase inter-rater reliability and data collection efficiency [O'Neill et al., 2017]. We incorporate this feedback into model training through a contrastive loss; specifically, triplet loss [Chechik et al., 2010].

The main goal of this paper is to demonstrate the effectiveness this loss based on just two exemplars. Therefore, we use an existing feature annotated dataset to identify good and bad explanations to demonstrate suitability of our proposal. In a real-world interactive learning scenario where end users have to choose the good and bad explanations, active learning approaches can be used to reduce the pool of explanations users have to choose the explanations from.

The main contributions of this paper are:

1. We propose the first type of eXplanation Based Learning (XBL) that can learn from only two exemplary explanations of two training images;
2. We adopt triplet loss for XBL to incorporate the two exemplary explanations into an explanation loss;

3. In addition to showing that XBL can be implemented with just two instances, our experiments demonstrate that our proposed method achieves improved explanations and comparable classification performance when compared against a baseline model.

## 2 Related Work

Based on the approach utilised to incorporate user feedback into model training, XBL methods can be generally categorised into two: (1) augmenting loss functions; and (2) augmenting training datasets using user feedback by removing confounding or spurious regions identified by users.

**Augmenting Loss Functions.** XBL methods that fall under this category follow the approach introduced in Equation 1 by adding an explanation loss to a model’s training to refine it to focus on image regions that are considered relevant by user(s) or to ignore confounding regions. One example of this category is Right for the Right Reasons (RRR) [Ross et al., 2017] that penalises a model with high input gradient model explanations on the wrong image regions based on user annotation. It uses,

$$L_{expl} = \sum_n^N \left[ M_n \frac{\partial}{\partial x_n} \sum_{k=1}^K \log \hat{y}_{nk} \right]^2 \quad (3)$$

for a function  $f(X|\theta) = \hat{y} \in \mathbb{R}^{N \times K}$  trained on images  $x_n$  of size  $N$  with  $K$  categories, where  $M_n \in \{0, 1\}$  is user annotation of image regions that should be avoided by the model.

Similarly, Right for Better Reasons (RBR) [Shao et al., 2021] uses Influence Functions (IF) in place of input gradients to correct a model’s behaviour. Contextual Decomposition Explanation Penalisation (CDEP) [Rieger et al., 2020] penalises features and feature interactions.

User feedback in XBL experiments can be either: (1) telling the model to ignore non-salient image regions; or (2) instructing the model to focus on important image regions in a training dataset [Hagos et al., 2022b]. While the XBL methods presented above refine a model by using the first feedback type, Human Importance-aware Network Tuning (HINT) does the opposite by teaching a model to focus on important image parts using GradCAM model explanations [Selvaraju et al., 2019].

**Augmenting Training Dataset.** In addition to augmenting loss functions, XBL can also be implemented by augmenting a training dataset based on user feedback. Instance relabelling [Teso et al., 2021], counterexamples generation [Teso and Kersting, 2019], and using user feedback as new training instances [Popordanoska et al., 2020] are some of the methods that augment a dataset to incorporate user feedback into XBL.

While XBL approaches show promise in unlearning spurious correlations that a model might have learned by giving attention to non-relevant or confounding image regions [Hagos et al., 2022a, Pfeuffer et al., 2023], they all need a lot of effort from users. In order to unlearn spurious correlations from a classifier, [Pfeuffer et al., 2023] collected feature annotation on 3000 chest x-ray images. This kind of demanding task hinders practical deployment and domain transferability of XBL. For this reason, it is of paramount importance to build an XBL method that can refine a trained model using a limited amount of user interaction in order to achieve a plausible and domain transferable implementation. To the best of our knowledge, this area of XBL is completely unexplored.

## 3 Exemplary eXplanation Based Learning

As is illustrated by Equations 2 and 3, for typical XBL approaches, user annotation of image features, or  $M$ , is an important prerequisite. We introduce Exemplary eXplanation Based Learning (eXBL) to mitigate the time and resource complexity caused by the feature annotation process. In eXBL, we propose to simplify the

expensive feature annotation requirement and replace it with two exemplary explanations: *Good GradCAM explanation* ( $C_{good}$ ) and *Bad GradCAM explanation* ( $C_{bad}$ ). However, even if this replaces feature annotation with two labels, categorising explanations would still be expensive if it's to be performed for all training instances whose size could be in the thousands. For this reason, we only use one  $C_{good}$  and one  $C_{bad}$ .

We choose to use GradCAM model explanations because they have been found to be more sensitive to training label reshuffling and model parameter randomisation than other saliency based explanations [Adebayo et al., 2018]. To select the good and bad explanations from a list of generated GradCAM explanations, we use an objective explanation metric, called Activation Recall (AR). AR measures how much of the actual relevant parts of test images,  $M$ , are considered relevant by a model. While a larger AR value means a model is giving higher attention to relevant image regions, a smaller AR would mean the model is not focusing on relevant image parts for its prediction. AR is formulated as follows,

$$AR_{x \in X} = \frac{\text{GradCAM}(x) * M}{M} \quad (4)$$

We then assign products of input instances and GradCAM explanation to  $C_{bad}$  and  $C_{good}$  using the instances with maximum and minimum AR values, as follows,

$$C_{good} := i \cdot \text{GradCAM}(i), \max_{x \in X}(AR(x)) := AR_i \quad (5)$$

$$C_{bad} := j \cdot \text{GradCAM}(j), \min_{x \in X}(AR(x)) := AR_j \quad (6)$$

The product of the input instance and the Grad-CAM explanation is used instead of just the Grad-CAM explanation because taking only the GradCAM outputs to be the good/ bad explanations could lead to biased exemplary explanations as it would mean we are only taking the model's focus or attention into consideration.

We then take inspiration from triplet loss to incorporate  $C_{good}$  and  $C_{bad}$  into our explanation loss. The main purpose of our explanation loss is to penalise a trainer according to its distance from  $C_{good}$  and  $C_{bad}$ : The closest to  $C_{good}$  and the furthest from  $C_{bad}$ , the lower the loss.

For the product of the training instances  $x \in X$ , and their corresponding GradCAM outputs,  $x \cdot \text{GradCAM}(x)$ , we compute the euclidean distances  $d_{xg}$  and  $d_{xb}$ , which represent distances from  $C_{good}$  and  $C_{bad}$  as follows,

$$d_{xg} := d(x \cdot \text{GradCAM}(x), C_{good}) \quad (7)$$

$$d_{xb} := d(x \cdot \text{GradCAM}(x), C_{bad}) \quad (8)$$

We train the model  $f$  to achieve  $d_{xg} \ll d_{xb}$  for all  $x$ . We do this by adding a  $\text{margin} = 1.0$ ;  $d_{xg} - d_{xb} + \text{margin} < 0$ .

We then compute the explanation loss as follows,

$$L_{expl} = \sum_i^N \max(d_{x_i g} - d_{x_i b} + \text{margin}, 0) \quad (9)$$

In addition to correctly classifying the training images, which is achieved through  $L_{CE}$ , this  $L_{expl}$  (Equation 9) would train  $f$  to output GradCAM values that resemble the good explanations and that differ from the bad explanations.

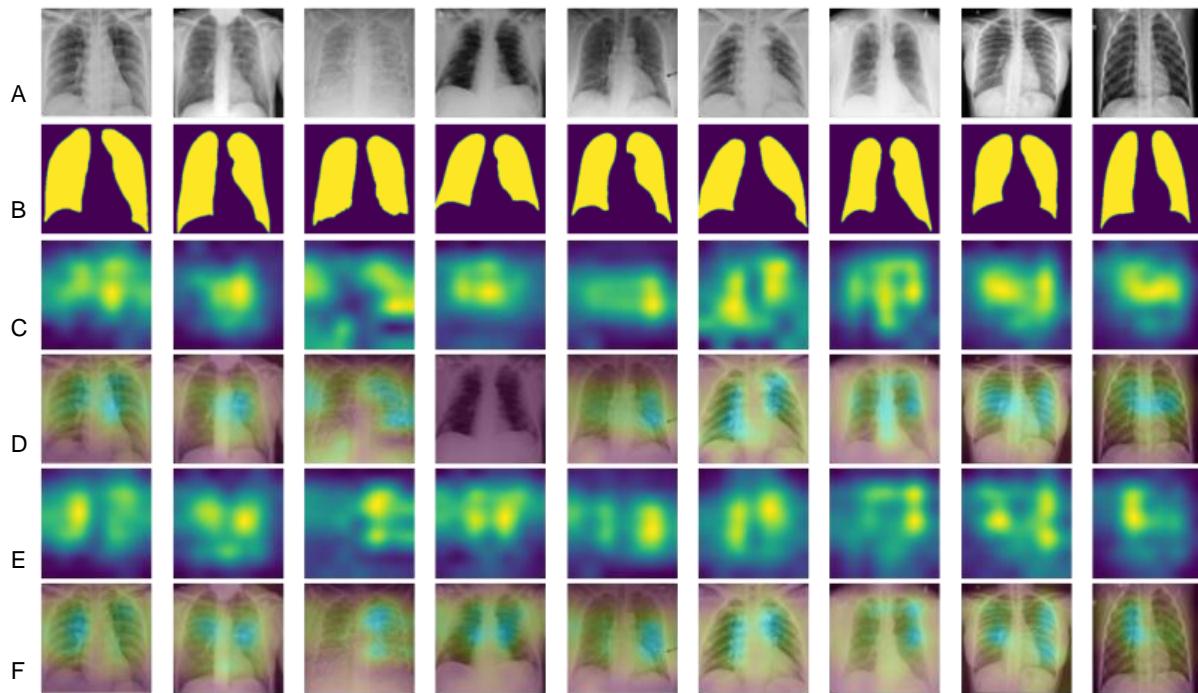


Figure 2: (A) Input images. (B) Feature annotations masks. (C) GradCAM explanations of the Unrefined model. (D) GradCAM outputs of Unrefined model overlaid over input images. (E) GradCAM explanations of eXBL. (F) GradCAM outputs of eXBL model overlaid over input images.

## 4 Experiments

### 4.1 Data Collection and Preparation

We use the Covid-19 Radiography Database dataset [Chowdhury et al., 2020, Rahman et al., 2021]<sup>1</sup> which contains chest x-ray images of four categories: covid, normal, lung opacity, and viral pneumonia. We downsample the dataset to circumnavigate class imbalance. For model training we used 800 x-ray images per category totalling 3200 images. For validation and testing, we collected 1200 and 800 total images. We resize all images to  $224 \times 224$  pixels. The dataset is also accompanied with feature annotation masks that show the relevant regions for each of the x-ray images collected from radiologists [Rahman et al., 2021].

Even though the exact number of effected images is unknown, the dataset contains confounding regions, such as marks, texts, and timestamps in many of the images.

### 4.2 Model Training

We followed a transfer learning approach using a pre-trained MobileNetV2 model [Sandler et al., 2018]. We chose to use MobileNetV2 because it achieved better performance at the chest x-ray images classification task at a reduced computational cost after comparison against pre-trained models available at the Keras website<sup>2</sup>. In order for the training process to affect the GradCAM explanation outputs, we only freeze and reuse the first 50 layers of MobileNetV2 and retrain the rest of the convolutional layers with a classifier layer (256 nodes with a ReLu activation with a 50% dropout followed by a Softmax layer with 4 nodes) that we added.

<sup>1</sup><https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database>

<sup>2</sup><https://keras.io/api/applications/>

We first trained the MobileNetV2 to categorise the training set into the four classes using categorical cross entropy. It was trained for 60 epochs<sup>3</sup> using Adam optimiser. We refer to this model as the Unrefined model. We use the Unrefined model to extract good and bad GradCAM explanations. Next, we employ our eXBL algorithm using the good and bad explanations to teach the Unrefined model to focus on relevant image regions by tuning its explanations to look like the good explanations and differ from the bad explanations as much as possible. We refer to this model as the eXBL model and it was trained for 100 epochs using the same early stopping, learning rate, and optimiser as the Unrefined model.

## 5 Results

Tables 1 and 2 show classification performance of the Unrefined and eXBL refined models. While the average AR score of GradCAM explanations produced using the eXBL model is 0.705, the explanations of the Unrefined model score an average AR of 0.685. Sample test images, masks, GradCAM outputs, and overlaid GradCAM visualisations of both the Unrefined and eXBL models are displayed in Figure 2. From the sample outputs, we observe that the eXBL model was able to produce more accurate explanations that capture the relevant image regions presented with annotation masks. However, the superior explanations of the eXBL model come with a classification performance loss on half of the categories as is summarised in Table 2.

METRIC	UNREFINED MODEL	EXBL
ACCURACY	0.950	0.910
PRECISION	0.947	0.912
RECALL	0.945	0.902

Table 1: Summary of classification performances of the Unrefined and eXBL models.

CATEGORY	UNREFINED MODEL	EXBL
COVID	0.925	0.855
NORMAL	0.930	0.955
LUNG OPACITY	0.955	0.955
VIRAL PNEUMONIA	0.975	0.945

Table 2: Classification performance into four categories.

## 6 Conclusion

In this work, we have presented an approach to simplify the demanding task of feature annotation in XBL to an identification of only two model explanations. Our approach, Exemplary eXplanation-based Learning (eXBL) can tune a model’s attention to focus on relevant image regions, thereby improving the saliency-based model explanations. We believe our approach is domain transferable and shows potential for real-world implementation of interactive learning using XBL.

Even though the eXBL model achieved comparable classification performance when compared against the Unrefined model (especially in categorising the Normal and Lung opacity categories, in which it scored better and equal to the Unrefined model, respectively), as is presented in Tables 1 and 2, we observed that there is a classification performance loss when retraining the Unrefined model with eXBL to produce good explanations. We attribute this to the accuracy-interpretability trade-off. Although the existence of this trade-off is debated [Rudin, 2019, Dziugaite et al., 2020], performance loss after retraining a model could mean that the initial model was exploiting confounding regions in the training instances. It could also mean that our selection of good and bad explanations may not have been optimal and that the two exemplary explanations may be degrading model performance.

<sup>3</sup>The model was trained with an early stop monitoring the validation loss at a patience of five epochs and a decaying learning rate = 1e-04.

The two exemplary explanations are selected using an objective evaluation metric, AR, and an existing dataset of annotation masks. For system development and experiment purposes, we use the masks as base knowledge. Although we believe our work presents a simple approach to implement XBL on other domains, future work should involve domain experts when picking the good and bad explanations. However, when involving end users, since the pool of explanations to choose the exemplary explanations from could be large, active learning approaches should be explored to select a subset of model explanations to prompt domain experts for feedback.

## Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Adebayo et al., 2018] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.
- [Budd et al., 2021] Budd, S., Robinson, E. C., and Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062.
- [Chechik et al., 2010] Chechik, G., Sharma, V., Shalit, U., and Bengio, S. (2010). Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3).
- [Chowdhury et al., 2020] Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Al Emadi, N., et al. (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676.
- [Dziugaite et al., 2020] Dziugaite, G. K., Ben-David, S., and Roy, D. M. (2020). Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*.
- [Fails and Olsen Jr, 2003] Fails, J. A. and Olsen Jr, D. R. (2003). Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 39–45.
- [Hagos et al., 2022a] Hagos, M. T., Curran, K. M., and Mac Namee, B. (2022a). Identifying spurious correlations and correcting them with an explanation-based learning. *arXiv preprint arXiv:2211.08285*.
- [Hagos et al., 2022b] Hagos, M. T., Curran, K. M., and Mac Namee, B. (2022b). Impact of feedback type on explanatory interactive learning. In *International Symposium on Methodologies for Intelligent Systems*, pages 127–137. Springer.
- [Kulesza et al., 2015] Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 126–137.
- [O’Neill et al., 2017] O’Neill, J., Delany, S. J., and Mac Namee, B. (2017). Rating by ranking: An improved scale for judgement-based labels. In *IntRS@ RecSys*, pages 24–29.
- [Pfeuffer et al., 2023] Pfeuffer, N., Baum, L., Stammer, W., Abdel-Karim, B. M., Schramowski, P., Bucher, A. M., Hügel, C., Rohde, G., Kersting, K., and Hinz, O. (2023). Explanatory interactive machine learning. *Business & Information Systems Engineering*, pages 1–25.

- [Popordanoska et al., 2020] Popordanoska, T., Kumar, M., and Teso, S. (2020). Machine guides, human supervises: Interactive learning with global explanations. *arXiv preprint arXiv:2009.09723*.
- [Rahman et al., 2021] Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., Al Maadeed, S., Zughraier, S. M., Khan, M. S., et al. (2021). Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319.
- [Rieger et al., 2020] Rieger, L., Singh, C., Murdoch, W., and Yu, B. (2020). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR.
- [Ross et al., 2017] Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*.
- [Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- [Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.
- [Schramowski et al., 2020] Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486.
- [Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.
- [Selvaraju et al., 2019] Selvaraju, R. R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., and Parikh, D. (2019). Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2591–2600.
- [Shao et al., 2021] Shao, X., Skryagin, A., Stammer, W., Schramowski, P., and Kersting, K. (2021). Right for better reasons: Training differentiable models by constraining their influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9533–9540.
- [Stumpf et al., 2009] Stumpf, S., Rajaram, V., Li, L., Wong, W.-K., Burnett, M., Dietterich, T., Sullivan, E., and Herlocker, J. (2009). Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662.
- [Teso et al., 2023] Teso, S., Alkan, , Stammer, W., and Daly, E. (2023). Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence*, 6.
- [Teso et al., 2021] Teso, S., Bontempelli, A., Giunchiglia, F., and Passerini, A. (2021). Interactive label cleaning with example-based explanations. *Advances in Neural Information Processing Systems*, 34:12966–12977.
- [Teso and Kersting, 2019] Teso, S. and Kersting, K. (2019). Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245.
- [Zlateski et al., 2018] Zlateski, A., Jaroensri, R., Sharma, P., and Durand, F. (2018). On the importance of label quality for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1479–1487.

# DF-Net: The Digital Forensics Network for Image Forgery Detection

David Fischinger and Martin Boyer

*Austrian Institute of Technology*

## Abstract

The orchestrated manipulation of public opinion, particularly through manipulated images, often spread via online social networks (OSN), has become a serious threat to society. In this paper we introduce the Digital Forensics Net (DF-Net), a deep neural network for pixel-wise image forgery detection. The released model outperforms several state-of-the-art methods on four established benchmark datasets. Most notably, DF-Net's detection is robust against lossy image operations (*e.g.* resizing, compression) as they are automatically performed by social networks.

**Keywords:** Image Manipulation Detection and Localization, Digital Forensics, DF-Net

## 1 Introduction

"Fake News" poses an ever-growing challenge in our society as technological advancements facilitate the production of high-quality forgeries in digital media like audio, video, and images. This impact ranges from satirical memes to orchestrated political Fake News campaigns that aim to manipulate public opinion. In this paper, we introduce an effective approach to identify manipulated regions in images. This enables institutions such as media organizations and interested citizens to get a better indication of whether specific images may have been manipulated.

Over the past decade, various methods have been proposed to detect different categories of image forgery, including: copy-move, splicing, inpainting, and various enhancement techniques. However, these approaches often concentrate on specific features of each manipulation type. In recent years, more general approaches for multiple manipulation types were developed, such as [Wu et al., 2019] and [Wu et al., 2022]. Each of them promotes sophisticated and problem-specific network architectures and concepts, like modeling known and unknown noise on images that result from transmission to Online Social Networks (OSN).

In this paper we present the DF-Net, an image forgery detector trained on the DF2023 dataset [Fischinger and Boyer, 2023]. To be more specific, our main contributions are as follows:

- **Model:** Our proposed forgery detection model outperforms several state-of-the-art methods. We show its evaluation on four benchmark datasets. The model's deep learning network architecture combines the strengths of two specialized sub-models, trained from scratch on our DF2023 dataset.
- **OSN robustness:** We show the robustness of our model against lossy operations (*e.g.* resizing, compression) as automatically done by online social networks in an extensive evaluation.
- **Speed:** We present a processing time comparison with a SOTA approach that shows a significant reduction in time, especially for larger images.

## 2 Related Work

Many methods of detecting and localizing image forgery were published (see, for example, the review of [Verdoliva, 2020] and references therein) in order to ensure visual information authenticity. Some of these forensic techniques are designed to detect specific forms of tampering, such as splicing [Lyu et al., 2013],

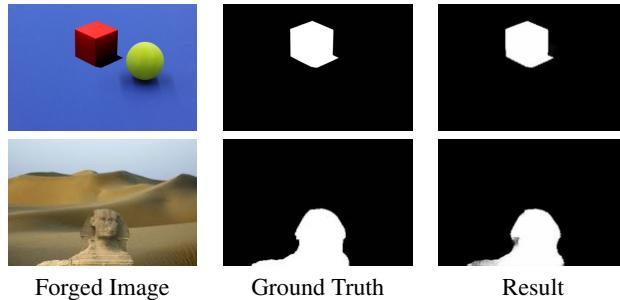


Figure 1: Forgery detection results of our network. Example images are taken from the CASIA [Dong et al., 2013] and the NIST [National Institute of Standards and Technology (NIST), 2016] datasets.

copy-move [Wang et al., 2017, Mahmood et al., 2017, Ouyang et al., 2019, Zedan et al., 2021, Zhong and Pun, 2020], and inpainting [Li et al., 2017]. Unfortunately, these forensic approaches can only be applied to detect specific tampering manipulations.

In recent years, deep learning-based methods were developed to address the problem of detecting general (compound) types of forgeries. Notably, [Wu et al., 2019] proposes a unified deep neural architecture called ManTra-Net, which is an end-to-end network that performs both detection and localization without extra preprocessing and postprocessing. ManTra-Net is a fully convolutional network which can handle images of arbitrary sizes and many known – and even unknown – forgery types. Furthermore, the authors design a self-supervised learning task to learn robust image manipulation features, formulate the forgery localization problem as a local anomaly detection problem, and propose a long short-term memory (LSTM) solution to assess local anomalies.

The work of [Zhuang et al., 2021] addresses the issue of tampering localization by focusing on the detection of commonly used editing tools and operations in Photoshop. A fully convolutional encoder-decoder architecture is designed, as well as a training data generation strategy by resorting to Photoshop scripting.

The widespread availability of online social networks (OSN), e.g., Twitter, Facebook, Whatsapp, etc., makes them the dominant channels for transmitting forged images. However, almost all OSN manipulate the uploaded images in a lossy fashion (including format conversion, resizing, enhancement filtering and JPEG compression). The noise introduced by these lossy operations could severely affect the effectiveness of forensic methods. In a recent paper by [Wu et al., 2022], the problem of OSN-shared image forgeries is tackled by employing a dedicated training scheme. A baseline detector is presented, which is based on a modified U-Net [Ronneberger et al., 2015] as the backbone architecture. Next, an analysis of the noise introduced by OSN is conducted, and the noise is decoupled into two parts, i.e., predictable noise and unseen noise. These are then modelled separately and the modelled noise is further incorporated into the training framework.

*Outline:* The rest of this paper is structured as follows: In section 3, we present the DF-Net, evaluate different model design choices and investigate combinations of the models. In section 4, our proposed model is evaluated and compared to state-of-the-art methods, specifically for OSN transmitted images. Final remarks are made in section 5.

### 3 Network Architecture

#### 3.1 Architecture

The DF-Net (available from <https://zenodo.org/record/8142658>) is designed to detect and localize image forgeries of various types. Essentially, this is a binary segmentation problem in which each pixel of an image is classified as either pristine or forged, resulting in a binary mask  $M$ .

Our proposed network comprises two sub-networks ( $M_1, M_2$ ). Both networks use U-Net [Ronneberger et al., 2015] implementations, an architecture commonly used in the area of image segmentation.

The U-Net architecture of M1 is depicted in Fig. 2: U-Nets are Convolutional Neural Networks (CNNs) which consist of an encoding part where the spatial dimensions are down-scaled (downsampling), and a de-coding part where the spatial dimensions are increased (upsampling). On the first four sampling stages, skip connections are used which provide multi-channel feature maps from the encoding part directly to the decoding stage with the same spatial dimensions. Our U-Net implementation takes RGB images of size (256,256,3) as input. In each scaling step, we use two times a building block consisting of a 3x3 convolution, a batch normalization layer and a Relu activation layer, followed by a spatial and channel Squeeze & Excitation (scSE) block [Roy et al., 2018] and a (2x2) Max-pooling (downscaling) or a (3x3) Conv2DTranspose layer (upscaling). In the upscaling phase, the skip features are concatenated with the output of the Conv2DTranspose layer. The scSE layer can be seen as a re-calibration method for the network with a relatively small overhead regarding computing resources. During the training process, these blocks amplify spatial areas and channels which contribute more to better solutions and diminish the influence of worse performing network parts. At the final layer of the network, a 1x1 convolution with sigmoid activation function is used to calculate a value in [0, 1] which indicates how likely each pixel is manipulated.

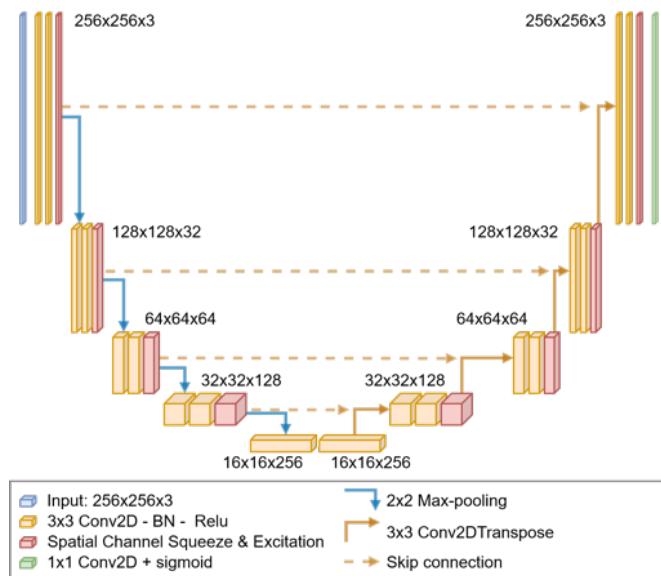


Figure 2: Network architecture of submodel M1: A U-Net architecture with 4 skip connections and spatial channel Squeeze & Excitation (scSE) extension. A more detailed description can be found in section 3.1.

The architecture of model M2 slightly deviates from M1: for each Conv2D block of M2, the kernel size is set to 5x5 instead of 3x3 for exactly 4 filters. The output feature maps of the filters with different-sized kernels are concatenated again before the batch normalization is done.

We listed major evolutionary steps of the DF-Net in Tab. 1, where we compared networks on four benchmark datasets. Considerable performance boosts were accomplished when adding the spatial channel Squeeze & Excitation (scSE) calibration blocks (plus of 0.082) and after the input size for training and prediction was increased from (224x224) to (256x256) (0.029). Model M2 was only trained on splicing images from DF2023, but still performed best out of all single sub-networks. The architectural modification replacing four 3x3 kernels with 5x5 kernels in each convolution gave an additional boost of 0.014. A considerable performance gain was reached by combining networks. With the maximum operator, better results were achieved compared to averaging the predictions of two single networks. Taking the maximum prediction value from M1 and M2 for each image pixel, resulted in an average value of 0.583 which exceeds the best performing sub-model M2, which only achieved 0.535 overall. A deeper analysis showed that the maximum-operator-based model performs

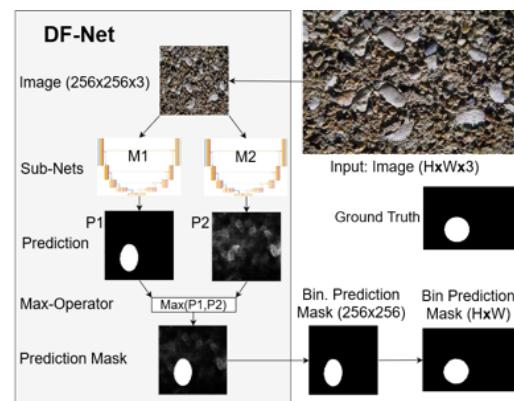


Figure 3: Network architecture of DF-Net: Example of Model combination for image Sp\_S\_NNN\_C\_txt0019\_txt0019\_0019.jpg from the CASIA\_V1 dataset.

worse than the better sub-model, for most images. For the CASIA dataset, the AUC value for max(M1,M2) is higher than the better performing submodule for only 86 out of 920 images. But the AUC performance is considerably higher than the average AUC value of the two sub-modules. In other words: By applying the maximum operator, the models prediction on each image is generally almost as good as the better-performing sub-module. So, the final DF-Net architecture depicted in Fig. 3 combines the strengths of the separately trained models M1 and M2 by taking the maximum prediction value per pixel from the outputs of both sub-nets.

Model Arch.	Test Datasets																	
	Average(metrics)				CASIA			Columbia			DSO			NIST				
	Mean	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU		
U-Net (224 <sup>2</sup> )	<b>.390</b>	.701	.263	.206	.75	.21	.18	.80	.52	.41	.60	.10	.06	.65	.22	.17		
M1 (224 <sup>2</sup> )	<b>.472</b>	.797	.337	.282	.85	.43	.37	.83	.50	.43	.73	.15	.11	.78	.26	.22		
M1 (256 <sup>2</sup> )	<b>.501</b>	.782	.396	.326	.85	.53	.47	.79	.54	.45	.71	.24	.16	.77	.28	.23		
M2: M1-Ar.	<b>.521</b>	.790	.425	.348	.82	.42	.36	.87	.70	.61	.74	.32	.22	.74	.26	.20		
M2	<b>.535</b>	.804	.439	.363	.82	.43	.37	.89	.74	.67	.76	.31	.21	.75	.27	.20		
avg(M1,M2)	<b>.574</b>	.842	.474	.405	.91	.60	.54	.88	.69	.63	.78	.30	.21	.80	.30	.24		
max(M1,M2)	<b>.583</b>	.837	.501	.411	.91	.59	.50	.88	.76	.68	.77	.36	.25	.79	.30	.23		

Table 1: Comparison of different network implementations we have trained and evaluated. The first two networks were trained on images of size 224x224, all other networks on images of size 256x256. Column **Mean** shows the average of the AUC, F1 and IoU metrics over all 4 datasets. The last row shows results from the DF-Net. More details are given in Sec. 3.1

Our experiments show that the combination of separately trained sub-networks results in a considerable performance improvement (see Tab. 1). Equally beneficiary is the advantage of splitting the whole model to sub-models that can be trained separately. This reduces time because researchers can faster assess if changes in the network architecture, the training data or the training parameters should be discarded or pursued. Furthermore, it allows to overcome hardware limitations which we see as a major advantage of this architecture.

### 3.2 Implementation Details

Model specification and training were done in the deep learning framework Tensorflow. For training and detection, the images were resized to (256,256) pixels. An Nvidia GeForce GTX 1080 Ti GPU with 11G memory was used for training, with batch size set to 32. We used the Adam optimizer [Kingma and Ba, 2015] and performed 1000 steps per epoch and stopped after the loss value of the validation dataset did not improve for 35 epochs. Training starts with a learning rate of 0.0001, which is halved after 10 epochs without improvement. M1 was first trained on all manipulation types and images. Subsequently it was refined by training on the 200.000 copy-move forgeries from DF2023 [Fischinger and Boyer, 2023]. Model M2 was trained only on the 400K splicing images from DF2023.

## 4 Evaluation

We compare the DF-Net to four state-of-the-art methods: ForSim [Mayer and Stamm, 2019], DFCN [Zhuang et al., 2021], ManTra-Net [Wu et al., 2019] and the work of Wu *et al.* [Wu et al., 2022], abbreviated below as Wu22. The approaches are evaluated on the four benchmark datasets CASIA\_V1 [Dong et al., 2013], Columbia [Hsu and Chang, 2006], DSO [Carvalho et al., 2013] and NIST16 [National Institute of Standards and Technology (NIST), 2016]. See Table 2 for an overview of the datsets.

### 4.1 Online Social Networks

The popularity of online social networks (OSN) makes them the dominating channels for the distribution of manipulated images in the context of entertainment, but also for fake news, disinformation and propaganda.

Dataset	# Images	Format	t-WU22	t-DF-N
CASIA [Dong et al., 2013]	920	jpg	169	<b>155</b>
Columbia [Hsu and Chang, 2006]	160	tif	120	<b>28</b>
DSO [Carvalho et al., 2013]	100	png	701	<b>27</b>
NIST [National Institute of Standards and Technology (NIST), 2016]	564	jpg	15250	<b>235</b>

Table 2: Benchmark datasets with processing times (t) for Wu22 [Wu et al., 2022] and DF-Net (ours) in seconds for predictions per benchmark dataset. For datasets with huge images such as NIST (images of size up-to 5616×3744 pixels), tile-based approaches take considerably longer than approaches performing pre-scaling.

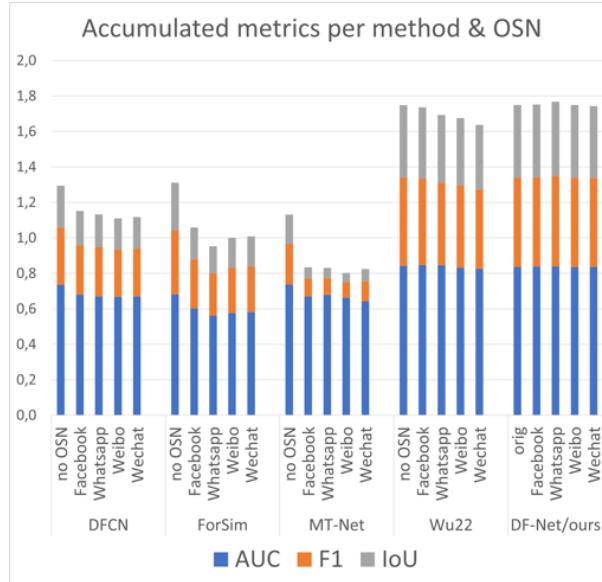


Figure 4: Metrics (AUC, F1, IoU) averaged over 4 benchmark datasets in accumulated presentation. Each column represents the combination of a method and the OSN used for dataset modification

Unfortunately, OSN automatically apply operations like compression and resizing, which reduce valuable information for image forgery detection. To show the robustness of DF-Net against these lossy operations, all the SOTA methods are tested against OSN adapted versions of the four benchmark datasets. In [Wu et al., 2022], the authors transmitted the images of the benchmark datasets via the social online platforms Facebook, Whatsapp, Weibo, and Wechat and made the collected datasets and their evaluation of several state-of-the-art methods available to the research community. We could reproduce the results for Wu22 [Wu et al., 2022]. For ManTra-Net [Wu et al., 2019], we tested the officially released TensorFlow model which is different from the model used for the authors' evaluation as the authors stated in [Wu, 2023], and a public PyTorch re-implementation [Abecidan, 2023] as well. Here we could reproduce the ManTraNet results as stated in [Wu et al., 2022]. For ForSim [Mayer and Stamm, 2019] and DFCN [Zhuang et al., 2021] results from the evaluation in [Wu et al., 2022] are stated in Tab. 3, together with the evaluation of the proposed DF-Net.

#### 4.1.1 Evaluation Criteria:

We adopt three metrics commonly used in the area of image forgery detection: Area under the receiver operating characteristic curve (AUC), F1-score and Intersection over Union (IoU). The metrics are calculated on a pixel-level. For IoU and F1-score, the threshold for the output of the trained networks is set to 0.5.

Models	OSN	Test Datasets																		
		CASIA						Columbia						DSO						Average
		AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	Mean
DFCN	-	.654	.192	.119	.789	.541	.395	.724	.303	.227	.778	.250	.204	.736	.322	.236	.431			
FSim	-	.554	.169	.102	.731	.604	.474	.796	<b>.487</b>	<b>.371</b>	.642	.188	.123	.681	.362	.268	.437			
MNet	-	.776	.130	.086	.747	.357	.258	.795	.344	.253	.634	.088	.054	.738	.230	.163	.377			
Wu22	-	.873	.509	.465	.862	.707	.608	<b>.854</b>	.436	.308	.783	<b>.332</b>	<b>.255</b>	<b>.843</b>	.496	.409	.5827			
<b>DF-Net</b>	-	<b>.906</b>	<b>.589</b>	<b>.496</b>	<b>.880</b>	<b>.757</b>	<b>.679</b>	.769	.360	.246	<b>.793</b>	.299	.226	.837	<b>.501</b>	<b>.411</b>	<b>.5832</b>			
DFCN	Facebook	.654	.190	.116	.687	.479	.338	.673	.238	.184	.705	.207	.138	.680	.278	.194	.384			
FSim	Facebook	.537	.157	.094	.607	.450	.304	.689	.356	.238	.580	.140	.085	.603	.276	.180	.353			
MNet	Facebook	.763	.102	.065	.626	.103	.056	.638	.109	.071	.652	.095	.057	.670	.102	.062	.278			
Wu22	Facebook	.862	.462	.417	<b>.883</b>	.714	.611	<b>.859</b>	<b>.447</b>	<b>.320</b>	.783	<b>.329</b>	<b>.253</b>	<b>.847</b>	.488	.400	.578			
<b>DF-Net</b>	Facebook	<b>.905</b>	<b>.587</b>	<b>.492</b>	.883	<b>.760</b>	<b>.681</b>	.770	.359	.245	<b>.795</b>	.304	.229	.838	<b>.502</b>	<b>.412</b>	<b>.584</b>			
DFCN	Wechat	.651	.193	.119	.676	.487	.344	.653	.221	.137	.701	.176	.114	.670	.269	.179	.373			
FSim	Wechat	.532	.153	.091	.650	.496	.354	.564	.247	.147	.581	.136	.082	.582	.258	.168	.336			
MNet	Wechat	.724	.080	.048	.613	.199	.125	.582	.076	.045	.654	.095	.057	.643	.113	.069	.275			
Wu22	Wechat	.833	.405	.358	<b>.883</b>	.727	.631	<b>.823</b>	<b>.366</b>	<b>.252</b>	.764	.286	.214	.826	.446	.364	.545			
<b>DF-Net</b>	Wechat	<b>.902</b>	<b>.564</b>	<b>.467</b>	.881	<b>.759</b>	<b>.681</b>	.765	.358	.245	<b>.799</b>	<b>.314</b>	<b>.238</b>	<b>.837</b>	<b>.499</b>	<b>.408</b>	<b>.581</b>			
DFCN	Whatsapp	.655	.191	.117	.692	.471	.331	.645	.264	.162	.689	.187	.125	.670	.278	.184	.377			
FSim	Whatsapp	.525	.151	.091	.595	.436	.294	.542	.233	.139	.586	.137	.082	.562	.239	.152	.318			
MNet	Whatsapp	.763	.099	.063	.630	.098	.052	.616	.081	.052	.702	.101	.062	.678	.095	.057	.277			
Wu22	Whatsapp	.866	.478	.431	<b>.889</b>	.727	.628	<b>.839</b>	.341	.233	.785	.313	.239	<b>.845</b>	.465	.383	.564			
<b>DF-Net</b>	Whatsapp	<b>.905</b>	<b>.588</b>	<b>.495</b>	.883	<b>.762</b>	<b>.685</b>	.765	<b>.361</b>	<b>.249</b>	<b>.803</b>	<b>.324</b>	<b>.247</b>	.839	<b>.509</b>	<b>.419</b>	<b>.589</b>			
DFCN	Weibo	.653	.191	.117	.676	.458	.319	.639	.227	.140	.706	.192	.125	.668	.267	.175	.370			
FSim	Weibo	.542	.165	.100	.610	.453	.312	.568	.260	.165	.581	.150	.094	.575	.257	.168	.333			
MNet	Weibo	.754	.099	.063	.620	.103	.056	.606	.057	.036	.671	.088	.053	.663	.087	.052	.267			
Wu22	Weibo	.858	.466	.421	.883	.724	.626	<b>.808</b>	<b>.370</b>	<b>.253</b>	.780	.294	.219	.832	.463	.380	.558			
<b>DF-Net</b>	Weibo	<b>.902</b>	<b>.584</b>	<b>.490</b>	<b>.890</b>	<b>.766</b>	<b>.684</b>	.759	.354	.245	<b>.791</b>	<b>.303</b>	<b>.230</b>	<b>.836</b>	<b>.502</b>	<b>.412</b>	<b>.583</b>			

Table 3: Comparison of the state-of-the-art approaches DFCN [Zhuang et al., 2021], ForSim [Mayer and Stamm, 2019] (FSim), ManTra-Net [Wu et al., 2019] (MNet), Wu22 [Wu et al., 2022] and our proposed DF-Net. Highest metric values per benchmark dataset and OSN are marked **bold**.

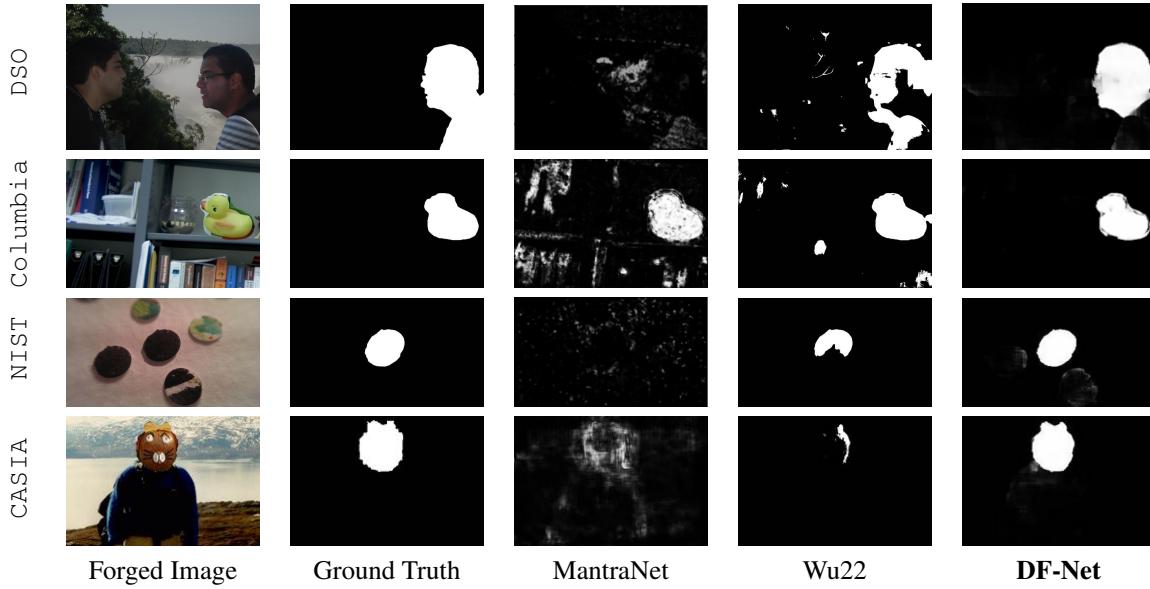


Figure 5: Examples of qualitative comparison of MantraNet [Wu et al., 2019], Wu22 [Wu et al., 2022] and our proposed DF-Net. Each line shows one example image for each of the four benchmark datasets DSO [Carvalho et al., 2013], Columbia [Hsu and Chang, 2006], NIST [National Institute of Standards and Technology (NIST), 2016], CASIA [Dong et al., 2013]. The five columns show: the forged image (input), manipulated area (ground truth), results (output) from MantraNet, Wu22 and DF-Net. We show example results of the M1 sub-model for the DSO and the NIST dataset.

## 4.2 Quantitative Comparison

As shown in Tab. 3, the DF-Net could clearly outperform ForSim, DFCN and ManTra-Net on the original benchmark datasets CASIA V1, Columbia and NIST16, and on the overall average of the metrics. The method Wu22 performed similar to the DF-Net. The overall average of DF-Net (0.5832) is just 0.0005 higher. Yet the situation for the OSN modified datasets has to be noted: in Fig. 4 we visualize the sum of AUC, F1 and IoU per dataset. The methods ForSim, DFCN and ManTra-Net show a large performance decrease for the modified datasets. Robustness against OSN transmitted data was the key contribution of Wu22 [Wu et al., 2022], hence their method performs only moderately worse on OSN transmitted images compared to the original ones. DF-Net on the other hand does not show a significant performance drop at all. This can be explained by the training process with image pre-scaling to 256x256 pixels. This forces the DF-Network to learn manipulation traces which are even included in downsampled images.

Curiously enough, the metrics for DF-Net do sometimes even improve on the OSN-transmitted dataset versions. The effect is strongest for WhatsApp transmitted image data, where the average of the three metrics over all benchmark datasets increases by 0.00578. This effect can also be found for Wu22 [Wu et al., 2022], where all metrics are higher for the Facebook transmitted Columbia dataset compared to the original (non-transmitted) data.

## 5 Conclusion

In this paper, we propose a lightweight network architecture for image manipulation detection. We share our model, the Digital Forensics Network (DF-Net, with the community. This model shows a better performance than several state-of-the-art methods on four well-established benchmark datasets. In particular, DF-Net outperforms its competitors for images transmitted over popular social networks such as Facebook or WhatsApp. With a simple and practical training concept, the DF-Net addresses the challenges of lossy operations (downscaling, filtering) and focuses on robust manipulation features. In extensive evaluations, we show that DF-Net has virtually no performance degradation on OSN-transmitted images, which is a unique feature compared to competitors in the field of image forgery detection. Furthermore, the detection speed is significantly higher than that of its closest competitor since DF-Net does not rely on a tiling process (see Tab. 2).

## Acknowledgement



Co-funded by  
the European Union

Project 101083573 — GADMO

## References

- [Abecidan, 2023] Abecidan, R. (2023). ManTraNet pytorch implementation. <https://github.com/RonyAbecidan/Mantranet-pytorch>. Accessed: 2023-03-06.
- [Carvalho et al., 2013] Carvalho, T., Riess, C., Angelopoulou, E., Pedrini, H., and Rocha, A. R. (2013). Exposing digital image forgeries by illumination color classification. *IEEE Trans. Inf. Forensics and Security*, 8(7):1182–1194.
- [Dong et al., 2013] Dong, J., Wang, W., and Tan, T. (2013). Casia image tampering detection evaluation database. In *IEEE China Summit Inter. Conf. Signal Info. Proc.*, pages 422–426. IEEE.
- [Fischinger and Boyer, 2023] Fischinger, D. and Boyer, M. (2023). DF2023: The digital forensics 2023 dataset for image forgery detection. *Irish Machine Vision and Image Processing conference*.
- [Hsu and Chang, 2006] Hsu, Y. and Chang, S. (2006). Detecting image splicing using geometry invariants and camera characteristics consistency. In *IEEE Inter. Conf. Multim. Expo*, pages 549–552. IEEE.

- [Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*.
- [Li et al., 2017] Li, H., Luo, W., and Huang, J. (2017). Localization of diffusion-based inpainting in digital images. *IEEE Transactions on Information Forensics and Security*, 12(12):3050–3064.
- [Lyu et al., 2013] Lyu, S., Pan, X., and Zhang, X. (2013). Exposing region splicing forgeries with blind local noise estimation. *International Journal of Computer Vision*, 110:202–221.
- [Mahmood et al., 2017] Mahmood, T., Irtaza, A., Mehmood, Z., and Mahmood, M. (2017). Copy-move forgery detection through stationary wavelets and local binary pattern variance for forensic analysis in digital images. *Forensic Science International*, Elsevier, 279:8–21.
- [Mayer and Stamm, 2019] Mayer, O. and Stamm, M. C. (2019). Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*.
- [National Institute of Standards and Technology (NIST), 2016] National Institute of Standards and Technology (NIST) (2016). Nist nimble 2016 datasets. <https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation/>.
- [Ouyang et al., 2019] Ouyang, J., Liu, Y., and Liao, M. (2019). Robust copy-move forgery detection method using pyramid model and zernike moments. *Multimedia Tools and Applications*, 78:1–19.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*.
- [Roy et al., 2018] Roy, A. G., Navab, N., and Wachinger, C. (2018). Recalibrating fully convolutional networks with spatial and channel 'squeeze & excitation' blocks. In *Medical Imaging*, pages 540–549.
- [Verdoliva, 2020] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932.
- [Wang et al., 2017] Wang, Y., Tian, L., and Li, C. (2017). Lbp-svd based copy move forgery detection algorithm. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 553–556.
- [Wu et al., 2022] Wu, H., Zhou, J., Tian, J., Liu, J., and Qiao, Y. (2022). Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*.
- [Wu, 2023] Wu, Y. (2023). ManTraNet github repository. <https://github.com/ISICV/ManTraNet>. Accessed: 2023-06-09.
- [Wu et al., 2019] Wu, Y., AbdAlmageed, W., and Natarajan, P. (2019). Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9535–9544.
- [Zedan et al., 2021] Zedan, I. A., Soliman, M. M., Elsayed, K. M., and Onsi, H. M. (2021). Copy move forgery detection techniques: A comprehensive survey of challenges and future directions. *International Journal of Advanced Computer Science and Applications*, 12(7).
- [Zhong and Pun, 2020] Zhong, J.-L. and Pun, C.-M. (2020). An end-to-end dense-inceptionnet for image copy-move forgery detection. *IEEE Transactions on Information Forensics and Security*, 15:2134–2146.
- [Zhuang et al., 2021] Zhuang, P., Li, H., Tan, S., Li, B., and Huang, J. (2021). Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security*, 16:2986–2999.

# DF2023: The Digital Forensics 2023 Dataset for Image Forgery Detection

David Fischinger and Martin Boyer

*Austrian Institute of Technology*

## Abstract

The deliberate manipulation of public opinion, especially through altered images, which are frequently disseminated through online social networks, poses a significant danger to society. To fight this issue on a technical level we support the research community by releasing the Digital Forensics 2023 (DF2023) training and validation dataset, comprising one million images from four major forgery categories: splicing, copy-move, enhancement and removal. This dataset enables an objective comparison of network architectures and can significantly reduce the time and effort of researchers preparing datasets.

**Keywords:** Image Manipulation Detection, Training Dataset, Benchmark Dataset, DF2023

## 1 Introduction

The proliferation of fake news presents a mounting concern in our society. Advances in technology have facilitated the swift and seamless production of convincing counterfeit digital media content, encompassing audio, video, and images. This impact spans from humorous satirical memes to organized political campaigns that disseminate fabricated news in order to manipulate public sentiment.

This paper addresses the issue of identifying local image forgeries. Over the past decade, several methods have been proposed in order to detect the main categories of image forgery: copy-move [Li et al., 2013], splicing [Lyu et al., 2013], inpainting [Li et al., 2017] and other specific filtering techniques, subsumed as enhancement [Sun et al., 2018]. However, these detection methods often concentrate on specific characteristics of each manipulation type. In recent years, more comprehensive approaches capable of detecting multiple types of manipulation have emerged, such as those presented in [Wu et al., 2019] and [Wu et al., 2022].

However, the research community is still in need of a large and more generalized dataset which enables training and hence an objective comparison of network architectures for the issue of image forgery detection. In this paper, we close this gap by introducing the Digital Forensics 2023 (DF2023) dataset. This training dataset is comprised of one million manipulated images specifically designed for image forgery detection and localization. By making the DF2023 dataset publicly available, it provides the research community with the means to conduct unbiased comparisons of network architectures and reduces the time and effort required for preparing training data.

## 2 Related Work

Many methods of detecting and localizing image forgery were published (see, for example, the reviews of [Zanardelli et al., 2022] and [Verdoliva, 2020] and references therein) in order to ensure visual information authenticity.

While there are a number of established benchmark datasets in the field of image forgery detection [Dong et al., 2013, Hsu and Chang, 2006, Carvalho et al., 2013, National Institute of Standards and Technology (NIST),

2016], proposed datasets are limited in size and manipulation diversity, and are therefore not appropriate as training datasets. Table 1, partly taken from [Novozamsky et al., 2020], gives an overview of available datasets designed for image forgery detection. Proposed datasets from literature which are not accessible anymore were removed from the table. As shown, the tampCoco [Kwon et al., 2022] dataset and the Defacto [Mahfoudi et al.,

<b>Dataset of manipulated images</b>	<b>Size</b>	<b>Manip-Types</b>
Coverage [Wen et al., 2016]	100	C
CoMoFoD [Tralic et al., 2013]	260	C
DSO [Carvalho et al., 2013]	100	SE
Columbia [Hsu and Chang, 2006]	160	S
CASIA [Dong et al., 2013]	920	SCE
CASIA v2.0 [Dong et al., 2013]	5,123	SCE
MICC-F220, MICC-F2000 [Amerini et al., 2011]	2,200	C
Zhou et al. [Zhou et al., 2017]	3,410	SE
NIST16 [National Institute of Standards and Technology (NIST), 2016]	564	SCR
OpenMFC20_Image_MD [Guan et al., 2019]	16,075	SCR
OpenMFC22_SpliceImage_MD [Guan et al., 2019]	2,000	S
IMD2020 Manually Created [Novozamsky et al., 2020]	2,010	SCRE
IMD2020 [Novozamsky et al., 2020]	35,000	R
Defacto [Mahfoudi et al., 2019]	189,387	SCR
tampCoco [Kwon et al., 2022]	800,000	SC
<b>DF2023 Training (proposed)</b>	1,000,000	SCRE
<b>DF2023 Validation (proposed)</b>	5,000	SCRE

Table 1: Examples of datasets designed for image manipulation detection with number of tampered images and manipulation types: (S)plicing, (C)opy-Move, (R)evelopment, (E)nhancement

2019] dataset are by far the largest available datasets. The Defacto [Mahfoudi et al., 2019] dataset has about 190,000 images. However, 39,800 images of this dataset are very specific face morphing forgeries. The forgery type enhancement, on the other side, was not specifically included in the dataset. The tampCoco dataset has just been released on Kaggle on March 28, 2023. The dataset is derived from the MS-COCO dataset [Lin et al., 2014] and was generated by applying the manipulation techniques of splicing and copy-move operations.

Considering the typical volume of training data required for deep neural networks to tackle complex tasks, the overview provided in Table 1 highlights the necessity for a sufficiently large training dataset that encompasses a diverse range of manipulations.

### 3 Digital Forensics Dataset - DF2023

The benefits of a large, diverse and public training dataset for detection of image forgeries are manifold: Researchers can save significant time by avoiding data collection, scripting and data generation. Using a pre-existing dataset prevents from consciously or unconsciously adjusting the training dataset to become too similar to the evaluation sets. Most importantly, such a dataset allows the decoupled evaluation and comparison of deep learning network architectures in an objective, transparent and (rather) reproducible way. For this reason, we introduce the Digital Forensics 2023 (DF2023) dataset, available from here: DF2023. The DF2023 training dataset contains one million forged images of the four main manipulation types. Specifically, the training dataset consists of 100K forged images produced by removal operations, 200K images produced by various enhancement modifications, 300K copy-move manipulated images and 400K spliced images. This distribution was selected based on our experience regarding the positive impact of each manipulation type on improving forgery detectors. The MS-COCO [Lin et al., 2014] 2017 training and validation datasets with 118K/5K images were facilitated as the source of pristine and donor images. Many other publicly available datasets in this research domain often lack comprehensive documentation, we on the other hand have chosen to provide a detailed description in the following sections on the meticulous process of creating the DF2023 dataset.

### 3.1 DF2023 - Dataset generation

#### 1. Selection of pristine image:

A pristine image  $\mathcal{I}_P$  was randomly selected from the MS-COCO 2017 training dataset, and respectively from the validation dataset. For the few images with width  $W$  or height  $H$  smaller than 256 pixels, the image was resized to the size  $(\max(W, 256), \max(H, 256))$ . For 50% of the images  $\mathcal{I}_P$  in the training dataset, a proportion-preserving downscale was executed. This avoided extracting only small portions of larger images (like a monochrome patch depicting a part of the sky from the original image). The scaling of an image  $\mathcal{I}_P$  with size  $(W, H)$  to  $(W_{new}, H_{new})$  was done as follows:

$$\begin{aligned} W_{new} &= \max(\lfloor (\frac{256 \cdot W}{\min(W, H)}) \rfloor, 256) \\ H_{new} &= \max(\lfloor (\frac{256 \cdot H}{\min(W, H)}) \rfloor, 256) \\ \mathcal{I}_P &= \mathcal{I}_P.\text{resize}((W_{new}, H_{new})) \end{aligned} \quad (1)$$

Next, a patch of size (256, 256) pixels was randomly chosen from the image  $\mathcal{I}_P$  and used as a pristine image patch  $\mathcal{P}$ .

#### 2. Selection of donor image:

A donor image  $\mathcal{I}_D$  from MS-COCO (training/validation) was selected. For the splicing operation, a random image other than the pristine image  $\mathcal{I}_P$  was selected. For the copy-move, removal and enhancement manipulations, the same pristine image was selected as a donor image ( $\mathcal{I}_D = \mathcal{I}_P$ ).

#### 3. Pre-processing of donor image:

Table 2 shows which preprocessing steps may be applied to the donor image  $\mathcal{I}_D$  for each manipulation type. **Resample** rescaled the height and the width image dimensions independently by 70 to 130 percent. The size of the resulting image is at least (256, 256). The preprocessing step **Flip** flipped the donor image horizontally with a likelihood of 50%, while **Rotate** rotated the image by either 90, 180 or 270 degrees with a likelihood factor controlled by a predefined parameter (for the DF2023 dataset, 30% of the donor images were rotated). **Blur** blurred the donor image with a likelihood of 50%. In case the blurring filter was applied, either `ImageFilter.BoxBlur` or `ImageFilter.GaussianBlur` from the Python package `PIL` were used, both with equal probabilities. The blur radius was set randomly between 1 and 7 pixels. **Contrast** used one of the `ImageFilters` `EDGE_ENHANCE`, `EDGE_ENHANCE_MORE`, `SHARPEN`, `UnsharpMask` or `ImageEnhance.Contrast` from the Python package `PIL`. **Noise** added Gaussian noise with mean and standard deviation  $(\mu, \sigma) = (0, 12)$  with a likelihood of 1 out of 3. The **Brightness** was changed with a probability of 50% by a factor uniformly chosen in the range [0.5-1.5]. With 50% probability, a **JPEG-Compression** with quality factor  $10x$  for  $x \in [1, 2, 3, 4, 5, 6, 7]$  was employed. If the manipulation type was **Removal**, an inpainting filter from OpenCV [Bradski, 2000] was applied (either `cv2.INPAINT_TELEA` or `cv2.INPAINT_NS`) on the manipulation mask defined in step 5.

In case the chosen manipulation type was **Enhance** and none of the filters (blur, contrast, noise, brightness, JPEG compression) were applied to the donor image  $\mathcal{I}_D$ , the process was repeated.

Manipulation	C	S	R	E	Pos.	values	e.g.
Resample	x	x	-	-	1	0/1	0
Flip	x	x	-	-	2	0/1	0
Rotate	x	x	-	-	3	0/1/2/3	0
Blur	-	-	-	x	4-5	B/G, 0-9	G4
Contrast	-	-	-	x	6	0-5	0
Noise	-	-	-	x	7	0/1	1
Brightness	-	-	-	x	8	0/1	1
JPEG-Compression	-	-	-	x	9	0-9	7

Table 2: Preprocessing steps for donor image per manipulation types: Copy-Move (**C**), Splicing (**S**), Removal (**R**) and Enhancement (**E**). The column **Pos.** indicates the filenames encoding position for the corresponding manipulation (starting to count at the position for the manipulation type) as explained in Section 3.2. Column **values** and column **e.g.** show possible values and an example value, respectively, for the position in the name. For this example, a filename could be: COCO\_DF\_E000G40117\_00200620.jpg

#### 4. Cropping of donor patch:

Then, a donor patch  $\mathcal{D}$  of size (256,256) was randomly cropped from  $\mathcal{I}_D$ . For enhancement and removal (inpainting) manipulations, the donor patch  $\mathcal{D}$  and the pristine patch  $\mathcal{P}$  share the same location in  $\mathcal{I}_D = \mathcal{I}_P$ .

#### 5. Creation of a binary manipulation mask:

Seven types of binary masks  $\mathcal{M}$  were used to define the image region where manipulations were executed (see Table 3). In Table 4, various examples of the masks created and the resulting forged images are shown. Despite the five mask types which are based on geometric forms, we used Python's image processing toolbox scikit-image to segment the donor patch into Superpixels [Achanta et al., 2012] of appropriate size, and selected one Superpixel (connected set of pixels) as the defining mask where the manipulations would be applied on. Furthermore, the "object segmentation" used the segmentation ground truth from the MS-COCO dataset. All pixels from a donor image patch  $\mathcal{D}$  which were marked corresponding to a specific object class (e.g. person) were selected and used as splicing input. The object category was randomly selected from the possible categories of the donor image, hence MS-COCO images with no labeled objects were excluded in case of object based mask creation.

Shape of Mask	Parameters	Impact
Triangle	p1, p2, p3	3 random points
Rounded Rectangle	X, Y, r	2 points for Bbox; radius of the corners
Ellipse	X, Y	2 points to define the bounding box
Polygon with 5 vertices	p1,...,p5	sequence of 5 random points
Ellipse + Polygon with 4 vertices	X, Y, p1,...,p4	ellipse + 4 vertex polygon
Superpixel Segmentation	[min, max]	range for number of Superpixels per image
Object Segmentation	obj. category	object category for segmentation (e.g. person)

Table 3: Types of mask shapes generated for local image manipulation

#### 6. Creation of a non-binary manipulation mask:

For a smooth gradient at the edges of manipulation and as preparation for alpha blending, the manipulation masks  $\mathcal{M}$  were blurred half of the time for splicing, copy-move and enhancement operations. This way, the transition from pristine image to manipulated patch is smooth and the forgery detection network is forced not to solely rely on sharp edges for identifying manipulated regions. Additionally, we applied alpha blending to make splicing manipulations more realistic and harder to detect. We achieved this by randomly setting an alpha value in the range [0.94, 1.0] and multiplying the manipulation mask with this floating point scalar value.

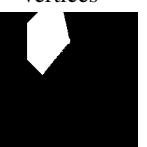
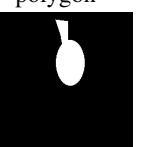
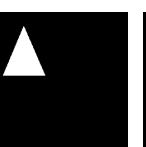
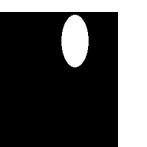
Forgery Type	Copy-Move	Splicing	Removal	Enhance	Removal	Enhance	Copy-Move
							
Shape of Forgery Mask	superpixel segmentat.	object segmen-	polygon 5 vertices	ellipse + 4V polygon	triangle	rounded rectangle	ellipse
							

Table 4: Examples from the Digital Forensics 2023 (DF2023) dataset: The upper row shows the forged images and the applied manipulation type. The second row shows the corresponding manipulation mask and its shape.

Considerably stronger alpha blending led to worse results in our experiments. Finally, masks were recalculated if their size was below 5% or above 40% of the image patch.

#### 7. Generation of forged image:

Given a pristine patch  $\mathcal{P}$ , a donor patch  $\mathcal{D}$ , a manipulation  $m$  and a binary manipulation mask  $\mathcal{M}$ , the forged image  $\mathcal{X}$  is represented as

$$\mathcal{X} = (1 - \mathcal{M}) \cdot \mathcal{P} + \mathcal{M} \cdot m(\mathcal{D}) \quad (2)$$

meaning that each pixel of the resulting image  $\mathcal{X}$  is taken either from the pristine patch  $\mathcal{P}$  or the manipulated donor patch  $\mathcal{D}$ , depending on the binary mask  $\mathcal{M}$ . For alpha blending with a non-binary mask  $\mathcal{M}$ , the formula is still valid and combines pixels from the pristine and the donor image according to the mask values in the range  $[0, 1]$ . In case of a copy-move manipulation, an additional translation of the copied image part  $(1 - \mathcal{M}) \cdot m(\mathcal{D})$  is made towards another position in the pristine image patch.

#### 8. Generation of ground truth:

Ground truth masks  $\mathcal{M}_{GT}$  of (non-binary) manipulation masks  $\mathcal{M}$ , are defined as binary masks counting each non-zero value as 1, or as a Boolean matrix in NumPy notation:

$$\mathcal{M}_{GT} = (\mathcal{M} > 0) \quad (3)$$

### 3.2 DF2023 - Naming convention

The naming convention of DF2023 encodes information about the applied manipulations. The following convention is used for the image names:

COCO\_DF\_0123456789\_NNNNNNNN.{EXT}

For example:

COCO\_DF\_E000G40117\_00200620.jpg

After the identifier of the image data source ("COCO") and the self-reference to the Digital Forensics ("DF") dataset, there are 10 digits as placeholders for the manipulation. Position 0 defines the manipulation types copy-move, splicing, removal, enhancement ([C,S,R,E]). The following digits 1-9 represent donor patch manipulations according to column *Pos.* in Table 2. For positions [1,2,7,8] (resample, flip, noise and brightness),

a binary value indicates if this manipulation was applied to the donor image patch. In Position 3 (rotate) the values 0-3 indicate if the rotation was executed by 0, 90, 180 or 270 degrees. Position 4 defines if BoxBlur (B) or GaussianBlur (G) was used. Position 5 specifies the blurring radius. A value of 0 indicates that no blurring was executed. Position 6 indicates which one of the Python-PIL contrast filters EDGE\_ENHANCE, EDGE\_ENHANCE\_MORE, SHARPEN, UnsharpMask or ImageEnhance (values 1-5) was applied. If none of them was applied, this value is set to 0. Finally, position 9 is set to the JPEG compression factor modulo 10, where a value of 0 indicates that no JPEG compression was applied. The 8 characters NNNNNNNN in the image name template stand for a running number of the images.

## 4 Experimental results

For experimental results and in-depth evaluation, we refer to the publication [Fischinger and Boyer, 2023]. Here, the authors explain how using a simple network trained on the DF2023 dataset has led to state-of-the-art results in the area of image forgery detection.

## 5 Conclusion

This paper addresses the existing gap in the research area of image forgery detection and localization by providing a comprehensive and publicly accessible training dataset that encompasses a wide array of image manipulation types: We present the Digital Forensics 2023 (DF2023) dataset for training and validation (available from <https://zenodo.org/record/7326540>), comprised of more than one million images with diverse manipulations. We firmly believe that the availability of this dataset will not only save researchers valuable time but also facilitate easier and more transparent comparisons of network architectures.

## Acknowledgement



Co-funded by  
the European Union

Project 101083573 — GADMO

## References

- [Achanta et al., 2012] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282.
- [Amerini et al., 2011] Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., and Serra, G. (2011). A sift-based forensic method for copy-move attack detection and transformation recovery. *IEEE transactions on information forensics and security*, 6(3):1099–1110.
- [Bradski, 2000] Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123.
- [Carvalho et al., 2013] Carvalho, T., Riess, C., Angelopoulou, E., Pedrini, H., and Rocha, A. R. (2013). Exposing digital image forgeries by illumination color classification. *IEEE Trans. Inf. Forensics and Security*, 8(7):1182–1194.
- [Dong et al., 2013] Dong, J., Wang, W., and Tan, T. (2013). Casia image tampering detection evaluation database. In *IEEE China Summit Inter. Conf. Signal Info. Proc.*, pages 422–426. IEEE.
- [Fischinger and Boyer, 2023] Fischinger, D. and Boyer, M. (2023). DF-Net: The digital forensics network for image forgery detection. *Irish Machine Vision and Image Processing conference*.

- [Guan et al., 2019] Guan, H., Kozak, M., Robertson, E., Lee, Y., Yates, A. N., Delgado, A., Zhou, D., Kheykhah, T., Smith, J., and Fiscus, J. (2019). Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE.
- [Hsu and Chang, 2006] Hsu, Y. and Chang, S. (2006). Detecting image splicing using geometry invariants and camera characteristics consistency. In *IEEE Inter. Conf. Multim. Expo*, pages 549–552. IEEE.
- [Kwon et al., 2022] Kwon, M.-J., Nam, S.-H., Yu, I.-J., Lee, H.-K., and Kim, C. (2022). Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130:1875 – 1895.
- [Li et al., 2017] Li, H., Luo, W., and Huang, J. (2017). Localization of diffusion-based inpainting in digital images. *IEEE Transactions on Information Forensics and Security*, 12(12):3050–3064.
- [Li et al., 2013] Li, L., Li, S., Zhu, H., Chu, S.-C., Roddick, J., and Pan, J.-S. (2013). An efficient scheme for detecting copy-move forged images by local binary patterns. *Journal of Information Hiding and Multimedia Signal Processing*, 4:46–56.
- [Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, L. (2014). Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision.
- [Lyu et al., 2013] Lyu, S., Pan, X., and Zhang, X. (2013). Exposing region splicing forgeries with blind local noise estimation. *International Journal of Computer Vision*, 110:202–221.
- [Mahfoudi et al., 2019] Mahfoudi, G., Tajini, B., RETRAINT, F., Morain-Nicolier, F., Dugelay, J.-L., and Pic, M. (2019). Defacto: Image and face manipulation dataset. pages 1–5.
- [National Institute of Standards and Technology (NIST), 2016] National Institute of Standards and Technology (NIST) (2016). Nist nimble 2016 datasets. <https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation/>.
- [Novozamsky et al., 2020] Novozamsky, A., Mahdian, B., and Saic, S. (2020). Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 71–80.
- [Sun et al., 2018] Sun, J.-Y., Kim, S.-W., Lee, S.-W., and Ko, S.-J. (2018). A novel contrast enhancement forensics based on convolutional neural networks. *Signal Processing: Image Communication*, 63:149–160.
- [Tralic et al., 2013] Tralic, D., Zupancic, I., Grgic, S., and Grgic, M. (2013). Comofod—new database for copy-move forgery detection. In *Proceedings ELMAR-2013*, pages 49–54. IEEE.
- [Verdoliva, 2020] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932.
- [Wen et al., 2016] Wen, B., Zhu, Y., Subramanian, R., Ng, T.-T., Shen, X., and Winkler, S. (2016). Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE.
- [Wu et al., 2022] Wu, H., Zhou, J., Tian, J., Liu, J., and Qiao, Y. (2022). Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*.
- [Wu et al., 2019] Wu, Y., AbdAlmageed, W., and Natarajan, P. (2019). Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9535–9544.

[Zanardelli et al., 2022] Zanardelli, M., Guerrini, F., Leonardi, R., and Adami, N. (2022). Image forgery detection: a survey of recent deep-learning approaches. *Multimedia Tools and Applications*, 82:17521–17566.

[Zhou et al., 2017] Zhou, P., Han, X., Morariu, V. I., and Davis, L. S. (2017). Two-stream neural networks for tampered face detection. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1831–1839. IEEE.

# Development of a Classification-based Eye Gaze estimation technique using an Integrated Laptop Camera: two models are better than one

Jack Cribbin, Charles Markham  
*National University of Ireland, Maynooth*

## Abstract

While there are many purely software-based solutions available for live head pose tracking, the same is not true for gaze estimation which is usually done using specialised hardware; typically, eye-trackers that use infrared light sources and special camera. The challenge when using vision-based Machine Learning methods to estimate gaze from an image of a user is that the features of a user's eyes that vary with gaze position are small and difficult to track. This paper describes a classification-based Convolutional Neural Network (CNN) system that can, from images of a user's face, estimate where on a screen they are looking in real time. Labelled images of an individual looking at specified regions of a screen were collected, and these images were used to train a Deep Learning model. The final model trained was found to be accurate enough to correctly guess which section of the screen (3x3 grid) the user is looking in over 99% of the time. The current system requires strict input conditions on head orientation, a limitation on the system. The integration of head pose estimation and methods of improving resolution are discussed. The approach uses two classifiers, one for horizontal gaze (3 regions) and another for vertical gaze (3 regions) combined to estimate gaze location. This approach was found to be more effective than a single classifier for nine separate regions. Using both eyes, excluding the nose region and other facial features of the input images also improved performance.

**Keywords:** Gaze Estimation, Machine Learning, Machine Vision, Convolutional Neural Networks

## 1 Introduction

Currently, live and accurate gaze estimation is usually achieved using specialist hardware that utilizes infrared light sources and cameras to track the position of a user's corneas [Guestrin et al. 2006]. Gaze estimation has many applications, including computer vision [Krafka et al. 2016], human-computer interfacing [Jacob et al. 2013], marketing research [Wedel et al. 2008] and medical diagnosis [Holzman et al. 1974]. There is interest in the development of a gaze estimation system that is inexpensive, easy to use and does not require a high level of setup.

The upfront cost involved in purchasing this hardware, as well as the requirement to calibrate it for each user, makes it a barrier for those who wish to use it in some research or commercial applications. The less expensive face tracking alternatives that include eye tracking make use of advanced cameras, such as that of iOS Face ID. This camera utilises an infrared projector and infrared camera [Apple 2017].

In this work, Machine Learning was used to train a Convolutional Neural Network (CNN) to label the region of a screen that a user is looking at. This is done using an image of their face collected on a standard laptop webcam. CNN's are well-suited for the task of image classification. This approach requires consistent features that can be identified, tracked and correlate with the section of the screen the user is looking at. Using this approach, a purely software-based method for gaze estimation was developed that does not require structured light sources. This reduces the hardware barrier for users wishing to do gaze estimation to a PC to run the software and a webcam of sufficient resolution. An application using the gaze estimator was developed to demonstrate its utility.

Current gaze estimation software requires personal calibration by the user to accurately map their gaze. While the CNN model developed in this paper requires individual datasets to calibrate for each user, a feature of these neural networks is that, with a broad enough dataset, they can be trained to recognize features that are generalized. This means that it may be possible to make a model that requires no calibration for low resolution operation. This would require a larger training dataset consisted of a large enough group of subjects to ensure that the model does not rely on features present only in some subjects, and instead uses features that are present in most subjects.

## 2 State of the Art

Typically, gaze estimation is done using structured light sources (usually infrared light sources) which allow an infrared camera to triangulate the location of a user's cornea in a scene, and from the portion of the cornea that is visible estimate where on the screen the user is looking. Usually this is done with multiple infrared sources, such as with the iOS Face ID [Apple 2017]. Commercial eye trackers also make use of specialist equipment [Tobii] [Gazepoint] which can be mounted alongside a screen to estimate eye gaze location.

While purely software-based gaze estimators that only make use of visible light cameras have been developed, in most cases this has been attempted using Support Vector Machines to track the movement of eye features [Papoutsaki et al. 2016] [Quirós et al. 2014] [Wu et al. 2012]. This approach is limited by camera resolution and lighting conditions as the minute features of the eye and their changes over time can be difficult to track accurately outside of ideal conditions. Typically, their accuracy is lacking compared to hardware-based solutions, with typical accuracy values ranging from 65.3% to 73.6% [Quirós et al. 2014] or mean error of ~100 pixels [Papoutsaki et al. 2016], with licensed software having an accuracy of 1-1.5 degrees under ideal conditions [GazeSense SDK] [GazeRecorder]. This is opposed to a mean error of ~30 pixels for typical hardware-based solutions [Tobii].

In recent years, attempts have been made to utilize Convolutional Neural Networks to classify images of a user's face in relation to where on a screen they are looking [Akinyelu et al. 2022] [Palmero et al. 2018]. As these networks are well suited to the problems of machine vision and image classification, there are high expectations for their applications in gaze estimation. In practice, most implementations do not control for head orientation and instead try to make their models head orientation invariant. Most CNN models are trained using either full-face images [Akinyelu et al. 2022] or at least a band of the image spanning from one eye to the other, including the middle of the user's face [Palmero et al. 2018]. Additionally, these tasks are usually approached as a regression problem instead of a classification problem, assigning continuous coordinate values to images instead of classifying it to a discreet location.

This paper outlines a novel, classification-based approach to the problem of gaze estimation. The paper also explores improvements that can be made to the techniques currently in use by examining the feature maps the models produce. Preprocessing the image, by extraction of eye regions, as well as utilizing multiple models was shown to improve the overall accuracy

## 3 System Development

The models trained were developed using the TensorFlow Python module [TensorFlow] using a dataset collected over the course of development. The final system overview is outlined in Figure 1.

Training code was developed to allow the collection of a dataset of labelled images of the user looking at regions of the screen. The user was asked to gaze at highlighted horizontal or vertical regions of the screen displayed using the PyGame module [PyGame] and hold down spacebar for the duration. For each frame of video recorded, a series of Haar Cascade Classifiers from the OpenCV library [OpenCV] were used to identify the regions of the image

likely to contain an eye. These regions were then extracted and combined to form a single image. This image contained images of the left and right eye stitched together. These steps were then repeated until enough labelled images were collected for each region of the screen.

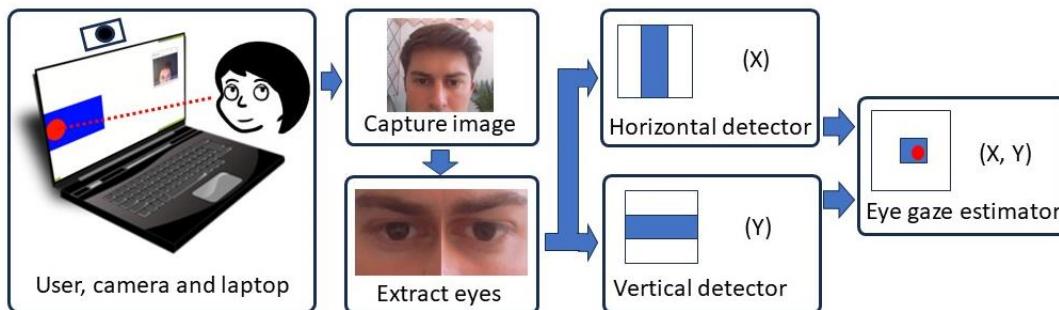


Figure 1: System Overview

These images were then used to train a Convolutional Neural Network for horizontal and vertical classification using the TensorFlow Keras library [[Tensorflow](#)]. Training was run until a sufficient accuracy rate and error rate was achieved. This was done by separating a subset of the image dataset into a validation set on which the models' accuracy and error rate was tested.

Finally, a live video feed was taken from the webcam, and for each frame read in the eye sections were isolated from the image as done during the dataset collection. Using inference, the horizontal and vertical models were then used to make a prediction of the label assigned to these new images. The models' best guess at the gaze location in the 3x3 grid is then displayed on the screen in real time, with guesses being made every ~0.5 seconds.

Most CNN techniques estimate a users' gaze using a single CNN model [[Akinyelu et al. 2022](#)] [[Palmero et al. 2018](#)]. This regressively computes a function to assign a continuous x and y value of gaze location, given an image of a user's face. This paper took a different approach, in that the screen was split into discrete segments for the user to look within. The models produced then classify images of the user's face with a discrete value for each segment. The model structure and techniques used mirrored those commonly used in previous papers, however changes were made that were found to significantly improve the accuracy of the models and the performance of the system overall.

The first design change made was to pre-emptively remove unnecessary sections of training images on which models were incorrectly identifying features. This was done after the feature maps generated by the models were examined. It was found that many of the misclassified images had features identified on the user's face or outside their eyes, effectively training on noise. To reduce this issue, the training images were cropped to only include a band spanning from one eye to the other. When it was found that the model was still identifying features around the user's nose, the middle section between the eyes was then cut out to reduce the input training images to just the two eye sections stitched together, see Figure 2.



Figure 2: Example of the input image and feature maps produced by 3 different models. (a) input image, (b) activation map generated by a single 3x3 model. (c) activation map generated by 1x3 model focused on determining vertical position. (d) activation map generated by 3x1 model focused on determining horizontal position. White pixels represent pixels of significance. Note for (c) and (d) the areas of significance around the sclera of the pupil, and the lack of weighting on other, irrelevant features.

The second change made was moving to a multi-model system for classifying the images instead of using a single model. During development, it was found that the feature maps for input images had trouble identifying both features that help discern the horizontal position and the vertical position of the user's gaze at the same time. To reduce this issue, the tasks were split between two models. One model focused entirely on discerning the horizontal gaze location and the other on the vertical location. This was found to improve the overall accuracy of the system from an accuracy of between 65% and 75% to between 99.3% and 100%. While this had the consequence of nearly doubling the training time for each system. Since the time needed for a model to produce a guess for an input image once it has been trained is approximately 0.04 - 0.05 seconds, doubling this prediction time had no noticeable effect for the end user of the software.



Figure 3: A diagram illustrating how the combination of a  $3 \times 1$  and a  $1 \times 3$  model is equivalent to a single  $3 \times 3$  model.

Finally, the software was adapted into an application to demonstrate the feasibility of the tool being developed into an API. The application allowed users to make a hot beverage using only their gaze (choice) and a single key (accept). The purpose of this application was to highlight how a gaze estimator could be used, especially by those with limited mobility, to interact with computers. Another simple application was made that displayed images of two similar products and tracked which product the user looked at (if either) in each frame. The purpose of this code was to demonstrate one possible commercial application of the gaze estimation software in product design and marketing, see Figure 4.

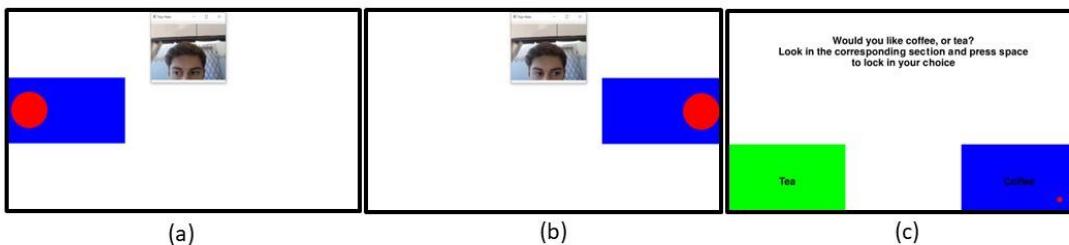


Figure 4: (a), (b) eye tracking demonstration. The blue segment represents the best guess of gaze region on screen. The location of the red dot (eye gaze) is found using the models' confidence levels for each section, to show a more accurate location than the system's actual resolution ( $3 \times 3$ ). (c) shows application demonstration.

All software developed can be found at: <https://github.com/JackCribbin/Gaze-Estimation-CNN>

## 4 Results

The accuracy of models developed was validated using a dataset of images separate to the ones used to train the models. The models were then live tested, however data was not collected for this testing as it was simply to validate the general accuracy of the models and check for possible mislabelling of training images. Figure 5 shows a graph of accuracy and loss for the horizontal model developed. As can be seen, accuracy rises sharply before plateauing at a value between 99.3% and 100%, whereas loss drops sharply and plateaus at a value between 1.6% and 1.8%.

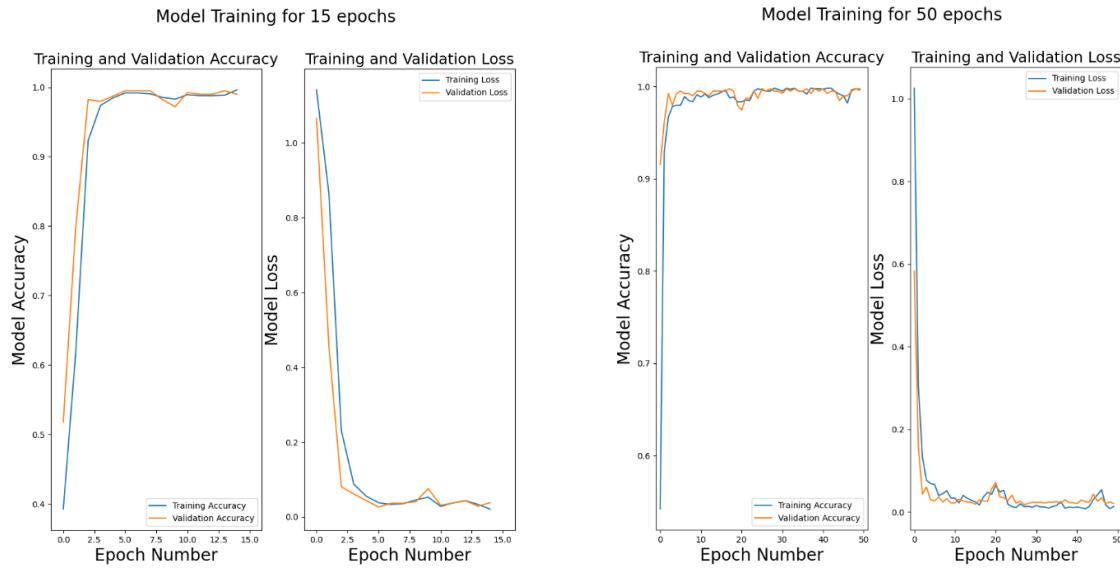


Figure 5: Plots of Model Accuracy and Loss for two example model trainings, one with 15 epochs (left) and one with 50 epochs (right). Model training was stopped when accuracy plateaued to avoid overtraining, which would make the models less elastic and worsen performance on new, unseen images.

This accuracy is limited to a resolution of 1/9<sup>th</sup> of the screen as the screen was split into 3x3 segments while collecting the datasets. Thus, the gaze estimator developed has an accuracy of over 99% within a region of 9.9 cm by 5.6 cm on a 29.7 cm by 16.7 cm screen at a typical viewing distance of approximately 40 cm.

However, this resolution is not necessarily an upper limit, it is simply the highest resolution tested by this paper. Further work may increase the dual models to two 4x1 models to reduce the segment size to 1/16<sup>th</sup> of the screen. Additionally, by making use of classification confidence levels, a specific point of gaze can be estimated within or between segments. This is done by using the models' confidence levels for multiple segments and multiplying these by the distance from the centre of the screen to show a more accurate guess than the system's actual resolution. See the red dot on Figure 4(a).

## 5 Future Work

Current work aims to train new models to improve the resolution and reduce the size of the segments the models use to classify images.

Additionally, integrating data on the user's head position and orientation during training for each image may reduce the limitations on the user to hold a fixed pose while the software is being used. This could be done by using a Support Vector Machine-based head tracker to produce pitch, yaw and roll data which would be included in the training data alongside the images.

Finally, additional datasets could be collected from various individuals to test the efficacy of a generalised gaze estimator that does not require individual calibration.

## 6 Conclusions

This paper demonstrates a method to estimate a user's gaze using only an integrated laptop camera and classification-based Convolutional Neural Network models. Adjustments were made to standard practices currently implemented to improve performance, such as approaching the task as a classification problem, limiting input data to solely relevant regions and making use of two models instead of one. The models developed serve as a proof of concept for the viability of developing a calibration-necessary, software-based gaze estimator, with single user accuracy above 99% within a resolution of 1/9<sup>th</sup> of the screen achieved after calibration. Future work with larger and more varied datasets from multiple users will show if the features identified for individuals can be generalised to produce a model that does not require calibration for each user.

## References:

- [Guestrin and Eizenman, 2006] Guestrin, E.D. and Eizenman, M., 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6), pp.1124-1133. Available at: <https://vemlab.github.io/127.0.0.1/static/attachment/36.pdf>
- [Krafcik et al., 2016] Krafcik, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W. and Torralba, A., 2016. Eye tracking for everyone. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2176-2184). Available at: [https://www.cvfoundation.org/openaccess/content\\_cvpr\\_2016/html/Krafcik\\_Eye\\_Tracking\\_for\\_CVPR\\_2016\\_paper.html](https://www.cvfoundation.org/openaccess/content_cvpr_2016/html/Krafcik_Eye_Tracking_for_CVPR_2016_paper.html)
- [Jacob and Karn, 2013] Jacob, R.J. and Karn, K.S., 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (pp. 573-605). North-Holland. Available at: <https://www.sciencedirect.com/science/article/abs/pii/B9780444510204500311?via%3Dhub>
- [Wedel and Pieters, 2008] Wedel, M. and Pieters, R., 2008. Eye tracking for visual marketing. *Foundations and Trends® in Marketing*, 1(4), pp.231-320. Available at: <https://www.nowpublishers.com/article/Details/MKT-011>
- [Holzman et al., 1974] Holzman, P.S., Proctor, L.R., Levy, D.L., Yasillo, N.J., Meltzer, H.Y. and Hurt, S.W., 1974. Eye-tracking dysfunctions in schizophrenic patients and their relatives. *Archives of general psychiatry*, 31(2), pp.143-151. Available at: <https://pubmed.ncbi.nlm.nih.gov/4851993/>
- [Apple 2017] Apple Inc. (2017) Apple awards Finisar \$390 million from its Advanced Manufacturing Fund, 13 December, 2017. Available at: <https://www.apple.com/newsroom/2017/12/apple-awards-finisar-390-million-from-its-advanced-manufacturing-fund/>
- [Tobii] Tobii, <https://www.tobii.com/products/eye-trackers>, accessed on 12/06/2023
- [GazePoint] GazePoint, <https://www.gazept.com/shop>, accessed on 12/06/2023[Quirós and Cristina, 2014] Cabrera Quirós, L.C., 2014. Eye gaze estimation using SVM for regression from face and eye detection results (Doctoral dissertation). Available at: <http://repositorio.conicit.go.cr:8080/xmlui/handle/123456789/88>
- [Wu et al., 2012] Wu, Y.L., Yeh, C.T., Hung, W.C. and Tang, C.Y., 2014. Gaze direction estimation using support vector machine with active appearance model. *Multimedia tools and applications*, 70, pp.2037-2062. Available at: [https://www.researchgate.net/publication/257627330\\_Gaze\\_direction\\_estimation\\_using\\_support\\_vector\\_machine\\_with\\_active\\_appearance\\_model](https://www.researchgate.net/publication/257627330_Gaze_direction_estimation_using_support_vector_machine_with_active_appearance_model)
- [Papoutsaki et al. 2016] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J. and Hays, J. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. 25<sup>th</sup> International Joint Conference on Artificial Intelligence. Available at:

[https://www.researchgate.net/publication/305375498\\_WebGazer\\_Scalable\\_Webcam\\_Eye\\_Tracking\\_Using\\_User\\_Interactions/references](https://www.researchgate.net/publication/305375498_WebGazer_Scalable_Webcam_Eye_Tracking_Using_User_Interactions/references)

[GazeSense SDK] GazeSense SDK, eyeware, <https://eyeware.tech/gazesense-eye-tracking-software-for-webcams-3d-sensor>, accessed on 12/06/2023

[GazeRecorder] GazeRecorder, <https://gazerecorder.com/webcam-eye-tracking-accuracy>, accessed on 12/06/2023

[Akinyelu and Blignaut, 2022] Akinyelu, A.A. and Blignaut, P., 2022. Convolutional Neural Network-Based Technique for Gaze Estimation on Mobile Devices. *Frontiers in Artificial Intelligence*, 4, p.796825. Available at: <https://www.frontiersin.org/articles/10.3389/frai.2021.796825/full>

[Palmero et al., 2018] Palmero, C., Selva, J., Bagheri, M.A. and Escalera, S., 2018. Recurrent cnn for 3d gaze estimation using appearance and shape cues. arXiv preprint arXiv:1805.03064. Available at: <http://www.bmva.org/bmvc/2018/contents/papers/0871.pdf>

[TensorFlow] TensorFlow, <https://tensorflow.org>, accessed on 12/06/2023

[OpenCV] OpenCV, <https://opencv.org>, accessed on 12/06/2023

[Pygame] Pygame, <https://pygame.org>, accessed on 12/06/2023

# A lightweight 3D dense facial landmark estimation model from position map data

Shubhajit Basak<sup>1,2,\*</sup>, Sathish Mangapuram<sup>2</sup>, Gabriel Costache<sup>2</sup>, Rachel McDonnell<sup>3</sup>, and Michael Schukat<sup>1</sup>

<sup>1</sup>*School Of Computer Science, University of Galway*

<sup>2</sup>*Xperi Corporation*

<sup>3</sup>*School of Computer Science and Statistics, Trinity College Dublin*

## Abstract

The incorporation of 3D data in facial analysis tasks has gained popularity in recent years. Though it provides a more accurate and detailed representation of the human face, accruing 3D face data is more complex and expensive than 2D face images. Either one has to rely on expensive 3D scanners or depth sensors which are prone to noise. An alternative option is the reconstruction of 3D faces from uncalibrated 2D images in an unsupervised way without any ground truth 3D data. However, such approaches are computationally expensive and the learned model size is not suitable for mobile or other edge device applications. Predicting dense 3D landmarks over the whole face can overcome this issue. As there is no public dataset available containing dense landmarks, we propose a pipeline to create a dense keypoint training dataset containing 520 key points across the whole face from an existing facial position map data. We train a lightweight MobileNet-based regressor model with the generated data. As we do not have access to any evaluation dataset with dense landmarks in it we evaluate our model against the 68 keypoint detection task. Experimental results show that our trained model outperforms many of the existing methods in spite of its lower model size and minimal computational cost. Also, the qualitative evaluation shows the efficiency of our trained models in extreme head pose angles as well as other facial variations and occlusions. Code is available at: <https://github.com/shubhajitbasak/dense3DFaceLandmarks>

**Keywords:** 3D Facial Landmarks, Position Map, Dense Landmarks

## 1 Introduction

Predicting the 3D features of the human face is the pre-requisite for many facial analysis tasks such as face reenactment and speech-driven animation, video dubbing, projection mapping, face replacement, facial animations, and many others [Zollhöfer et al., 2018]. Due to the limitation of depth sensors, it is difficult to capture high-frequency details through RGB-D data. Capturing high-quality 3D scans is expensive and often restricted because of ethical and privacy concerns. A popular alternative to these facial capturing methods is to estimate the face geometry from an uncalibrated 2D face image. However, this 3D-from-2D reconstruction of the human face is an ill-posed problem because of the complexity and variations of the human face. With the help of highly complex deep neural network models, we are able to recover the detailed face shape from uncalibrated face images. However, most of these methods depend on some kind of statistical priors of face shape like a 3D morphable model(3DMM) and the sparse face landmarks for face alignments. Some of the previous works also used additional signals beyond color images, like facial depth [Bao et al., 2021], optical flow [Cao et al., 2018], or multi-view stereo [Beeler et al., 2010], and then optimized their methods using geometric and temporal prior. Each of these methods can produce very detailed results, but take a very long time to compute. At the same time, the model size and the huge computational requirements make these approaches not suitable for real-time

applications in edge devices. Therefore it is still a very challenging task to implement a face modeling pipeline on limited computational cost systems such as mobile or embedded devices.

Estimating dense 3D landmarks on the face through facial landmark detection (FLD) can work as an alternative to estimating the face structure. The goal of FLD is to localize predefined feature points on the 2D human face such as the nose tip, mouth, eye corners, etc., which have anatomical importance. Almost all of the FLD tasks try to predict sparse landmarks on the face which comprises 68 key points both in 2D or 3D space. Specifically due to its robustness to illumination and pose variations, 3D FLD task has gained increasing attention among the computer vision community. Unfortunately, this set of sparse landmarks fails to encode most of the intricate facial features. So increasing the number of these landmarks can help to learn face geometry better. However, publicly available datasets mostly contain a sparse set of 68 facial landmarks, which fails to encode the full face structure. To achieve a high landmark density we take the existing approach of [Feng et al., 2018a] which produces a position map data of the face. We propose a sampling methodology to filter 520 key points from the whole face and create a dataset to train a lightweight regression model. As we do not have access to any public dataset which has dense landmarks, we evaluated the trained regression model against the 68 landmark estimation task. Experimental result shows that the trained regression model can produce comparable result in the 3D face alignment task.

## 2 Related Work

Annotating a real face with dense landmarks is highly ambiguous and expensive. Some of the previous methods, like Wood et al.[Wood et al., 2021], rely on synthetic data alone. Though the authors have detailed ground truth annotations like albedo, normals, depth, and dense landmarks, none of these data is publicly available. The authors also proposed a method [Wood et al., 2022] to learn the dense landmarks as a Gaussian uncertainty from those synthetic data and fit a 3DMM model from those dense key points only. Some other methods [Deng et al., 2020, Feng et al., 2018a] use pseudo-labels model-fitting approaches like fitting an existing 3DMM model to generate synthetic landmarks. [Jeni et al., 2015] predicted dense frontal face landmarks with cascade regressions. Through an iterative method, they created 1024 dense 3D landmark annotations from 3D scan datasets [Zhang et al., 2014]. In contrast, Kartynnik et al. [Kartynnik et al., 2019] used a predefined mesh topology of 468 points arranged in fixed quads and fit a 3DMM model to a large set of in-the-wild images to create ground truth 3D dense annotations of key points. They later employed direct regression to predict these landmarks from face images. Some other methods [Alp Guler et al., 2017, Feng et al., 2018a] used a different method to unwrap each pixel of the face as a position map and regress the position in 3D space. They created the position map by fitting the Basel Face Model (BFM) [Paysan et al., 2009] from the 300WLP dataset [Zhu et al., 2016], which has the 3DMM parameters associated with more than 60k of wild images. As we don't have access to such massive 3D scan data, the same position map data can be an option to create the ground truth dense landmark.

## 3 Methodology

As discussed in the above section, we don't have access to large 3D scan data. So, generating position maps similar to [Feng et al., 2018a] can be an alternative. The position map records the 3D shape of the complete face in UV space as a 2D representation, where each pixel value has the 3D position information of that pixel. It provides correspondence to the semantic meaning of each point on the UV space. Their method aligns a 3D face model to the corresponding 2D face image and stores the 3D position of the points. We can apply the same to extract dense key points to create the ground truth data, before using direct regression to train a model that can predict those dense landmarks in 3D space. The following sections provide further details:

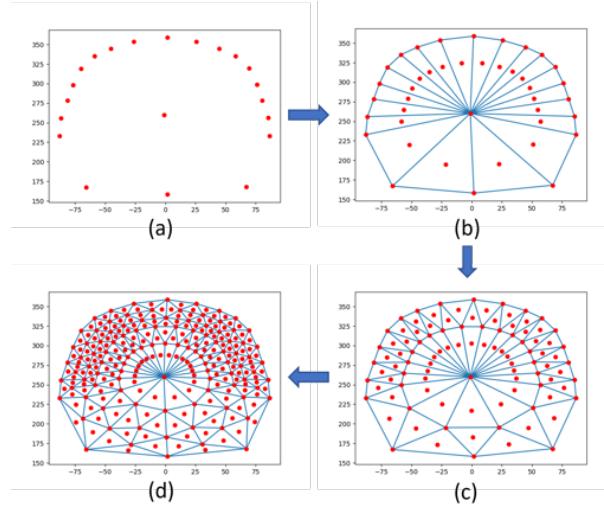


Figure 1: Selection of key points through Delaunay Triangulation. (a) Initial selected key points across the jaw, forehead, and nose tip (b) First iteration of Delaunay triangulation and centroid selection (c) Second iteration of Delaunay triangulation and centroid selection (d) Third iteration of Delaunay triangulation and centroid selection

### 3.1 Dense Facial Landmark Data Generation from UV Map

As stated above, [Feng et al., 2018a] proposed a 3D facial representation based on the UV position map. They used the UV space to store the 3D position points from the 3D face model aligned with the 2D facial image. They assume the projection from the 3D model on the 2D image as a weak perspective projection and define the 3D facial position as a Left-hand Cartesian coordinate system. The ground truth 3D facial shape exactly matches the 2D image when projected to the x-y plane. They define the position map as  $Pos(u_i, v_i) = (x_i, y_i, z_i)$ , where  $(u_i, v_i)$  represents the  $i$ th point in face surface and  $(x_i, y_i, z_i)$  represents the corresponding 3D position of facial mesh with  $(x_i, y_i)$  being the corresponding 2D position of the face in the input RGB image and  $z_i$  representing the depth value of the corresponding point.

We followed the same representation and used their pipeline to build the raw data from the 300W-LP [Zhu et al., 2016] dataset. This contains more than 60k unconstrained face images with fitted 3DMM parameters which are based on the Basel Face Model. They used the parameterized UV coordinates from [Bas et al., 2017], which compute a Tutte embedding [Floater, 1997] with conformal Laplacian weight and then map the mesh boundary to a square. So we can filter this UV position map data to create a dense face landmark. The 3DMM face template that was used by [Feng et al., 2018a] has a total of 43867 vertices. Out of these, we have sampled 520 vertices. To sample, we have followed the following steps as shown in figure 1 -

- First, we have selected 18 key points across the jaw and one key point on the nose tip from the 68 key points provided.
- Then we run the Delaunay triangulation on the selected points and select the centroids of the three vertices of each triangle.
- We repeat the same step another two times and have the final key points.
- Finally, we select these key points across the template mesh and manually select the rest of the key points and rectify some of the already selected key points in Blender.

After these iterations, the final version of the ground truth data has the RGB face images and their corresponding 68 face key points and the selected 520 key points. The whole dataset contains around 61k pairs of

ground truth images and their corresponding ground truth position map data saved in numpy format. Further, we expanded the data by applying a horizontal flip which made the total dataset size to 120k of paired images and their position map data.

### 3.2 Dense Facial Landmark Prediction using Regression

As we have around 120k pairs of ground truth face images in the wild and their corresponding ground truth 520 facial key points, we formulate the problem as a direct regression of those 520 key points from a monocular face image. We build a model with a standard feature extractor with a regressor head. The trained model will predict a continuous value of three positions (x,y,z) for those 520 3D landmarks. We choose the total number of classes as  $520 \times 3 = 1560$ . As the feature extractor, we have chosen two popular backbones, Resnet50 and MobilenetV2.

The standard loss function that is typically used for any landmark estimator is the  $L1$ , and  $L2$  loss or the Mean Square Error loss. But the  $L2$  loss ( $L2(x) = x^2/2$ ) function is sensitive to outliers, therefore [Rashid et al., 2017] used *smoothL1* loss where they updated the  $L2$  loss value with  $|x| - 1/2$  for  $x >= 1$ .

Both  $L1(L1(x) = |x|)$  and *smoothL1* perform well for outliers, but they produce a very small value for small landmark differences. This hinders the network training for small errors. To solve this issue, [Feng et al., 2018b] proposed a new loss called Wing loss which pays more attention to small and medium errors. They combined the  $L1$  loss for the large landmark deviations and  $\ln(\cdot)$  for small deviations as follows -

$$wing(x) = \begin{cases} w\ln(1+|x|/\epsilon), & \text{if } |x| < w \\ |x| - C, & \text{otherwise} \end{cases} \quad (1)$$

where  $C = w - w\ln(1+w/\epsilon)$ ,  $w$  and  $\epsilon$  are the hyperparameters ( $w = 15$ ,  $\epsilon = 3$  in the paper). In this work as well we combined the Meas Square Error loss with the Wing loss to define a hybrid loss function as -

$$L = w_1 L_{Wing} + w_2 L_{MSE} \quad (2)$$

Through an empirical study, we set the weight of these two loss terms as  $w_1 = 1.5$  and  $w_2 = 0.5$ .

## 4 Evaluation

As we don't have any separate evaluation or test dataset that has the 3DMM parameters or the position map data available, we evaluated our trained model on the 3D face alignment task. To measure the face alignment quantitatively, we use the normalized mean error (NME) as the evaluation metric. NME is computed as the normalized mean Euclidean distance between each set of corresponding landmarks in the predicted result  $l$  and the ground truth  $l'$  :

$$NME = \frac{1}{N} \sum_{i=1}^N \frac{\|l_i - l'_i\|_2}{d} \quad (3)$$

Following the previous works [Ruan et al., 2021], the normalization factor  $d$  is computed as  $\sqrt{h * w}$ , where  $h$  and  $w$  are the height and width of the bounding box, respectively. Similar to [Feng et al., 2018a] and [Ruan et al., 2021] for 2D and 3D sparse alignment, we consider all 68 landmark points. We divide the dataset based on the yaw angles (0,30), (30,60) and (60,90) and a balanced subset created by taking a random sample from the whole dataset. We benchmarked our model on the widely used AFLW2000-3D dataset. It is an in-the-wild face dataset with a large variation in illumination, pose, occlusion, and expression. It has 2000 images with 68 3D face landmark points for face alignment. The results are shown in table 1.

Following 3DDFA-V2 [Guo et al., 2020], we have also evaluated our work using the AFLW full set (21K test images with 21-point landmarks). We followed the same split and showed the results for different angles in table 2.

Table 1: Quantitative evaluation on AFLW2000-3D dataset on facial alignment task Metrics - NME (Lower is better) for different Head Pose Bins

Method	0 to 30	30 to 60	60 to 90	All
3DDFA [Zhu et al., 2016]	3.43	4.24	7.17	4.94
3DSTN [Bhagavatula et al., 2017]	3.15	4.33	5.98	4.49
3D-FAN [Bulat and Tzimiropoulos, 2017]	3.16	3.53	4.60	3.76
3DDFA TPAMI [Zhu et al., 2017]	2.84	3.57	4.96	3.79
PRNet [Feng et al., 2018a]	2.75	3.51	4.61	3.62
2DASL [Tu et al., 2020]	2.75	3.46	4.45	3.55
3DDFA V2[Guo et al., 2020]	2.63	3.420	4.48	3.51
Ours	2.86	3.68	4.76	3.77

Table 2: Quantitative evaluation on AFLW dataset with 21-point landmark definition on facial alignment task Metrics - NME (Lower is better) for different Head Pose Bins

Method	0 to 30	30 to 60	60 to 90	All
ESR [Cao et al., 2014]	5.66	7.12	11.94	8.24
3DDFA [Zhu et al., 2016]	4.75	4.83	6.39	5.32
3D-FAN [Bulat and Tzimiropoulos, 2017]	4.40	4.52	5.17	4.69
3DDFA TPAMI [Zhu et al., 2017]	4.11	4.38	5.16	4.55
PRNet [Feng et al., 2018a]	4.19	4.69	5.45	4.77
3DDFA V2[Guo et al., 2020]	3.98	4.31	4.99	4.43
Ours	4.04	4.45	5.2	4.57

Table 3: Comparative analysis with two different backbones Mobilenet-V2 and Resnet-18 of quantitative results on AFLW-3D dataset on facial alignment task and the computational requirement (gMac, gFlop, # params - Number of parameters) Metrics - NME (Lower is better) for different Head Pose Bins

Backbone	0 to 30	30 to 60	60 to 90	All	gMac	gFlop	# Params
Resnet-18	2.88	3.72	4.82	3.83	5.13	2.56	16.03M
Mobilenet-V2	2.86	3.68	4.76	3.77	0.39	0.19	4.18M

## 5 Discussion

As we do not have access to any dense landmark evaluation dataset, we evaluated our trained network against the 3D face alignment task for 68 points FLD on AFLW2000 and for 21 points FLD on AFLW dataset. The experimental results in table 1 and 2 show that our model is able to outperform most of the previous methods. Also as we have used the MobilenetV2-based model, the overall model size is comparatively small and requires less amount of computational resources. Table 3 shows a comparative analysis of the Rensent and Mobilenet backbone-based networks in terms of their computational resource requirement. We have also conducted a study on the effect of the hybrid loss function. Figure 3 shows the cumulative error distribution curves based on NME for the AFLW-3D and AFLW datasets. In both cases, a combination of Wing Loss and MSE performs better than the rest. Figure 2 shows some of the qualitative results on some samples from the AFLW dataset. The model shows a comparative result on extreme head pose angles as well as with occlusions and other facial variations.

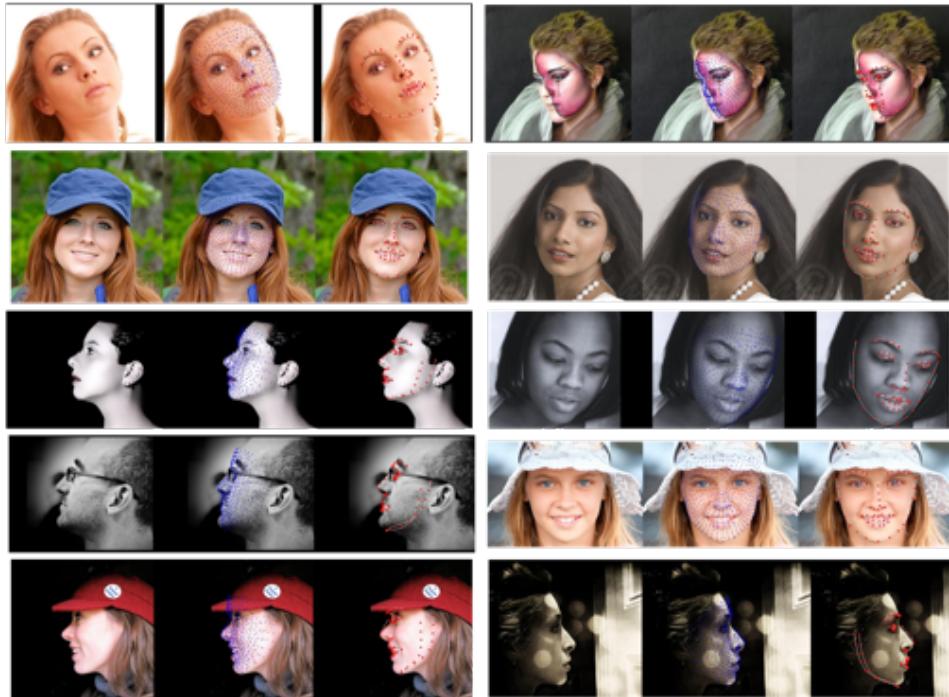


Figure 2: Qualitative results: the first image shows the ground truth image, the second image shows the 520 key points and the third image shows the 68 key points predicted by the model

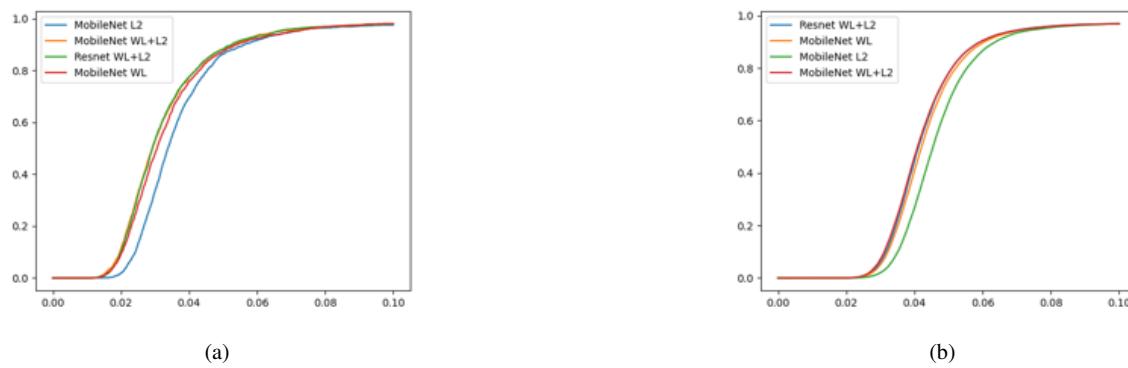


Figure 3: Different loss function study on Cumulative Errors (NME) Distribution (CED) curves on (a) AFLW2000-3D with 68 point landmarks and (b) AFLW with 21 point landmarks. The backbone and loss functions are also shown in the legend. WL stands for Wing Loss, and L2 stands for MSE loss.

## 6 Conclusion

In this work, we have presented a methodology to create a dense landmark dataset that has 520 key points generated from the UV position map data. With the help of generated dataset, we have trained an FLD regressor network with two different backbones, Resnet18 and MobileNetV2. As we do not have access to any other dense landmark evaluation dataset we have evaluated our trained model on a 68 points FLD task. Experimental results show our trained model is able to outperform most of the existing landmark detection methods while using fewer computational resources. The qualitative results show the robustness of our model and provide

better results in extreme head pose angles. Though visually, the results of the model look good, due to the lack of ground truth test data, we are only able to evaluate the model against the 3D facial alignment task. In the future, we can extend this work and use those 520 key points to fit an existing statistical (e.g., 3DMM) model to the face and evaluate the full face reconstruction benchmark.

## Acknowledgments

This work was supported by The Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality(d-real) under Grant No.18/CRT/6224. Some of this work is done as a part of an Internship at Xperi Corporation.

## References

- [Alp Guler et al., 2017] Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., and Kokkinos, I. (2017). Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6799–6808.
- [Bao et al., 2021] Bao, L., Lin, X., Chen, Y., Zhang, H., Wang, S., Zhe, X., Kang, D., Huang, H., Jiang, X., Wang, J., et al. (2021). High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1):1–21.
- [Bas et al., 2017] Bas, A., Huber, P., Smith, W. A., Awais, M., and Kittler, J. (2017). 3d morphable models as spatial transformer networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 904–912.
- [Beeler et al., 2010] Beeler, T., Bickel, B., Beardsley, P., Sumner, B., and Gross, M. (2010). High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9.
- [Bhagavatula et al., 2017] Bhagavatula, C., Zhu, C., Luu, K., and Savvides, M. (2017). Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3980–3989.
- [Bulat and Tzimiropoulos, 2017] Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030.
- [Cao et al., 2018] Cao, C., Chai, M., Woodford, O., and Luo, L. (2018). Stabilized real-time face tracking via a learned dynamic rigidity prior. *ACM Transactions on Graphics (TOG)*, 37(6):1–11.
- [Cao et al., 2014] Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International journal of computer vision*, 107:177–190.
- [Deng et al., 2020] Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.
- [Feng et al., 2018a] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018a). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551.
- [Feng et al., 2018b] Feng, Z.-H., Kittler, J., Awais, M., Huber, P., and Wu, X.-J. (2018b). Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245.

- [Floater, 1997] Floater, M. S. (1997). Parametrization and smooth approximation of surface triangulations. *Computer aided geometric design*, 14(3):231–250.
- [Guo et al., 2020] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., and Li, S. Z. (2020). Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*, pages 152–168. Springer.
- [Jeni et al., 2015] Jeni, L. A., Cohn, J. F., and Kanade, T. (2015). Dense 3d face alignment from 2d videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE.
- [Kartynnik et al., 2019] Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*.
- [Paysan et al., 2009] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee.
- [Rashid et al., 2017] Rashid, M., Gu, X., and Jae Lee, Y. (2017). Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6894–6903.
- [Ruan et al., 2021] Ruan, Z., Zou, C., Wu, L., Wu, G., and Wang, L. (2021). Sadnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Transactions on Image Processing*, 30:5793–5806.
- [Tu et al., 2020] Tu, X., Zhao, J., Xie, M., Jiang, Z., Balamurugan, A., Luo, Y., Zhao, Y., He, L., Ma, Z., and Feng, J. (2020). 3d face reconstruction from a single image assisted by 2d face images in the wild. *IEEE Transactions on Multimedia*, 23:1160–1172.
- [Wood et al., 2021] Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., and Shotton, J. (2021). Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691.
- [Wood et al., 2022] Wood, E., Baltrušaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljević, N., Wilde, D., Garbin, S., Sharp, T., Stojiljković, I., et al. (2022). 3d face reconstruction with dense landmarks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 160–177. Springer.
- [Zhang et al., 2014] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706.
- [Zhu et al., 2016] Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155.
- [Zhu et al., 2017] Zhu, X., Liu, X., Lei, Z., and Li, S. Z. (2017). Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92.
- [Zollhöfer et al., 2018] Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., and Theobalt, C. (2018). State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer graphics forum*, volume 37, pages 523–550. Wiley Online Library.

# YUDO: YOLO for Uniform Directed Object Detection

Dorđe Nedeljković

*Independent Researcher*

## Abstract

This paper presents an efficient way of detecting directed objects by predicting their center coordinates and direction angle. Since the objects are of uniform size, the proposed model works without predicting the object's width and height. The dataset used for this problem is presented in Honeybee Segmentation and Tracking Datasets project [Bozek et al., 2017]. One of the contributions of this work is an examination of the ability of the standard real-time object detection architecture like YoloV7 [Wang et al., 2022] to be customized for position and direction detection. A very efficient, tiny version of the architecture is used in this approach. Moreover, only one of three detection heads without anchors is sufficient for this task. We also introduce the extended Skew Intersection over Union (SkewIoU) [Ma et al., 2017] calculation for rotated boxes - directed IoU (**DirIoU**), which includes an absolute angle difference. **DirIoU** is used both in the matching procedure of target and predicted bounding boxes for mAP calculation, and in the NMS filtering procedure. The code and models are available at <https://github.com/djordjened92/yudo>.

**Keywords:** Object detection, Horizontal detection, Rotated object detection, Directed object

## 1 Introduction

The original dataset from the paper [Bozek et al., 2017], which inspired this work was intended to be used for resolving the general multi-object tracking in a crowded environment like a beehive. Each bee was manually annotated with  $(x, y, t, \theta)$ , where  $x$  and  $y$  represent coordinates of the body center,  $t$  class label (1 when the full body is visible and 2 when the bee is inside a comb cell) and  $\theta$  is the angle of body direction against the vertical bottom-up axis with values from the range  $[0, 2\pi]$  calculated clockwise. An angle is annotated only for the bee class ( $t = 1$ ) and ignored for the abdomen class ( $t = 2$ ). In the original work, the localization of the tracked honeybee bodies was implemented through a segmentation approach using U-Net [Ronneberger et al., 2015], in combination with angle distance loss and temporal component, so authors generated an approximation of bee body size annotations (around the annotated center)

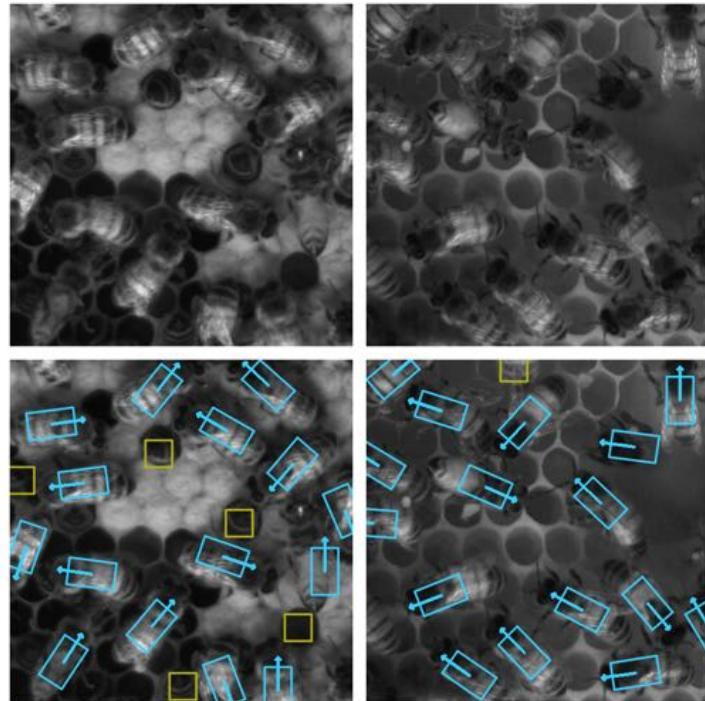


Figure 1: Dataset image samples with annotation visualizations.

using ellipses for  $t = 1$  with semi-minor axis  $r1 = 20$  pixels and semi-major axis  $r2 = 35$  pixels, and for  $t = 2$  the model is a circle with  $r = 20$ . We aim to show that modern object detection architectures can predict the center and direction angle of uniformly sized objects in a dense configuration. Therefore, the neural network is not provided with object size during the training time. The beforementioned approximation of the bee body is transformed into a more suitable bounding box format which encloses an ellipse or circle for the abdomen label. These bounding boxes are used in the process of non-maximum suppression filtering of predictions and prediction-target matching in evaluation time. Two samples of images (top) and their annotations (bottom) are shown in Figure 1. The blue arrowed boxes represent the bee class and the yellow squares represent the abdomen class.

## 2 Related Work

Object detection is one of the most important and one of the most comprehensively developed research areas in computer vision, and nowadays deep learning methods are a dominant approach for this task. According to the alignment of the labeled and detected boxes, this field of research is divided into more general horizontal detection and rotated object detection. Horizontal object detection, by which all boxes are aligned to horizontal and vertical axes, is more suitable for natural scene images. For the purpose of more accurate object positioning, like scene text detection or aerial image analysis, rotated detection is a more suitable approach.

### 2.1 Horizontal and Rotated Object Detection

**Horizontal Object Detection:** Horizontal object detection is the default object detection subfield which assumes horizontal bounding boxes as object representations. The dominant backbones so far in this field are ConvNets [LeCun et al., 1989] of various forms.

*Two stage detection* is a family of approaches which are dominantly region based. In the first stage, they generate category-independent proposals of bounding boxes, and then after feature extraction of those regions, they apply classification and regression in the second stage. Fast RCNN [Girshick, 2015], Faster RCNN [Ren et al., 2015] and R-FCN [Dai et al., 2016] belong to this category and they don't exploit the hierarchical structure of ConvNets but rely on the single feature map instead, in both stages. FPN [Lin et al., 2016] is also a two stage approach, but it exploits all stages of the hierarchical backbone with top-down and lateral connections. The FPN design inspired many recent object detection architectures.

*Single stage detection* methods aim to find regions of interest, regress coordinates, and apply classification in one neural network pass. Due to their efficiency, most of the real-time solutions rely on this approach. SSD [Liu et al., 2015] generate class scores and bounding box parameters per feature map location on different scales of convolution layers. Therefore, this has been one of the first methods leveraging the hierarchical nature of ConvNets. RetinaNet [Lin et al., 2017] is trying to address background/foreground class imbalance by implementing the FocalLoss. The most popular family of single stage detection is YOLO [Redmon et al., 2015]. Although there has been a lot of progress in the YOLO-type detectors over time, the principle of coordinates regression and class scores prediction directly from feature map(s) of ConvNets remains throughout all versions.

Most of the mentioned approaches use bounding box suggestions called anchors in the prediction stage, but there are also so-called anchor-free approaches like FCOS [Tian et al., 2019], CornerNet [Law and Deng, 2018], CenterNet [Zhou et al., 2019]. Nowadays, transformer is the most popular and widely spread architecture. Object detection is no exception, so DETR [Carion et al., 2020] and ViTDet [Li et al., 2022] are representatives of this category.

**YoloV7:** Besides a few trainable bag-of-freebies, which are tricks used in training proposed by other authors, this work introduced some original architectural improvements. The YoloV7 [Wang et al., 2022] architecture

belongs to the concatenation-based type of model, which usually consists of computational and transition blocks. The computational blocks are the main components which are aimed to distillate qualitative feature maps, using parallel branches of inference. The results are aggregated using a concatenation. The transition layers are helper elements, usually built from convolution and pooling operations, used to maintain output resolution and the number of channels. Inspired by VoVNet [Lee et al., 2019] architecture, the authors suggested an improved computational block called E-ELAN (Extended Efficient Layer Aggregation Networks). A group convolution is used to expand the channel and cardinality of computational blocks. Output feature maps of these groups are shuffled, concatenated, and finally added.

An additional contribution of this approach is the specific scaling of models. Increasing depth in a concatenation-based model implies that the output width of a computational block also increases. YoloV7 [Wang et al., 2022] deals with this by scaling up the width of transition layers.

There are few designed basic models of YoloV7 (for edge GPU, normal GPU, and cloud GPU). We tried to apply our approach to the most efficient, edge GPU version of architecture - YOLOv7-tiny.

**Rotated Object Detection:** The mainstream approaches of rotated detection are mostly adapting the horizontal detection paradigm by representing the bounding boxes with five parameters: coordinates of the bounding box center, bounding box width, bounding box height, and rotation angle -  $(x, y, w, h, \theta)$ . The rotation angle can be determined by the  $x$ -axis and closest rectangle edge in the  $90^\circ$  range  $\theta \in [-90^\circ, 0^\circ]$ , or by the long side of the rectangle and the  $x$ -axis in the  $180^\circ$  range  $\theta \in [-90^\circ, 90^\circ]$ . The notation is that the clockwise direction is negative.

The region based approaches usually regress these five parameters, implementing different variants of  $l_n$ -norm losses (such as smooth  $l_1$  loss) like R-DFPN [Yang et al., 2018], R<sup>3</sup>Det [Yang et al., 2019a], RSDet [Qian et al., 2019], CSL [Yang et al., 2020], or exploiting differentiable approximation of IoU loss like SCRDet [Yang et al., 2019b], GWD (Gaussian Wasserstein distance) [Yang et al., 2021].

Either directly or implicitly, all these methods rely on the regression of the rotation bounding:

$$\begin{aligned} t_x &= (x - x_a)/w_a, t_y = (y - y_a)/h_a \\ t_w &= \log(w/w_a), t_h = \log(h/h_a), t_\theta = f(\theta - \theta_a) \\ t'_x &= (x' - x_a)/w_a, t'_y = (y' - y_a)/h_a \\ t'_w &= \log(w'/w_a), t'_h = \log(h'/h_a), t'_\theta = f(\theta' - \theta_a) \end{aligned} \quad (1)$$

where  $x, x_a, x'$  are for the ground-truth box, anchor box and predicted box, respectively (likewise  $y, w, h, \theta$ ). The function  $f(\cdot)$  is usually used to deal with angular periodicity such as trigonometric functions, modulo (like for the  $l_{mr}^{5p}$  loss in the RSDet [Qian et al., 2019]), etc. The regression loss usually has a general form:

$$L_{reg} = l_n - norm(\Delta t_x, \Delta t_y, \Delta t_w, \Delta t_h, \Delta t_\theta) \quad (2)$$

where  $\Delta t_j = |t_j - t'_j|$  for  $j \in \{x, y, w, h, \theta\}$ .

All aforementioned methods contain this loss component, extended in different ways to address **boundary discontinuity problems** mainly caused by the periodicity of angular (PoA). Put simply, when the target and predicted angle are close to the opposite ends of the range, the regression loss suddenly increases. In the case of  $180^\circ$  range shown in Figure 2, the ground-truth bounding box is close to the  $\pi/2$ , and the predicted box is close to the  $-\pi/2$ . As rotated object

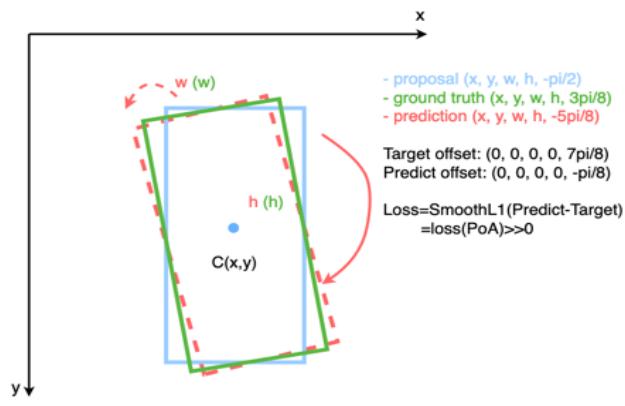


Figure 2: The boundary discontinuity caused by PoA on the  $180^\circ$  range example [Yang et al., 2020].

detection is agnostic to the direction, then these 2 boxes are visually very similar despite of high Smooth L1 loss.

### 3 Directed Object Detection

Since the standard rotated object detection is agnostic to the direction of objects, we introduce a method for directed object detection in this work. The angle of a directed object is measured relative to only one certain edge, in contrast to rotated object detection. For the main backbone architecture, we use YOLOv7-tiny [Wang et al., 2022].

#### 3.1 Bounding Box Parametrization

The bounding box model of a honeybee is shown in Figure 3. The angle of the bee's direction  $\theta$  is determined by the vertical axis (bottom-up orientation) and the bee's head position. Its value can be from the range  $[0, 2\pi)$  calculated in the clockwise direction. Central point coordinates are  $(x, y)$ .

#### 3.2 Regression method

The concept of YOLO architecture is to map detection head output into the grid which covers the input image. Each cell provides  $N_a$  prediction vectors where each of them represents one predicted bounding box, and  $N_a$  is the number of anchors. Thus each cell's output nodes related to the coordinates  $(x, y)$  represent the offset from the cell top-left corner. These values are in the normalized form, relative to the width and height of the cell's receptive field. It's intuitive why these nodes have a sigmoidal activation function since the possible position of the predicted bounding box is from the top-left to the bottom-right border point of the cell. Our simplifications:

- $(w, h)$  nodes for bounding box width and height are omitted from the prediction vector since those values are predefined and uniform.
- We used an anchor-free approach where each cell has one possible prediction, which is justified when overlapping of objects is not expected and if the output grid is dense enough in order not to miss prediction in the crowded parts of the scene.
- Instead of three detection heads (with different grid sizes and assigned sets of anchors) only one is kept with grid size 16x16.

Before loss calculation, all ground-truth bounding boxes are assigned to the specific cell according to the central point position. The regression loss for the central point is the mean squared error:

$$L_{xy} = \frac{1}{N} \sum_{i=1}^N ((\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2) \quad (3)$$

where  $t_i$  is the relative coordinate offset of ground-truth box,  $\hat{t}_i$  is the coordinate of assigned prediction for  $i \in \{x, y\}$ .  $N$  is total number of bounding boxes.

Alongside  $(x, y)$  nodes in the output vector for each cell, there is a dedicated node for the angle  $\theta$  prediction. The aforementioned **boundary discontinuity problem** caused by PoA appears in this setup, too. An example

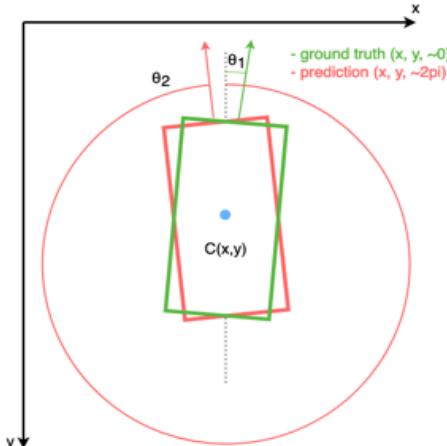


Figure 3: Bounding box model of honeybee.

is presented in Figure 3, where the target bounding box has the angle  $\theta_1 \approx 0$  and the predicted angle is  $\theta_2 \approx 2\pi$ . In the standard  $L_1$  loss, the absolute difference  $\approx 2\pi$  would cause sudden raise of regression loss, although it is obvious these bounding boxes have close angles of direction. One possible way to deal with this issue is to use cosine distance:  $1 - \cos(\Delta\theta)$ , which has the lowest value of 0 for the case of a small difference in direction angles, and a maximum value of 2 in the case when directions are opposite. The output angle node has ReLU activation function and afterward update of value  $\theta = \theta \pmod{2\pi}$ . The cosine distance loss has the form:

$$L_\theta = \frac{1}{N} \sum_{i=1}^N (1 - \cos(\hat{\theta}_i - \theta_i)) \quad (4)$$

where  $\theta$  is the target angle of direction,  $\hat{\theta}$  is the predicted angle and  $N$  is number of boxes.

Therefore, the output vector for each grid cell is defined as  $(x, y, \theta, obj, abd_{cls}, bee_{cls})$ . The node  $obj$  represents the objectness node responsible for the sureness of prediction existence, and nodes  $abd_{cls}, bee_{cls}$  are classification nodes for the abdomen and regular bee class respectively. Activations for classes are independent logistic classifiers. The total loss is defined as:

$$L = \lambda_{xy} L_{xy} + \lambda_\theta L_\theta + \lambda_{cls} L_{cls} + \lambda_{obj} L_{obj} \quad (5)$$

where  $\lambda_i$  is a weight for specific loss in the total sum.  $L_{cls}$  and  $L_{obj}$  are classification and objectness losses in the form of standard binary cross-entropy losses. The loss and activations related to objectness and classification are kept from the original implementation of YoloV7 [Wang et al., 2022].

### 3.3 Directed IoU

Skew Intersection over Union (SkewIoU), introduced by [Ma et al., 2017], is an IoU version adapted for rotated bounding boxes. The premise of this property adopted from horizontal bounding boxes is the same, with suggestions for intersection area calculation. It includes the difference in bounding box orientation, thus it is sufficient for describing oriented bounding box detection. However, this formulation of IoU suffers from a lack of explicit taking into account the difference in **direction angle** of a bounding box which has a certain "head" edge, for which the direction angle is calculated. We upgraded the IoU for rotated bounding boxes in Equation (6) with the correction factor, which emphasizes the difference in direction.

$$\begin{aligned} DirCorr(\Delta\theta) &= \frac{1 + \cos(\Delta\theta)}{2} \\ DirIoU &= IoU \cdot DirCorr \end{aligned} \quad (6)$$

For the calculation of the rotated bounding box IoU, we used detectronv2 [Wu et al., 2019] implementation. Figure 4 depicts the effect of the correction factor ( $DirCorr(\theta)$  in Equation 6) on the rotated bounding box IoU. Two bounding boxes (in this case two bee class boxes with  $w = 40, h = 70$ ), which have aligned center positions, reach maximal  $IoU = 1$  for an angle difference of  $0^\circ$  or  $180^\circ$ . To emphasize the importance of direction, our  $DirIoU$  has maximal value only when boxes are aligned and the angle difference is  $0^\circ$ . The lowest value of  $DirIoU$  is reached when boxes are not intersecting or when their directions are opposite. In Figure 4  $DirIoU$  has a value 0 for the angle difference of  $180^\circ$ , where boxes are overlapping completely, but their directions are opposite. It is evident that this criterion of IoU is more rigid, so we used the threshold of 0.3 instead of the default value of 0.5 for mAP calculation and NMS filtering.

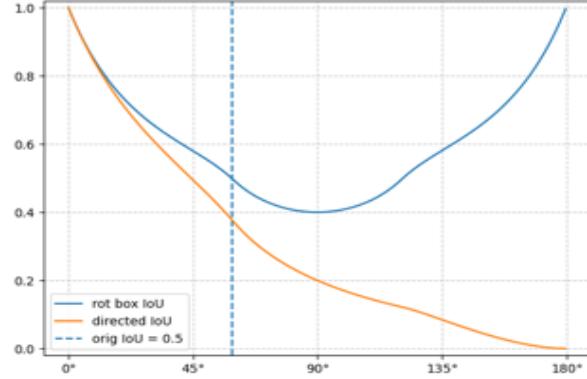


Figure 4: Rotated bounding box IoU and Directed IoU.

## 4 Experiment

**Dataset:** As suggested in the original work [Bozek et al., 2017], we randomly sampled test data in equal proportions from both image pools of 30FPS and 70FPS recordings. The model is trained on 13908 images of size 512x512 and validated on 1392 images of the same size. We applied a few augmentation techniques during the training, such as vertical and horizontal flips, 2x2 mosaic image arrangement, and HSV random parameter changes. Since the original images have blurred regions, we applied a sharpening filter also. On image loading, we filter it by 3x3 kernel with center value 9 and -1 for other cells.

**Model and Training:** We customized the YoloV7 [Wang et al., 2022] tiny architecture as a detection model. As mentioned before, this customization implies an anchor-free version of the model where only one of three detection heads is kept. This neural network has ~6 million parameters or 13.1 GFLOPS. The model is trained for 200 epochs with batch size 4, applying the one-cycle learning rate schedule [He et al., 2018] with an initial learning rate of 0.001. Initial model weights are from the pre-trained model available at the official github repository. We use Stochastic Gradient Descent as an optimizer and Exponential Moving Average (EMA) of model weights update for stabilized training. Chosen loss weights are:  $\lambda_{xy} = 0.1, \lambda_\theta = 0.1, \lambda_{cls} = 0.3, \lambda_{obj} = 1.0$ . The training is conducted using the GeForce RTX2070 Super GPU.

**Results:** The best checkpoint mAP@30 result is **70.5**, and class-specific metrics are presented in Table 1. For the bee class, we're applying directed object detection with the center and the angle prediction, but for the abdomen class, the center is predicted only. So in terms of the result discussion, the bee class is much more important, where obtained numbers are pretty satisfying. The abdomen class appears problematic in terms of results, and one of the causes might be non-consistent annotations noticed during the error analysis.

	Labels	Precision	Recall	AP <sub>30</sub>
bee	21163	82.3	88.2	85.1
abdomen	2940	58.5	60.0	55.9

Table 1: Detection results.

## 5 Conclusion

We propose the detection method for directed, uniform objects and the corresponding customization of modern horizontal object detectors. This approach meets the requirements of the problem with complete omission of objects' width and height. The directed bee class is detected with mAP@30=**85.1** using only one YOLOv7-tiny [Wang et al., 2022] detection head in simplified anchor-free form. The proposed angle loss component overcomes the problem of angle periodicity.

In addition, we define a proper metric (DirIoU) for overlapping measurement of directed objects. It emphasizes the angle difference, introducing an additional factor alongside the standard IoU, which is focused on the object areas only. That periodic function cancels completely any area overlapping if the directions are opposite ( $\Delta\theta \approx \pi$ ) and keeps the IoU value in the case when directions are similar ( $\Delta\theta \approx 0$ ).

## A Examples of detections and problematic abdomen annotations

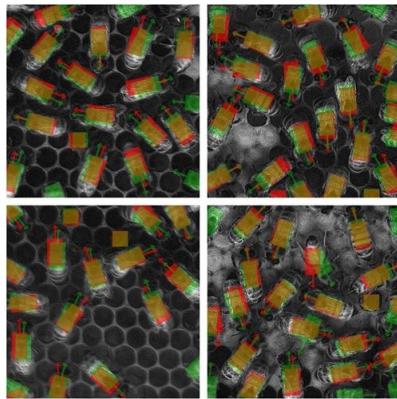


Figure 5: Detection examples. Arrowed rectangles are bee class and squares are abdomen class. Red color represents target and green is prediction (overlaps are yellow).

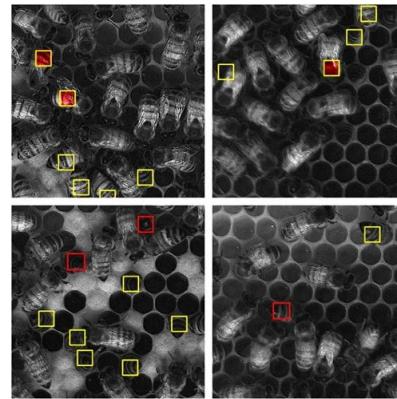


Figure 6: Problematic abdomen annotations. On the top two images wrong abdomen labels contain red fill. Bottom images have missed abdomen labels shown with red squares.

## References

- [Bozek et al., 2017] Bozek, K., Hebert, L., Mikheyev, A. S., and Stephens, G. J. (2017). Towards dense object tracking in a 2d honeybee hive. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4185–4193.
- [Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. *ArXiv*, abs/2005.12872.
- [Dai et al., 2016] Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *ArXiv*, abs/1605.06409.
- [Girshick, 2015] Girshick, R. B. (2015). Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.
- [He et al., 2018] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. (2018). Bag of tricks for image classification with convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 558–567.
- [Law and Deng, 2018] Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128:642–656.
- [LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.
- [Lee et al., 2019] Lee, Y., won Hwang, J., Lee, S., Bae, Y., and Park, J. (2019). An energy and gpu-computation efficient backbone network for real-time object detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 752–760.
- [Li et al., 2022] Li, Y., Mao, H., Girshick, R. B., and He, K. (2022). Exploring plain vision transformer backbones for object detection. *ArXiv*, abs/2203.16527.

- [Lin et al., 2016] Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., and Belongie, S. J. (2016). Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944.
- [Lin et al., 2017] Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017). Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.
- [Liu et al., 2015] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., and Berg, A. C. (2015). Ssd: Single shot multibox detector. In *European Conference on Computer Vision*.
- [Ma et al., 2017] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., and Xue, X. (2017). Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20:3111–3122.
- [Qian et al., 2019] Qian, W., Yang, X., Peng, S., Guo, Y., and Yan, C. (2019). Learning modulated loss for rotated object detection. *ArXiv*, abs/1911.08299.
- [Redmon et al., 2015] Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2015). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.
- [Ren et al., 2015] Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.
- [Tian et al., 2019] Tian, Z., Shen, C., Chen, H., and He, T. (2019). Fcos: Fully convolutional one-stage object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635.
- [Wang et al., 2022] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *ArXiv*, abs/2207.02696.
- [Wu et al., 2019] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. <https://github.com/facebookresearch/detectron2>.
- [Yang et al., 2019a] Yang, X., Liu, Q., Yan, J., and Li, A. (2019a). R3det: Refined single-stage detector with feature refinement for rotating object. In *AAAI Conference on Artificial Intelligence*.
- [Yang et al., 2018] Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., and Guo, Z. (2018). Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote. Sens.*, 10:132.
- [Yang et al., 2020] Yang, X., Yan, J., and He, T. (2020). On the arbitrary-oriented object detection: Classification based approaches revisited. *International Journal of Computer Vision*, 130:1340 – 1365.
- [Yang et al., 2021] Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., and Tian, Q. (2021). Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning*.
- [Yang et al., 2019b] Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., and Fu, K. (2019b). Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8232–8241.
- [Zhou et al., 2019] Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *ArXiv*, abs/1904.07850.

# Accurate object detection using the sensor fusion of an event-based and a frame-based camera

Orla Sealy Phelan, Dara Molloy, Roshan George, Edward Jones, Martin Glavin, Brian Deegan

*University of Galway*

## Abstract

Efficient object detection is crucial to real-time monitoring applications such as autonomous driving. Modern RGB cameras can produce high-resolution images for accurate object detection. However, with increased resolution comes increased network latency and power consumption. To minimise this latency, CNNs often have a resolution limitation, requiring images to be down-sampled before inference, causing significant information loss. In this paper, we propose a neuromorphic vision approach based on biological vision, where images are cropped instead of down-sampled to meet the requirements of the CNN. The design implements the sensor fusion of an event-based camera and a frame-based RGB camera to implement an accurate, low-power monitoring system. The cameras are calibrated to create a multi-modal stereo vision system where pixel coordinates can be projected between the event camera and RGB camera image planes. Events are detected using clustering on the event-based data and projected to the RGB image plane using the calibration results. These projections identify important areas in the RGB image, directing the cropping areas. Using this implementation, the COCO AP is increased from 21.08 to 57.38 on bicycles in the RGB scene, with an overall increase from 37.93 to 46.89 for all classes tested.

**Keywords:** Sensor Fusion, Event-Based Vision, Camera Calibration, Object Detection, Image Processing

## 1 Introduction

Efficient monitoring systems are needed for a range of applications including autonomous vehicles, and security systems. Many of these applications are safety-critical systems that require high accuracy, and high temporal resolution to ensure that accurate responses are made in real-time. Modern cameras have improved significantly in recent years resulting in higher-quality images with improved resolution and dynamic range. Although this helps significantly with accurate object detection, it also comes at a cost. The power consumption and bandwidth needed to process these high-quality images are very high. Because of this, many convolutional neural networks (CNNs) have a resolution limitation to meet latency constraints. The resolution of images imputed into these networks is typically limited to 600x600 pixels or less, meaning that high-resolution images must be down-sampled, therefore reducing the accuracy of the detection, especially for smaller details in the image.

An event camera is a neuromorphic vision sensor, processing only the changes in a scene rather than the entire scene. In these sensors, each pixel asynchronously measures logarithmic changes in light intensity and records any changes that are above or below a pre-defined threshold. This removes a huge amount of redundant data which would be recorded in a regular camera, where data is recorded by every pixel at a pre-defined interval, wasting power and memory on static, unchanging scenes [Prophesee, 2023]. The resulting data recorded by these cameras is a sequence of events corresponding to areas of movement in the image in the form of the x, y coordinates of the pixel, the time, and the polarity of the light intensity change (increasing or decreasing brightness).

This design provides many advantages over state-of-the-art frame-based cameras. Event cameras can achieve temporal resolutions  $> 10,000$  fps (frames per second), with sub-millisecond latency. Also, the dynamic range achievable by event cameras is  $> 120$ dB making the sensors more robust to adverse conditions such as low light, or glare from the sun.

In this work, we propose an object detection system using the sensor fusion of an event-based camera with a frame-based camera. The design takes inspiration from the human visual system. The event camera acts as the low-resolution peripheral vision to detect movement in the scene, while the RGB camera acts as the fovea where high-resolution analysis of the image is completed. During this analysis, the RGB images will be cropped instead of down-sampled to meet the resolution limitations of the CNN used, allowing high-accuracy detection without forfeiting temporal resolution.

The main contributions of this research include: an adaptation to common camera calibration techniques which facilitates the calibration of an event-based camera with a frame-based camera using common automated software; a simple equation that can be used to project pixel coordinates from the image plane of one camera to the other with reasonable accuracy, when complex geometrical analysis is not feasible; a foveal vision approach to real-time object detection using an event camera to reduce information loss in high-resolution RGB images, showing considerable accuracy improvements for small objects.

## 2 State of the Art

### 2.1 Event-based vision

Numerous recent studies investigate the use of event cameras for object detection and tracking. Rebecq et al propose a recurrent neural network, which can be used to reconstruct high-resolution grayscale images from event streams [Rebecq et al., 2019]. The results show impressive quality and give good classification accuracy on off-the-shelf CNNs; however, this comes at an increased computational cost. Shariff et al demonstrated a proof of concept forward perception system using event-based YOLOv5 detection [Shariff et al., 2023]. The viability of raw event data in object detection models is shown by [Perot et al., 2020] and [Ryan et al., 2021], but the training of such models is limited by the availability of event-based training data, with [Ryan et al., 2021] using a simulated dataset for training. The use of unsupervised clustering algorithms combats this issue for tracking moving objects [Hinz et al., 2017, Mondal et al., 2021, Mondal and Das, 2021], but these algorithms do not provide sufficient information for object classification. In this work, the combination of data from an event-based camera and a frame-based camera allows common frame-based object detection techniques to be implemented and advanced using the characteristics of the event camera.

### 2.2 Sensor Fusion

There is a significant amount of research exploring the sensor fusion of RADAR (Radio Detection and Ranging) or LiDAR (Light Detection and Ranging) with regular cameras to gather depth information in a scene [Willis et al., 2009, Chai et al., 2018, Zhou et al., 2018, Kim and Park, 2019, Singandhupe and La, 2021], yet there is limited research on the multi-modal fusion of event-based cameras. The idea of combining event-based sensors with regular frame-based sensors is investigated for object detection [Perot et al., 2020, Tomy et al., 2022],

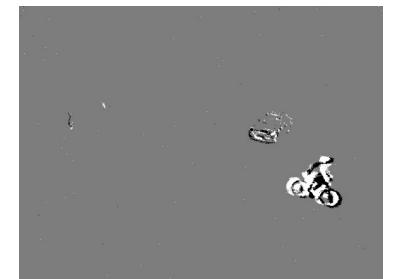


Figure 1: Example frame from the event camera (left) and RGB camera (right).

SLAM [Weikersdorfer et al., 2013, Vidal et al., 2018, Gao et al., 2022], optical flow estimation [Lee et al., 2022], and visual odometry [Chen et al., 2023]. The DAVIS sensor [Brandli et al., 2014] is used in many of the sensor fusion implementations [Zhu et al., 2018, Chen et al., 2023, Lee et al., 2022], which combines an asynchronous Dynamic Vision Sensor (DVS) with a synchronous frame-based sensor on a single photodiode. In this sensor, the event-based and frame-based cameras share a common reference frame making the calibration process easier, as standard calibration can be run on the frame-based image. Unfortunately, the low resolution of the sensor limits its functionality in object detection systems. Other approaches use stereo event camera systems in combination with high-resolution frame-based stereo systems to calculate depth in the 3D coordinate system [Vidal et al., 2018, Gehrig et al., 2021, Gao et al., 2022]. However, the implementation of a multi-modal stereo system where image points are projected between event-based and frame-based image planes of contrasting resolutions is limited to [Tomy et al., 2022]. Tomy et al show improvements in model robustness in adverse weather conditions using event-based data combined with frame-based data. However, the effects on the complexity of the resulting network are not evaluated. Each of these sensor fusion approaches involves data collection from both cameras simultaneously, and the event data is combined with each regular frame to evaluate the results. This increases the complexity of the system and does not take advantage of the low power consumption available with event cameras.

### 3 Methodology

#### 3.1 Data

The data was collected by a camera rig containing an event camera and an RGB camera, highlighted in Table 1. A thirty-second video sample is used for testing, which consists of vehicles, bikes, and pedestrians maneuvering around Dangan Car Park, Galway. An example frame from each camera is seen in Figure 1. The RGB data was recorded at a frame rate of 30 fps, giving an accumulation time of 33,333  $\mu$ s. This accumulation time is used to generate and visualise frames from the event data, which could be synchronised with the RGB video.

##### 3.1.1 Camera Calibration

In this study, a modified version of the method proposed by [Zhang, 2000] was used. The static checkerboard was replaced with a flashing checkerboard GIF (Graphics Interchange Format), and by displaying it on a large LCD screen, images can be generated from the event-based and frame-based cameras allowing stereo calibration using standard toolboxes. In this research, the Stereo Camera Calibrator App available from the MATLAB Computer Vision Toolbox is used to calculate the intrinsic and extrinsic parameters of each camera according to the following steps: create a flashing checkerboard GIF and display on a large screen; mount RGB camera on apparatus beside event camera and ensure apparatus is stable and secure; record video of flashing checkerboard using both cameras simultaneously and repeat for 10 – 20 different camera angles; generate event-based frames from event-camera data and extract corresponding frames from RGB video data; process images to ensure aspect ratios match and resize images to have the same resolution, by cropping the event frame to 640x338 pixels and downscaling the RGB image to this resolution; calibrate cameras using software to determine the intrinsic and extrinsic parameters of each camera.

Following successful calibration, Equation 1 can be used to project pixel coordinates from the event-camera frame,  $[x_2, y_2]$ , to the RGB camera frame,  $[x_1, y_1]$ . This equation uses the intrinsic parameters which remain

	Event Camera	RGB Camera
Name	PROPHASEE ONBOARD	Blackfly S
Spatial Resolution (pixels)	640 x 480	4096 x 2160
Temporal Resolution (fps)	>5000	42
Dynamic Range (dB)	>120	<50
Power Consumption (W)	0.026	3
Pixel Pitch ( $\mu$ m)	15	3.45

Table 1: Specifications of cameras used in the dataset.

constant for each camera, K1, K2, and the extrinsic parameters, R1, R2, and t1, t2, which change depending on the location of each camera. The extrinsic parameters of the cameras were not recorded at the time of the dataset collection, so the parameters used in this equation are the average relative extrinsic parameters of the two cameras,  $(R_E)_{av}$ . This can be derived according to Equation 1 for each set of extrinsic parameters and finding the mean over all angles used in the calibration. By using this average extrinsic value, this equation can be generalised to any situation if the cameras remain in the same position relative to each other.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{K_1}{K_2} \cdot (R_E)_{av} \cdot \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} \quad (1)$$

$$R_E = \begin{bmatrix} R_1 & t_1 \\ R_2 & t_2 \end{bmatrix} \quad (2)$$

### 3.2 Clustering & Object Detection

To detect and track moving objects in the event-camera data, clustering can be used to group events into objects. To implement this, event coordinates are processed into an array for each frame according to the accumulation time of the RGB camera. Prior to implementing the clustering, the following steps are run on the resulting array: remove any out-of-frame points to account for the change in aspect ratio used in the calibration; undistort the points using the OpenCV undistortPoints function with the intrinsic matrix and distortion coefficients of the event camera.

The DBSCAN clustering algorithm is then run on the formatted array to determine the locations of each moving object in the event camera data. To determine their location in the RGB frame, Equation 1 is used to project the resulting bounding box coordinates to the RGB image plane. This equation assumes that the images used are the same resolution as the images used in the calibration process, and that lens distortion has been removed. To convert the resulting points to the original RGB image plane, the coordinates must be distorted using the distortion coefficients of the RGB camera, according to Equation 3, and converted back to the original RGB camera resolution using Equation 4. These bounding box coordinates can then be used to crop the RGB frame, and object detection is run on the cropped images.

$$\begin{aligned} x' &= (1 + 2p_1 + p_2(r^2 + 2x^2) + k_1r^2 + k_2r^4)x \\ y' &= (1 + 2p_2 + p_1(r^2 + 2y^2) + k_1r^2 + k_2r^4)y \end{aligned} \quad (3)$$

$$x'' = \frac{4096x'}{640}, \quad y'' = \frac{2160y'}{338} \quad (4)$$

Where  $(x, y)$  are the projected points,  $(x', y')$  are the distorted points, and  $(x'', y'')$  are the points in the original RGB image.  $(k_1, k_2, p_1, p_2)$  are the distortion coefficients, and  $r^2 = x^2 + y^2$ . The object detection model used in this research is the YOLOv5s model [Jocker, 2020]. The resolution requirements for the model are for the input images to be 640x640 pixels. PyTorch is used for inference, and all required data augmentation is automatically completed on the image when inputted into the model.

### 3.3 Performance Evaluation

Performance is measured using mean average precision (mAP), which combines precision, recall, and bounding box accuracy using an intersection over union (IoU) threshold. The Microsoft COCO dataset [Lin et al., 2014] calculates the AP over a range of IoU thresholds. The standard COCO metric measures over ten thresholds from 0.5 – 0.95 with an increment of 0.05. The method used in the PASCAL VOC dataset [Everingham et al., 2010] calculates the average AP over each class for a fixed IoU threshold. AP50 uses a threshold of 50 %, while AP75 uses a threshold of 75 %. To investigate the advantages of the application, the performance of the system is compared against the baseline of running the same model on each full frame of the RGB video data.

## 4 Results & Discussion

The calibration accuracy is measured according to the reprojection error of the checkerboard corners for each calibration image. Table 2 shows the results achieved for the initial calibration, the final calibration, and from using Equation 1 to project the points. The original calibration gave the lowest mean reprojection error, however only 13 images were used, and the entire FOV of the cameras were not covered. When tested on real-world data, this calibration performed poorly at the edges of the image. More data was collected to cover a larger area of the FOV, and the calibration was repeated, giving an error of 0.24 pixels. Figure 2 shows accurate reprojection by the MATLAB software for a calibration image at the edge of the image, validating the calibration.

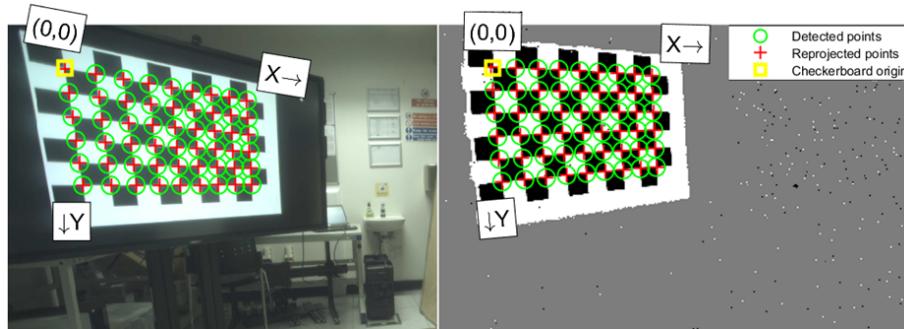


Figure 2: Reprojection results using MATLAB Stereo Camera Calibrator

Using these calibration results, Equation 1 was used to project corners from the event camera image to the RGB camera image, giving a mean reprojection error of 1.42 pixels, according to Table 2. This does show an increased error; however, it is still only slightly over 1 pixel, and it allows the projection to be generalised and used on data where the exact extrinsic parameters of the camera are unknown.

## 5 Clustering & Object Detection

Hyperparameter tuning of the DBSCAN clustering algorithm was performed through visual analysis. The hyperparameters chosen were Epsilon=20 and minPoints=10. The results of the DBSCAN algorithm and the projection of the resulting bounding boxes are shown for an example frame in Figure 3. Note that both images are undistorted in this figure.

Table 3 shows the results achieved on the full-frame implementation (FF), and the cropped implementation (Crop) using the sensor fusion approach. The accuracy results achieved using COCO AP, AP50, and AP75 are highlighted. The results are calculated for each class, and an average over all classes is calculated. The % tab shows the factor of improvement of the cropped detection over the full-frame detection according to Equation 5.

Method	# Images	Mean Reprojection Error
MATLAB Original Calibration	13	0.09
MATLAB Original Calibration	17	0.24
Projection using Equation 1	17	1.42

Table 2: Reprojection error for two calibration attempts, and error calculated using derived equation with parameters from final calibration.

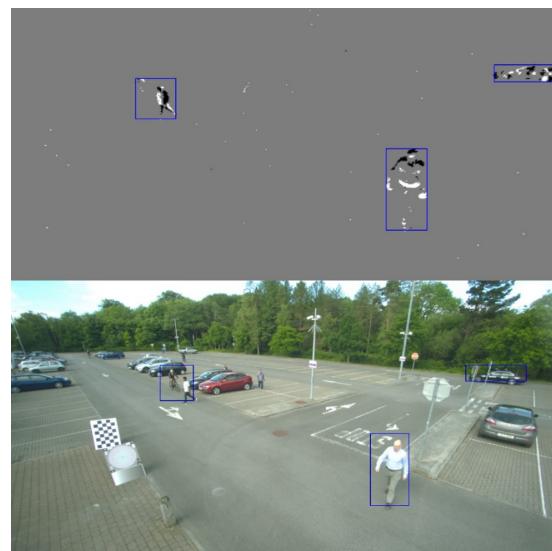


Figure 3: Results of DBSCAN clustering to and projection to RGB image using Equation 1

$$\% = \frac{Crop - FF}{FF} \quad (5)$$

Looking at the results achieved using COCO AP, the cropped implementation shows an improvement factor of 0.24 over all classes. The bicycle class accuracy improves by a factor of 1.72 from 21 % to 57 %. The person class also shows an improvement of 0.32. The car class sees a decrease in accuracy from 47.54 % to 23.55 %. It is important to note the limitations of this metric for the car class. On initial measurement, the accuracy of the car class was very low, giving only 7.83 % and 5.35 % COCO AP for the full-frame and cropped implementations, respectively. This application is interested in moving objects only, therefore any static objects in the frame were not annotated in the dataset. Unfortunately, this resulted in the parked cars in the image being marked as false positives during the accuracy evaluation resulting in very low precision scores. To avoid this, the background was blurred using a median filter mask. This increased precision for both implementations. However, it increased recall for the full frame implementation, but decreased recall for the cropped implementation. For this reason, the accuracy results cannot be reliably compared for the car class.

	COCO AP			AP50			AP75		
	FF	Crop	%	FF	Crop	%	FF	Crop	%
Car	47.54	23.55	-0.5	55.62	34.05	-0.39	53.88	23.59	-0.56
Person	45.16	59.73	0.32	76.15	84.19	0.11	45.05	69.08	0.53
Bicycle	21.08	57.38	1.72	39.33	70.61	0.8	22.86	67.00	1.93
Average	37.93	46.89	0.24	57.03	62.95	0.1	40.60	53.22	0.31

Table 3: Object detection results for full-frame (FF) and cropped (Crop) implementations

A similar relationship can be seen in the AP50 and AP75 results, with the bicycle class showing the highest improvement, and the car class showing a decrease in accuracy. The accuracy of the full-frame implementation reduces when the IoU threshold is increased from 0.5 – 0.75. For the person, bicycle, and average classes, the improvement factor of the crop over the full frame increases significantly for AP75: from 0.11 to 0.53 for the person class; 0.8 to 1.93 for the bicycle class; and 0.1 to 0.31 overall. This shows that the bounding box prediction is more accurate when using the cropped images, as fewer predictions are eliminated by the increased threshold.

## 6 Conclusions & Future Work

In this research, an ‘Artificial Fovea’ is proposed using the sensor fusion of an event camera with an RGB camera. We outline a method of calibrating an event camera with an RGB camera using a standard calibration toolbox and an equation is derived to allow projection between image planes of each camera without additional extrinsic calibration. The design implements the DBSCAN clustering algorithm on event-based data to determine the location of moving objects in a scene. These results are projected to the RGB camera image plane using the derived equation. The RGB frame is then cropped according to these results, and object detection is run on each of these cropped images. The system is evaluated by measuring the accuracy of the object detection achieved using the YOLOv5s algorithm and comparing these results with the same model run on each full frame of data from the RGB video. The cropping of the images shows good accuracy improvements over the full frame implementation with the COCO AP increasing from 37.93 % for the full frame implementation, to 46.89 % on the cropped frames. The highest accuracy improvement was seen for the bicycle class where the COCO AP increased from 21.08 % to 57.38 %. This shows the advantages of cropping the frame before inference, particularly for small objects in the image. This work highlights the accuracy improvements that can be achieved. Further analysis is required to address the low AP scores for the car class. It is likely, however, that the poor AP scores for the car class are due to data labeling issues and handling of static objects, rather than an intrinsic limitation of the proposed approach. Future work will explore the potential benefits of the proposed approach, including potential improvements in object detection accuracy, power consumption, frame rate, and image compression.

## Acknowledgments

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 P2 ([www.lero.ie](http://www.lero.ie)), and in part by Valeo Vision Systems.

## References

- [Brandli et al., 2014] Brandli, C., Berner, R., Yang, M., Liu, S. C., and Delbruck, T. (2014). A  $240 \times 180$  130 dB 3  $\mu$ s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49:2333–2341.
- [Chai et al., 2018] Chai, Z., Sun, Y., and Xiong, Z. (2018). A novel method for LiDAR camera calibration by plane fitting. In *2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 286–291. IEEE.
- [Chen et al., 2023] Chen, P., Guan, W., and Lu, P. (2023). Esvio: Event-based stereo visual inertial odometry. *IEEE Robotics and Automation Letters*.
- [Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.
- [Gao et al., 2022] Gao, L., Liang, Y., Yang, J., Wu, S., Wang, C., Chen, J., and Kneip, L. (2022). Vector: A versatile event-centric benchmark for multi-sensor SLAM. *IEEE Robotics and Automation Letters*, 7(3):8217–8224.
- [Gehrig et al., 2021] Gehrig, M., Aarents, W., Gehrig, D., and Scaramuzza, D. (2021). Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954.
- [Hinz et al., 2017] Hinz, G., Chen, G., Aafaque, M., Röhrbein, F., Conradt, J., Bing, Z., Qu, Z., Stechele, W., and Knoll, A. (2017). Online multi-object tracking-by-clustering for intelligent transportation system with neuromorphic vision sensor. In *KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25–29, 2017, Proceedings 40*, pages 142–154. Springer.
- [Jocker, 2020] Jocker, G. (2020). YOLOv5 by ultralytics. <https://github.com/ultralytics/yolov5>.
- [Kim and Park, 2019] Kim, E.-S. and Park, S.-Y. (2019). Extrinsic calibration of a camera-LiDAR multi sensor system using a planar chessboard. In *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 89–91. IEEE.
- [Lee et al., 2022] Lee, C., Kosta, A. K., and Roy, K. (2022). Fusion-flownet: Energy-efficient optical flow estimation using sensor fusion and deep fused spiking-analog network architectures. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6504–6510. IEEE.
- [Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- [Mondal and Das, 2021] Mondal, A. and Das, M. (2021). Moving object detection for event-based vision using K-means clustering. In *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–6. IEEE.

- [Mondal et al., 2021] Mondal, A., Giraldo, J. H., Bouwmans, T., Chowdhury, A. S., et al. (2021). Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 876–884.
- [Perot et al., 2020] Perot, E., De Tournemire, P., Nitti, D., Masci, J., and Sironi, A. (2020). Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652.
- [Prophesee, 2023] Prophesee (2023). Metavision for machines. <https://www.prophesee.ai/>.
- [Rebecq et al., 2019] Rebecq, H., Ranftl, R., Koltun, V., and Scaramuzza, D. (2019). High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980.
- [Ryan et al., 2021] Ryan, C., O’Sullivan, B., Elrasad, A., Cahill, A., Lemley, J., Kielty, P., Posch, C., and Perot, E. (2021). Real-time face & eye tracking and blink detection using event cameras. *Neural Networks*, 141:87–97.
- [Shariff et al., 2023] Shariff, W., Farooq, M. A., Lemley, J., and Corcoran, P. (2023). Event-based YOLO object detection: proof of concept for forward perception system. In Osten, W., Nikolaev, D. P., and Zhou, J. J., editors, *Fifteenth International Conference on Machine Vision (ICMV 2022)*, volume 12701, page 127010A. International Society for Optics and Photonics, SPIE.
- [Singandhupe and La, 2021] Singandhupe, A. and La, H. M. (2021). Single frame LiDAR and stereo camera calibration using registration of 3d planes. In *2021 Fifth IEEE International Conference on Robotic Computing (IRC)*, pages 115–118. IEEE.
- [Tomy et al., 2022] Tomy, A., Paigwar, A., Mann, K. S., Renzaglia, A., and Laugier, C. (2022). Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 933–939. IEEE.
- [Vidal et al., 2018] Vidal, A. R., Rebecq, H., Horstschafer, T., and Scaramuzza, D. (2018). Ultimate SLAM? combining events, images, and imu for robust visual SLAM in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001.
- [Weikersdorfer et al., 2013] Weikersdorfer, D., Hoffmann, R., and Conradt, J. (2013). Simultaneous localization and mapping for event-based vision systems. In *Computer Vision Systems: 9th International Conference, ICVS 2013, St. Petersburg, Russia, July 16-18, 2013. Proceedings 9*, pages 133–142. Springer.
- [Willis et al., 2009] Willis, A. R., Zapata, M. J., and Conrad, J. M. (2009). A linear method for calibrating LiDAR-and-camera systems. In *2009 IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems*, pages 1–3. IEEE.
- [Zhang, 2000] Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334.
- [Zhou et al., 2018] Zhou, L., Li, Z., and Kaess, M. (2018). Automatic extrinsic calibration of a camera and a 3d LiDAR using line and plane correspondences. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5562–5569. IEEE.
- [Zhu et al., 2018] Zhu, A. Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., and Daniilidis, K. (2018). The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039.

# A Comparative Analysis of Deep Learning Mobile Networks on Marine Vessel Detection

Dipak G Sharma, Michael O'Neill

*Natural Computing Research and Application Group (NCRA), School of Business,*

*University College Dublin*

## Abstract

Maritime vessel detection and classification is a challenging research domain in computer vision due to, for example, ever-changing weather patterns, tidal and seasonal variations. In recent years, deep learning or deep neural networks (DNN) have shown enormous potential in the field of computer vision and have steered us towards practical applications such as vessel identification and classification and so on. There have also been promising results in the accuracy of mobile network architectures to support the real time application of DNN's in a maritime environment. However, there is a gap in the literature in the comparative evaluation of mobile DNN architectures in the maritime domain. This paper presents a comparison of three high-performing networks, each of different size, on marine vessel datasets for marine object detection and classification task. The performance matrices are analysed for model accuracy and speed. We observe that among the models evaluated on the Singapore Maritime Dataset (SMD)-plus, in terms of accuracy, yolo-v8 has the best performance.

**Keywords:** Marine Vessels, Machine Vision, Detection, Localization, Deep Learning

## 1 Introduction

Maritime vessels are one of the important modes of passenger, freight, and cargo transportation for the modern world. Additionally, these vessels are actively used for pleasure activities and sport. In 2020 alone, 230 million passengers embarked and disembarked European ports, and 3.3 billion tonnes of freight were handled [Eurostat, 2022]. This large-scale movement of people and goods presents additional challenges as the increase in these activities will, for example, likely correspond with an increase in human trafficking, illicit drug trafficking, and illegal goods smuggling. One relatively recent report shows that most of the cocaine seized in the EU is mainly transported by maritime shipping containers [Europole, 2022]. This underlines the increasing need for a reliable, and automated system to address these issues.

Deep learning is becoming a widely accepted technology in almost every domain, whether it is to predict protein structure from sequence data, develop intelligent driving agents, or in natural language settings [Jumper et. al, 2021] [Tom et. Al, 2020]. Additionally, these innovations are advancing rapidly with the emergence of advanced hardware and a bounty of data. From this perspective we can sense that the development in deep learning technology to be gradually shifting towards two distinct directions. One towards the development of larger models, to generalize to wider application settings, such as represented by the growth in parameters of large language models (LLMs) such as growth in GPT-3 to GPT-4 [Brown et al., 2020] [OpenAI, 2023]. While the other direction is towards miniaturised versions so that they can be effectively implemented on low resource devices for numerous real-world applications “on the edge”, using for instance a drone platform. As a result, in

recent years, there has been extensive research in smaller networks architectures, which balance a trade-off between model accuracy and latency [Wang et. al]. These smaller models have shown state-of-the-art results in many publicly available datasets including Pascal VOC, ImageNet, and MS-Coco. However, there has not been enough open-sourced dataset to test the robustness of these model beyond the scope of general object detection, face detection or similar. Even with the few existing datasets that are suitable for the task, they have not been properly assessed with the latest models, which have been reported to have better trade-off between model accuracy and the computational latency.

This study aims to address this gap and is motivated by the urge to apply novel technologies in addressing those problems. To this end, we analysed smaller but multiple high performing network architectures “out of the box” with the default parameter settings on a domain specific problem of sea-vessel detection and localization. This will allow us to compare models in terms of both accuracy and speed for real world application using a drone platform. The models that we evaluated are mobilenet-v3 (large) with SSDlite, yolo-v8 (large), and mobilenet-v3 (large) with FRCNN [Howard et. al, 2019] [Jocher et. al, 2023] [Ren et. al, 2015]. We analyse performance on the Singapore maritime dataset (SMD)-plus [Kim et. al, 2022]. In the remainder of this paper we focus on a brief review of research in publicly available maritime datasets, and a review of general object detection and maritime vessels detection. We then outline our experimental design before discussing results and opportunities for future research.

## 2 Related Work

**Maritime Datasets:** Although DNN based approaches have shown promising results in recent times, one of the key issues with them is the requirement of a large-scale ground truth training dataset. Research on maritime vessel detection has been largely affected due to the lack of annotated datasets. There are several datasets available in the public domain, but they are limited by their intended aim of image classification but not localisation as most of them lack bounding box annotation for object localisation. For example, the Seagull dataset has multi-spectrum images for research on maritime environments however these datasets are limited to vessel detection and are not suitable for vessel localisation without further annotating them [Marques et. al]. Similarly large-scale image datasets of maritime vessels [Gundogdu et. al, 2016] include over 100 vessels categories; however, they also do not include bounding box annotation. Other publicly available datasets include a marine obstacle detection dataset [Kristan et. al, 2015], which holds data for large and small obstacles but does not accommodate neither proper classification of vessels nor the annotation for object localisation. Additionally [Moosbauer et. al, 2019] [Zhang et. al, 2021] lays out references for other multiple sources of publicly available datasets but they are limited to a reduced set of categories of carrier/vessels and does not supply the large video segment.

The Singapore Maritime Dataset (SMD) [Prasad et al., 2016], however, provides larger multi-spectrum video segments with 10 different classes of sea-vessels. The data are classified in two distinct categories as onshore data and onboard data. Onboard data is collected using a vision sensor mounted on vessels as they move along Singaporean ports, while the onshore data is collected from the seashore. SMD- plus [Kim et. al, 2022] is a refined version of SMD through the addition of annotation, tightening of the bounding boxes, and the classes are reduced to 7 from the original 10 in SMD. The SMD includes annotation for both object detection and classification. It includes both class labels and bounding boxes annotation for object localization. The dataset can be obtained from the link in its original paper [Kim et. al, 2022] (<https://github.com/kjunhwa/SMD-Plus>). The github link also provides reference to the original dataset by Prasad et. al. Figure 1 shows samples from SMD-plus video clips. We can see that the dataset is quite varied given the fact that it includes videos from different weather conditions.

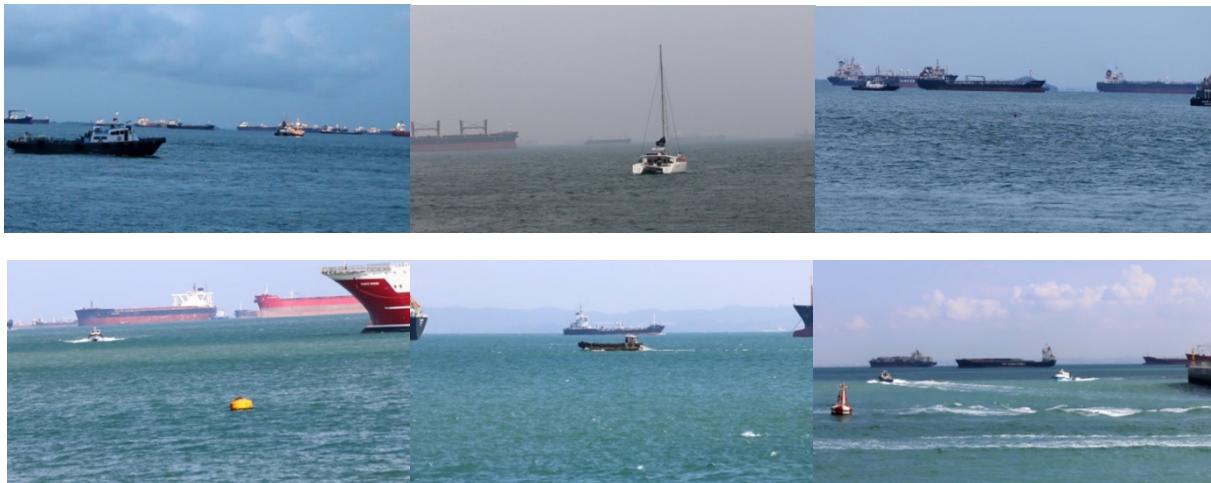


Figure-1: SMD-plus dataset samples

**Sea-vessel detection:** There has been much research in computer vision and several approaches have been investigated for detection, classification, and localization of marine objects. For example, [Dawkins et. al, 2014] uses the feature engineering technique to identify maritime objects from the ocean, which is further extended by [Maire et. al, 2013] in marine mammals' detection based on images collected in aerial surveys. In marine object detection we can see multiple works in horizon detection [Prasad et al., 2018] [Teutsch & Krüger, 2010] as a part of pre-processing. Prasad et. al with their SMD dataset evaluate multiple methods for background reduction, however, such approaches might not be robust enough in the context of sea or coastal environments as background changes are abrupt and very common in such a context [Moosbauer et. al, 2019]. Several other techniques such as Hidden Markov Models and Gaussian Mixture Models (GMM) have been used to overcome such issues [Westall et. al, 2009] [Frost & Tapamo, 2013], but deep learning approaches seem to be successful in more complex object detection and segmentation problems.

There have been few attempts at maritime object detection and classification using smaller architectures in deep learning. Most of those works have assessed multiple variants of yolo [Leela et. Al, 2020] [Yan et. Al, 2022]. However, we rarely observe the evaluation of these models with other variations like widely popular mobile-net variations or similar ones. For example, [Kim et. al, 2022] in their publication have evaluated SMD-plus with yolo-v5 and further used copy-paste technique to improve the accuracy of the model, however, it does not provide any overview of the speed and accuracy of the model in real-time scenarios. Similarly [Moosbauer et. al, 2019] has benchmarked SMD on two different state-of-the-art models, Faster-RCNN and Mask-RCNN, and has introduced additional techniques such as multiscale combinatorial grouping to generate pixel-wise segmentation from bounding box information but again the paper lacks comparison of those models with other architectures. Likewise [Tian et al., 2021] [Sun et al., 2022] have focused their research on either object classification, detection, or segmentation but it also lacks the analysis with models having fewer parameters. A recent paper by [Munin et. al, 2023] has evaluated yolo-v8 on the dataset of only 624 images for 13 different vessels and has achieved impressive results with mAP@50 of 98.9% for the similar task of detection and classification, however, the result is based on a limited dataset which hardly seems to contain objects of smaller (pixel) size. Therefore, the contribution of this paper could be seen as an effort to bridge the gap such that it attempts to; (i) evaluate existing smaller networks models using the SMD-plus dataset and, ii) benchmark the performance on more constrained computing resources.

### 3 Experimental Setup

As mentioned in the preceding section we use the SMD-plus dataset to evaluate the models. The dataset is provided in two categories of onboard and onshore data. Fig-2 shows the distribution of video frames on those two categories. Clearly, it has a higher number of video frames from the onboard camera which is mounted on sea vessels while it was slowly moving around the Singaporean port. This is beneficial in terms of model training as these onboard videos have high degree of tidal oscillation which helps in creating variations in the dataset. Fig-3 shows the distribution of the number of objects on each class on SMD-plus dataset. We can clearly observe that dataset has a class imbalance problem, such that, the vessel/ship category has relatively higher number of objects associated with it of nearly 71%. Therefore, we train the model in two different settings, (i) *including all 7 classes* and, (ii) *with only 6 classes* by removing the vessel/ship category as we want to observe the effect of the imbalanced dataset on the overall result of the model. The dataset is provided as a raw video frame in *avi* format. We used a total of 51 video frames for training. From each video file we created image frames each with a sample of 5 frames. A total of 4377 frames were generated from the provided video file. We further randomly subsample images to train/validation sets by distributing them in a ratio of 7:3. Those 30% frames in the validation set were vital for better generalization of the model as they help in preventing model overfitting. We further added data-augmentation technique in the training framework to increase the robustness in the dataset. Basically, we only used horizontal flip for generating additional image frames.

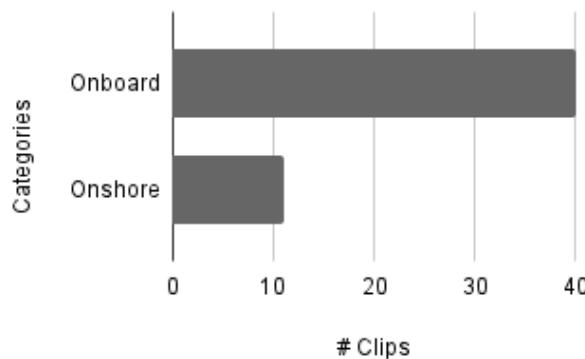


Figure-2: Distribution of onshore and onboard video clips in SMD-plus dataset

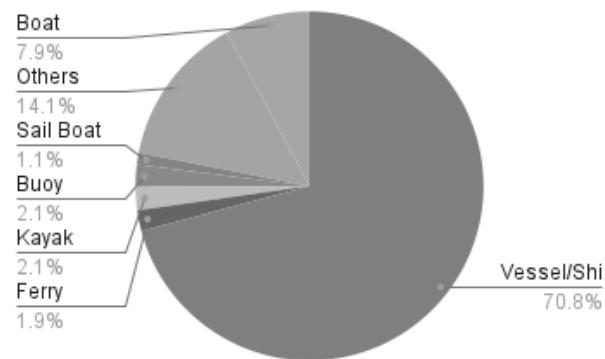


Figure-3: Distribution of objects associated with classes on SMD-plus dataset

Models were trained in two different frameworks. Mobilenet-v3 with SSDlite was trained on TensorFlow while yolo and mobilenet-v3 with FRCNN version were trained on Pytorch. We consider these models because (i) they are among the state of art model architectures for edge platforms [Wang et. al, 2022] [Terven et. al, 2023] whose performance have not yet been evaluated together in domain specific sea-vessel detection and, (ii) the model includes smaller architecture (like mobile-net) to much complex versions like yolo-v8 and halfway sized mobilenet-v3 with FRCNN, these variation in parameters would enable us to examine co-relation in between size of the model and its accuracy if any. SSDlite is the lighter version of SSD architecture [Liu et al., 2016] which might not generalize sufficiently in the complex object detection task therefore, keeping the feature extractor same (mobilenet-v3) with multiple detectors would allow us to confirm if lighter feature extractor (model) with more complex detector produces any better results. As we wanted to observe the effect of parameter sizes in this small architecture, we consider experimenting with models with varying parameter sizes in a way that are large enough to better generalize in domain specific sea-vessels detection but relatively small enough to run efficiently in a real-time scenario on edge platforms. Fig-6 shows the size of those three models. We can observe from the figure that mobilenet-v3 is relatively smaller than the remaining two models, whereas yolo-v8 has the higher number of parameters. Training was performed using NVIDIA RTX 4090 GPU. For consistency

batch size were set to 16 and were trained for 100 epochs. The input size of the image was 640x640. In total, we conducted at least 6 different training instances; first half being the case where models were trained and evaluated on all 7 classes, and the remaining half was performed on 6 classes excluding vessels/ship category.

## 4 Results

The performance of the model has been quantified in terms of both average precision to evaluate the accuracy of the model, and a count of average frames per second to measure computational latency. As mentioned in the previous section, we apportioned our evaluation into two sets. In the first set, assessment was staged on the dataset with all class categories. The chart illustrated in Fig-4 shows the evaluation result in the form of mean average precision (mAP) [Padilla et. al]. The figure shows that the yolo-v8 has better accuracy in sea-vessel detection and localization problem in contrast to the other two models. It scores 0.82 mAP with intersection of union (IoU) in between 0 to 50 while the scores shrink to 0.52 while evaluating it with IoU between 50 to 90. Similarly, on the second set, vessels/ship categories were excluded, and assessment was done with only 6 classes. Fig-5 underlines the outcome and we can witness that the accuracy drops overall. Unlike yolo in previous cases, for 6 classes the accuracy of yolo drops, however it is still the best performer in overall evaluation. Likewise, in case of mobilenet-v3 with SSDlite detector accuracy seems consistent in both cases although there is a small fluctuation on the scores while comparing the result with 6 classes. Mobilenet-v3 with frcnn detector has high mAP in comparison to the one with SSDlite but the score reduces while accessing it with 6 classes. Other interesting observations that can be made while evaluating Fig-4 and Fig-5 is that the accuracy in terms of average precision seems to increase as the complexity of the model increases. However, the observation does not correlate with the latter case.

The plot illustrated in Fig-6 demonstrates size verses speed performance of the models. Note that they were evaluated on an NVIDIA GPU RTX 3060Ti. As expected, we observed that the model with a smaller parameter size (mobilenet-v3-SSDLite) has the lowest latency with an average fps of 77. In contrast, yolo-v8 whose fps is recorded as 33 on average (with almost 47.5 million parameters) has the highest latency of the models examined. Mobilenet-v3 with the more complex FRCNN detector has similar inferencing speed to the SSDLite version.

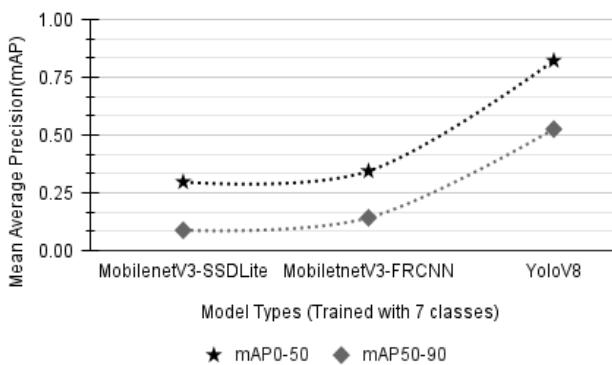


Fig-4: Mean average precision of the models trained on SMD-plus with 7 classes. The black dotted line with star mark indicate result for mAP50 and the latter gray mark indicates result for mAP90

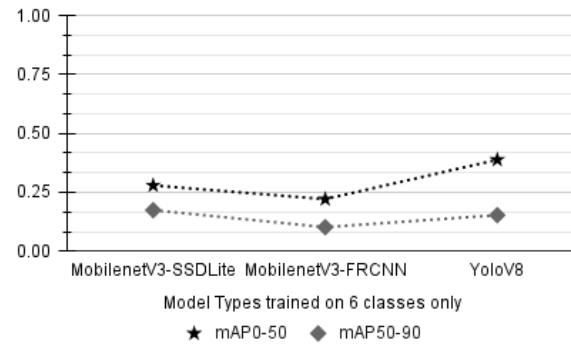


Fig-5: Mean average precision of the models trained on SMD-plus with 6 classes. In this case, we excluded vessels/ship class. The black dotted line with star mark indicate result for mAP50 and the latter gray mark indicates result for mAP90

## 5 Discussion and future work

One of the primary objectives of this evaluation was to identify the most accurate, but smaller, model to address the domain specific problem of marine vessel detection, which could be deployed on the edge. In so doing, we observe that in general yolo-v8 outperforms the two other models that were evaluated. During the experiment it was also observed that mobilenet-v3 with SSDlite was particularly good with smaller object detection, and yolo-v8 was better in detecting larger objects. However, further research is required to verify this outcome as the observations may be associated with the class imbalance issue. Another noticeable observation in this study is the imbalance of the data and its impact. As

more than two thirds of objects were associated with only a single class, we can see that there is a significant variation with the result of evaluation between only 6 classes (without the majority class) when compared with a dataset comprised of all 7 classes. The mAP significantly drops while the training was performed in only 6 classes. With half of the dataset, it might be possible that the number of datasets was simply not enough to allow model generalisation. Conversely, it might also be possible that the accuracy was elevated in case of all 7 classes since the number of majority class objects (vessel/ship) were far greater, and the detection of these object might have been better than other smaller objects such as sailing boats, buoys, or kayaks. In addition, we observe that the larger models (higher number of parameters) generalize well in sea-vessel detection, however the experiment excluding the majority class doesn't seem to enhance performance. Likewise, switching the detector of mobilenet-v3 with FRCNN does not clearly make any distinction on the result as the outcome varies significantly in between the cases.

There are multiple possible directions in which to continue this research into the future. For instance, since we are evaluating these models performances for time critical applications, detail analysis of evaluation on multiple edge platform is missing in this paper which is the immediate work that we are looking to accomplish going forward. Additionally, with the current experiment we can see that the dataset has been a major limitation as there is a huge difference in the ratio between the majority class objects and the others. Therefore, further exploration is needed in the data collection and curation so that a balanced dataset could be prepared for such domain specific problems. Another interesting direction could be to search the space of smaller architectures (network architecture search) using a neuro-evolutionary approach where an evolutionary algorithm is used to search network topologies with a view to find models having a better trade-off between accuracy and size. Likewise, we can see that some networks have very low computational latency and others have better accuracy, hence exploring a hybridisation between these networks would be another interesting area to explore in the future.

## Acknowledgement

This study was supported by the Government of Ireland through an Enterprise Ireland Disruptive Technologies Innovation Fund DT 2020 0268C (UCD)

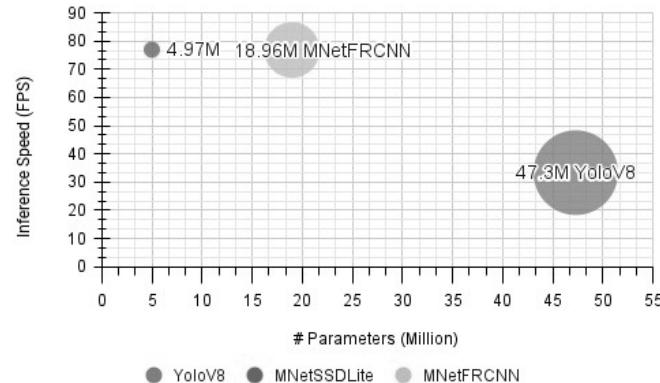


Figure-6: Model size versus computational latency (evaluated on RTX 3060Ti GPU)

## References

- [Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 1877-1901.
- [Dawkins et al., 2014] Dawkins, M., Sun, Z., Basharat, A., Perera, A., & Hoogs, A. (2014, June). Tracking nautical objects in real-time via layered saliency detection. In *Geospatial Info Fusion and Video Analytics IV; and Motion Imagery for ISR and Situational Awareness II* (Vol. 9089, pp. 10-19). SPIE.
- [Europole, 2022] Europe and the global cocaine trade in EU Drug Market. (2022). Cocaine – In-depth analysis by the EMCDDA and Europol, [https://www.emcdda.europa.eu/publications/eu-drug-markets/cocaine/europe-and-global-cocaine-trade\\_en#box\\_analysis\\_seized](https://www.emcdda.europa.eu/publications/eu-drug-markets/cocaine/europe-and-global-cocaine-trade_en#box_analysis_seized) (last accessed 2023/05/28)
- [Eurostat, 2022] Eurostat regional yearbook 2022 edition. 2022. Publications Office of the European Union, ISBN: [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Eurostat\\_regiona...\\_yearbook](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Eurostat_regiona..._yearbook) (last accessed: 2023/05/28)
- [Frost & Tapamo, 2013], Frost, D., & Tapamo, J. R. (2013). Detection and tracking of moving objects in a maritime environment using level set with shape priors. *EURASIP Journal on Image and Video Processing*, 2013(1), 1-16.
- [Gundogdu et al, 2016] Gundogdu, E., Solmaz, B., Yücesoy, V., & Koc, A. (2017). Marvel: A large-scale image dataset for maritime vessels. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13* (pp. 165-180). Springer International Publishing.
- [Howard et al., 2019] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1314-1324).
- [Jocher et al., 2023] Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics (Version 8.0.0) [Computer software]. <https://github.com/ultralytics/ultralytics>
- [Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
- [Kim et al., 2022] Kim, J. H., Kim, N., Park, Y. W., & Won, C. S. (2022). Object detection and classification based on YOLO-V5 with improved maritime dataset. *Journal of Marine Science and Engineering*, 10(3), 377.
- [Kristan et al., 2015] Kristan, M., Kenk, V. S., Kovačić, S., & Perš, J. (2015). Fast image-based obstacle detection from unmanned surface vehicles. *IEEE transactions on cybernetics*, 46(3), 641-654.
- [Liu et al., 2016] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.
- [Leela et al, 2020] Leela, S. J., Roh, M. I., & Ohb, M. J. (2020). Image-based ship detection using deep learning. *Ocean Systems Engineering*, 10.
- [Maire et al., 2013] Maire, F., Mejias, L., Hodgson, A., & Duclos, G. (2013, November). Detection of dugongs from unmanned aerial vehicles. 2013 IEEE. In *RSJ International Conference on Intelligent Robots and Systems* (3-8).
- [Marques et al, 2015] Marques, M. M., Dias, P., Santos, N. P., Lobo, V., Batista, R., Salgueiro, D., ... & Taiana, M. (2015, May). Unmanned Aircraft Systems in Maritime Operations: Challenges addressed in the scope of the SEAGULL project. In *OCEANS 2015-Genova* (pp. 1-6). IEEE.
- [Moosbauer et al., 2019] Moosbauer, S., Konig, D., Jakel, J., & Teutsch, M. (2019). A benchmark for deep learning based object detection in maritime environments. In *Proceedings of the IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition Workshops* (pp. 0-0).
- [Munin et al., 2023] Munin, A., Folarin, A., Munin-Doce, A., Alonso-Garcia, L., Diaz-Casas, V., Ferreno-Gonzalez, S., & Ciriano-Palacios, J. M. Real Time Vessel Detection Model Using Deep Learning Algorithms for Controlling a Barrier System. Available at SSRN 4423353.
- [Ni et al., 2020] Ni, J., Chen, Y., Chen, Y., Zhu, J., Ali, D., & Cao, W. (2020). A survey on theories and applications for self-driving cars based on deep learning methods. *Applied Sciences*, 10(8), 2749.
- [OpenAI, 2023]. GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf>
- [Padilla et al., 2020] Padilla, R., Netto, S. L., & Da Silva, E. A. (2020, July). A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)* (pp. 237-242). IEEE.
- [Prasad et al., 2016] Prasad, D. K., Prasath, C. K., Rajan, D., Rachmawati, L., Rajabally, E., & Quek, C. (2016). Challenges in video based object detection in maritime scenario using computer vision. *arXiv preprint arXiv:1608.01079*.
- [Prasad et al., 2018] Prasad, D. K., Prasath, C. K., Rajan, D., Rachmawati, L., Rajabally, E., & Quek, C. (2018). Object detection in a maritime environment: Performance evaluation of background subtraction methods. *IEEE Transactions on Intelligent Transportation Systems*, 20(5), 1787-1802.
- [Ren et al., 2015] Faster, R. C. N. N. (2015). Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 9199(10.5555), 2969239-2969250.
- [Sun et al., 2022] Sun, Y., Su, L., Luo, Y., Meng, H., Li, W., Zhang, Z., ... & Zhang, W. (2022). Global Mask R-CNN for marine ship instance segmentation. *Neurocomputing*, 480, 257-270.
- [Terven et al., 2023] Terven, J., & Cordova-Esparza, D. A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. *arXiv* 2023. *arXiv preprint arXiv:2304.00501*.
- [Teutsch & Krüger, 2010] Teutsch, M., & Krüger, W. (2010, November). Classification of small boats in infrared images for maritime surveillance. In *2010 International WaterSide Security Conference* (pp. 1-7). IEEE.
- [Tian et al., 2021] Tian, L., Cao, Y., He, B., Zhang, Y., He, C., & Li, D. (2021). Image enhancement driven by object characteristics and dense feature reuse network for ship target detection in remote sensing imagery. *Remote Sensing*, 13(7), 1327.
- [Tom et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901./
- [Wang et al., 2022] Wang, X., Zhao, F., Lin, P., & Chen, Y. (2022). Wang, X., Zhao, F., Lin, P., & Chen, Y. (2022). Evaluating computing performance of deep neural network models with different backbones on IoT-based edge and cloud platforms. *Internet of Things*, 20, 100609.
- [Westall et al, 2009] Westall, P., O'Shea, P., Ford, J. J., & Hrabar, S. (2009, June). Improved maritime target tracker using colour fusion. In *2009 International Conference on High Performance Computing & Simulation* (pp. 230-236). IEEE.
- [Yan et al., 2022] Yan, T., Sun, W., & Cui, K. (2022, August). Real-time Ship Object Detection with YOLOR. In *Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning* (pp. 203-210).
- [Zhang et al. 2021] Zhang, R., Li, S., Ji, G., Zhao, X., Li, J., & Pan, M. (2021). Survey on deep learning-based marine object detection. *Journal of Advanced Transportation*, 2021, 1-18.

# Towards the Use of Computer Vision Techniques on Streetscape Imagery to Empower Citizens in the Planning Enforcement Process

Sam Lynch and Paul Cuffe

*School of Electrical and Electronic Engineering, University College Dublin*

## Abstract

Urban streetscapes are often cluttered with intrusive advertising signage, which is typically erected without appropriate planning permission. This paper proposes the deployment of computer vision techniques to automatically identify this type of illicit signage within geotagged and timestamped digital images taken of an urban streetscape from a moving vehicle. Such object detection can underpin a semi-automated workflow for instigating planning enforcement complaints against offending signage at scale. The proposed method adapts deep learning models for object detection on a manually collated and labelled dataset of 1051 images containing illegal advertising signage. The system is evaluated on a batch of acquired streetscape images collected from various urban areas in Dublin, Ireland. These early results the broad feasibility of automatically detecting non-compliant vinyl banners and property signs. The main research contribution of this paper is illustrating the potential for computer vision techniques to mediate new relationships between citizens and local authorities.

**Keywords:** Machine Learning (ML), Deep Learning (DL), Single-Shot Detector (SSD)

## 1 Introduction

Ever since the invention of the printing press, outdoor advertising has been widespread in cities. This can greatly harm the visual amenity of urban spaces: such advertisements have been termed as *visual pollutants*. Research shows that urban spaces free of visual pollution increase the quality of life of people living in that environment [Voronych, 2013], while, also, inspiring a sense of pride and belonging in their area [Jensen et al., 2014].

The regulation of outdoor advertising varies from country to country; for example, in Southeast Asia there is little to nothing public administrations can do about what is built or assembled in urban spaces [Jana and De, 2015]. In many developed nations, such as in Ireland, to erect an advertising sign, one must obtain planning permission from the relevant local authority. Notwithstanding this, though, it is common sight to see large vinyl advertising banners hung on and around commercial premises, typically without any such planning permission.

Likewise, ubiquitous *For Sale* realty signage typically exceeds the modest dimensions allowed for small developments that are exempt from having to seek planning permission. For instance, typical requirements state that planning-exempt property signs are “*subject to a maximum area of 0.6sqm for a house and 1.2sqm for any other structure/land. There must be one sign only and it must be removed no later than 7 days after the sale/letting.*”.



Figure 1: Unauthorised advertising signage in Dublin, which cheapens the visual amenity of this otherwise pleasant canal bank setting

This research aims to tackle the problem of visual pollution by facilitating the reporting of these signs to authorities with greater scale and speed. In the Irish example, Local Authorities are bound by the *Planning & Development Act, 2000* (section 152) to investigate all written allegations of unauthorised developments within six weeks of receiving same. This project aims to deploy modern image processing techniques to activate this legal provision at much greater scale, by automating the creation of written complaints against specific planning breaches.

The ambition of this project is to deploy modern computer vision techniques on batches of streetscape imagery to mediate a new relationship between citizens and the planning enforcement officers within Local Authorities who are legally empowered to act against unauthorised advertising displays.

## 1.1 State of the Art

As computing hardware and computing techniques improved as costs went down through the 1990s, developments in subsets of machine learning such as convolutional neural networks (CNNs) and deep learning, which are inspired by neurons in the human brain, improved. A major breakthrough in computer vision occurred in 2012 with the release of a deep neural network called AlexNet [Krizhevsky et al., 2012].

The rapid evolution of such computer vision techniques has seen the technology being utilised extensively in many sectors. Impressive applications have been seen in transportation [Cao et al., 2021], security [Moeslund and Granum, 2001], automatic speech recognition [Yu and Deng, 2016] and agriculture [Bhargava and Bansal, 2021].

Early research in advertising detection used computer vision techniques to detect billboards in sports TV. For instance [Watve and Sural, 2008], where soccer videos were deterministically processed to detect pitch-side advertisement billboards. This paper used Hue slicing and canny edge detection to find a region of interest where the billboard lies in a shot.

The problem with edge and colour detection techniques is that all edges of the signage must visible and the background must be plain with a contrasting colour to the sign. In a more complex environment, such as at street-level, a supervised machine learning approach is required for advertisement detection. Modern research has shown that CNNs produce accurate results for outdoor image classification tasks such as advertising billboard detection. Architectures within this space include AlexNet, VGG, GoogLeNet, NiN, DenseNet and ResNet [Alom et al., 2018].

Work in [Rahmat et al., 2019] retrained AlexNet for billboard detection. This paper successfully made use of the technique of transfer learning, where the weights and biases of pre-trained models can be transferred and used as initial weights and biases for a new (unseen) dataset. Work in [Moffett et al., ] used streetscape imagery for automated detection of tobacco advertising. This research used Faster-RCNN (Region-based CNN) [Chen and Gupta, 2017] to detect the tobacco advertisements and this produced good results.

Work in [Bochkarev and Smirnov, 2019] deals with the problem of illegal advertising on building façades. This research focused on the regional-level laws that dictate the conditions for advertising on building façades in Saint-Petersburg, developing a set of checks for detected advertising objects to check their legality. Thus, although the motivation of the paper is the same as the present one it few transferable insights on how to tackle the problem. Work in [Jiang et al., 2020] developed an algorithm to automatically detect illegal billboards in Hebei Province. This paper finds that a Faster R-CNN detection framework is successful for detecting illegal billboards.

A very attractive feature of CNNs for object detection is transfer learning, where a pre-trained neural network model can be repurposed for a new image classification task with a new dataset. There exists a range of mature neural network models that can be re-trained on a new dataset using an open-source framework such as TensorFlow Object Detection API [Abadi et al., 2016]. The two major architectural types of object detection model are: one-stage detection, such as You Only Look Once (YOLO) [Redmon et al., 2016], and Single Shot Detection (SSD) [Liu et al., 2015]; and two-stage detection models, such as Faster R-CNN [Ren et al., 2015]. The key difference between the two types of models is that in the two-stage detection model, the region of interest is determined first and detection is performed on the region of interest only, resulting in a slower but

generally more accurate process that is computationally expensive.

In a recent study comparing the SSD model and the YOLO model [Ángel Morera et al., 2020] propose a robust method for the automatic detection of urban advertising panels in outdoor images. Their work usef a Single Shot Multiblox Detector (SSD) for advertising detection. Their paper states that the SSD model is most suited for this type of application as it does not re-sample features for bounding box hypotheses

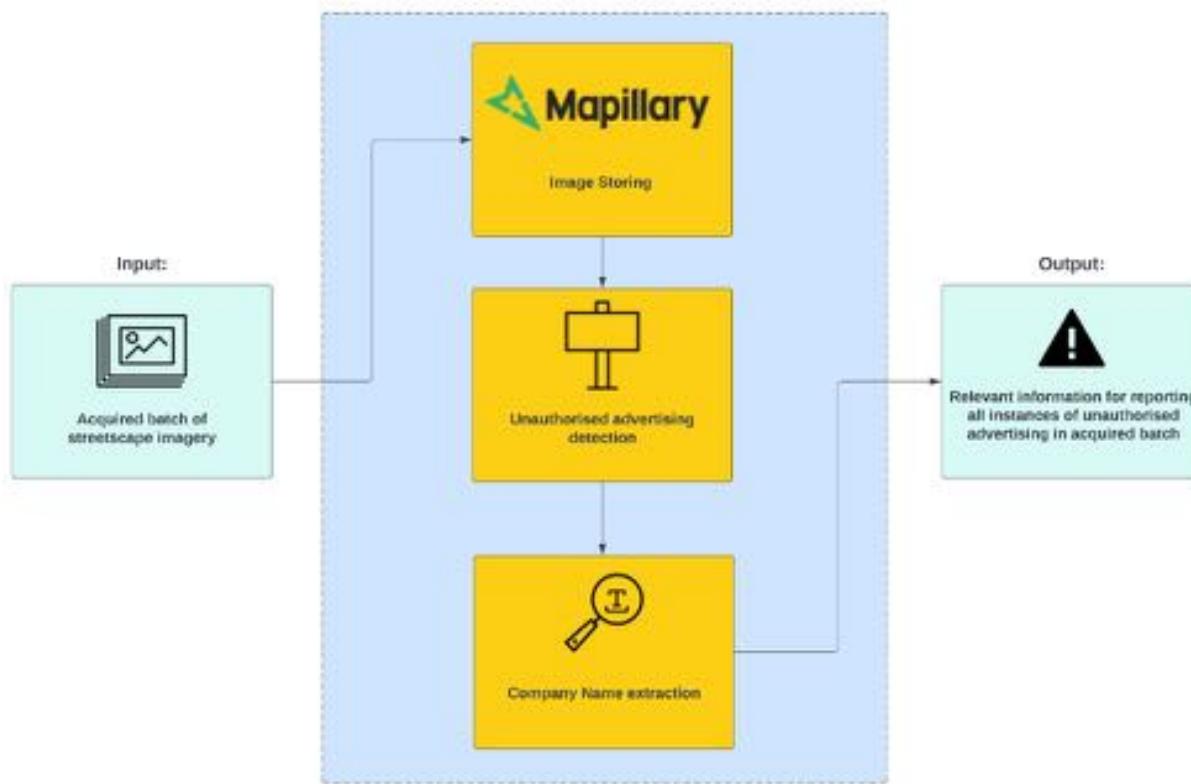


Figure 2: Broad flowchart of developed algorithm

Text extraction, to produce semantic data on the content of advertisements in scene images presents a more complex problem. Work in [Zhou et al., 2017] presents a fast and accurate solution for the detection of text in scenery with the EAST (Efficient and Accurate Scene Text) detector. With the regions of text in an image detected, the characters of the text in the region must be recognised to reveal what the text says. The Tesseract OCR engine [Smith, 2007] is used for the text recognition in this research.

## 2 Methodology

### 2.1 Image Acquisition

As outlined in figure 2, the input to the algorithm is a batch of streetscape imagery which is acquired on a drive through a city. The images acquired are stored in a suitable hosting infrastructure. This implementation utilises the open-source hosting infrastructure Mapillary. This platform is selected as it allows users to freely and conveniently upload images to Mapillary, where they are geotagged and stored with their street address (see a comparative analysis in [Mahabir et al., 2020]). Through interactions with the Mapillary API, the longitude and latitude coordinates of a queried image may be obtained. This research uses interactions through Python to obtain the coordinates of an image where a potentially unauthorised advertisement is situated.

## 2.2 Unauthorised Advertising Detection

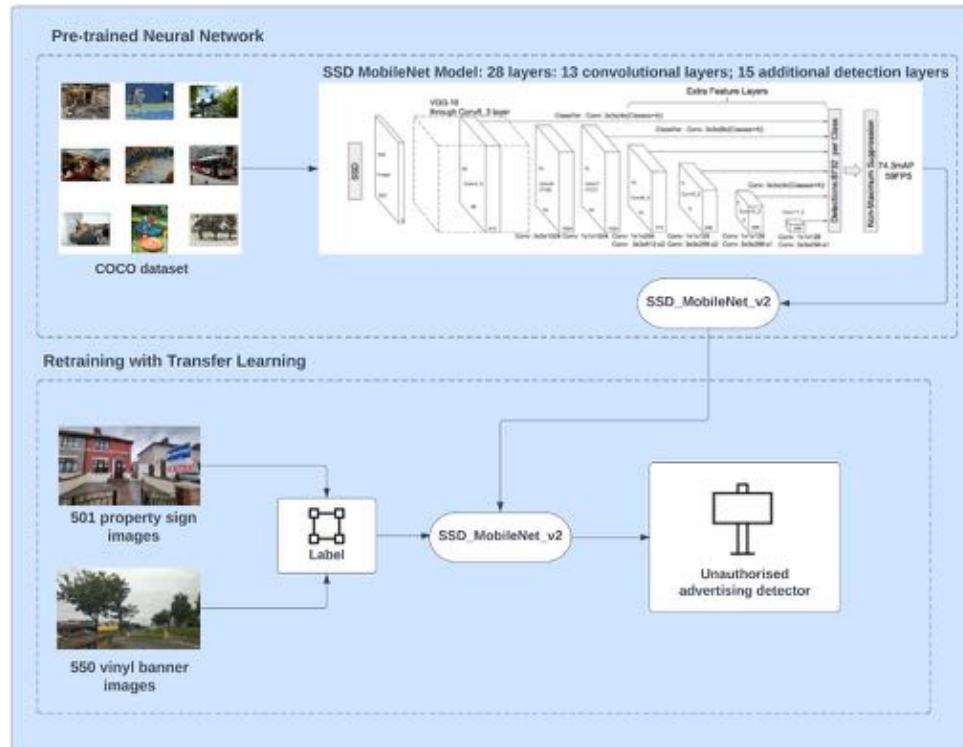


Figure 3: Process of transfer learning utilised to train unauthorised advertising detection model

### 2.2.1 Dataset Acquisition and Image Tagging

There are no public datasets of outdoor unauthorised advertising. Therefore, one task of this research endeavour was to create a suitable dataset to train the object detection model. It is very important to obtain a large and diverse set of training images for each class that are to be detected, to ensure that the object detection model performs well on new, unseen, images.

The classes for detection in this research are the types of unauthorised developments discussed previously: vinyl banners and property signs. Addressing the former, 550 images with vinyl banners were acquired from Visual Genome's [Krishna et al., 2017] VG\_100k dataset and the Mapillary Vistas dataset [Neuhold et al., 2017]. Acquiring the set of property sign images was more challenging as the property signs of interest are much less frequently found in the Mapillary Vistas dataset than vinyl banners. 501 images with property signs were acquired from Google and through the Mapillary API using Python, by filtering for advertisement detections. Combining these gives a total dataset with 1051 images.

Each image was labelled with object bounding boxes in Pascal VOC format by use of the open-source graphical image annotation tool LabelImg [Tzutalin, 2015]. The dataset was, then, split randomly into training images and test images with 80/20 Training/Test split.

### 2.2.2 Object Detection Model

The pre-trained SSD MobileNetV2 is implemented for this object detection task, on the TensorFlow Object Detection API framework. The MobileNet [Howard et al., 2017] is a object detector released in 2017 as an efficient CNN architecture designed for mobile and embedded vision application, it is chosen for this task due



Figure 4: Examples of model performance on test images

to its lean network and high accuracy. SSD , introduced by Liu et al. [Liu et al., 2015], is based on a feed-forward convolutional network that scores bounding box regions based on a confidence score for the presence of the object class in that box. This is followed by a non-maximum suppression step to produce the detections. The architecture of an SSD network uses two steps for object detection: (1) extraction of feature maps; (2) application of convolution filters to detect objects. For the extraction of feature maps, SSD uses the very deep convolutional network for image recognition VGG16, the SSD MobileNetV2 model uses the MobileNetV2 deep CNN. The object detection is carried out using the Conv4\_3 layer. Each prediction from the Conv4\_3 layer comprises of a bounding box and 21 scores for each object class (including no class); then, simply, the class with the highest confidence score is selected as the bounded object class. Conv4\_3 makes a total of  $38 \times 38 \times 4$  predictions: four predictions per cell.

### 2.2.3 Training the Model

The TensorFlow Object Detection API was used to re-train the SSD MobileNetV2 model, in Python. TensorFlow's object detection pipeline includes a configured file with configuration selection optimised for efficient model training. The parameters to be chosen were batch size and number of training epochs. 12 was chosen as the default batch size and 50,000 was chosen as the number of training epochs. The configuration file was updated to include the path to the label map, and the training and test TFRecords. The model was trained on Google Colab and exported to a local machine once training was complete.

### 2.2.4 Company Name Extraction

Text detection and recognition is carried out on the cropped detections from the unauthorised advertising detector. From the recognised text, the name of the company responsible for the erection of the sign can be extracted manually.

## 2.3 Collation of Output Information For Reporting At Scale

The output of the developed algorithm in this research is the relevant information required to instigate the removal of any instances of unauthorised advertising detected in the acquired input sequence. Such information is: a clear image with bounding boxes around the unauthorised advertising detected; cropped image(s) of the unauthorised advertising detected in the image; the address where the image is taken; and, where possible, the name of the company responsible for erecting the advertising. Local authorities outline enforcement measures for the removal of unauthorised advertising from both private and public areas, for example in the Dublin City Development Plan 2022-2028 [Dev, 2022]. Thus, the findings gathered from the algorithm can form many emails of complaint to the local authority reporting all detected instances and requesting their removal.

The output findings may include multiple detections of the same sign and may contain some incorrectly detected instances, e.g. square road signs. A script was written to automatically filter out any duplicate detections of the same sign. If any two detections are found in images taken closer than a pre-defined threshold of 5 metres, calculated by subtracting the longitude and latitude coordinates, the histograms of the cropped

images are compared using OpenCV's correlation comparison method. If the histograms are found to be sufficiently similar, meaning the same sign is pictured in both, the detection with higher confidence score is kept and the other detection is discarded.

The output detections should, then, be manually parsed to filter out any duplicate detections missed and any incorrect detections to eliminate the risk of redundant emails of complaint.

### 3 Results

The scripts implementing the functionality in Fig. 2, and the training image set, are available in an enduring online repository at [Cuffe and Lynch, 2023].

An example batch of streetscape imagery was acquired by use of the Mapillary app on a drive of approx. 16 km<sup>2</sup> through South County Dublin <sup>1</sup>. This batch contained 644 images.

The acquired imagery was uploaded to Mapillary where geotagging and privacy blurring was applied to each image. The computer vision scripts were executed locally against these with a confidence threshold of 60% set for detections.

After duplicates were automatically detected and filtered out, the output contained 77 potential instances of property signs and 45 potential instances of vinyl banners.

As shown in Table 1, analysis of this test acquisition shows that 56 property signs were correctly detected and classified, 21 instances were incorrectly classified as property signs and 8 property signs were not detected. This means that the algorithm precision for property signs is 72.7% and more importantly, due to the manual filtering of the output, the recall for property signs is 87.5%.

30 vinyl banners were correctly detected and classified, 15 instances were incorrectly classified as vinyl banners and 7 vinyl banners were not detected. Resulting in an algorithm precision for vinyl banners of 66.67% and more importantly a recall for vinyl banners of 81.1%.

	<b>True Positives</b>	<b>False Positives</b>	<b>False Negatives</b>	<b>Precision</b>	<b>Recall</b>
<b>Property Signs</b>	56	21	8	72.7%	87.5%
<b>Vinyl Banners</b>	30	15	7	66.67%	81.1%

Table 1: Property sign and vinyl banner algorithm results summary for test acquisition

### 4 Conclusion

This research describes a tool to empower ordinary citizens to initiate enforcement actions at scale against the proliferation of unauthorised advertising signage. The results show that the object detection model can accurately detect and locate instances of property signs and vinyl banners in streetscape images. The test acquisition obtained, surveyed an area of approximately 16 km<sup>2</sup>. In this batch of



Figure 5: Map of the area surveyed with pins depicting location of each potential property sign detected by algorithm

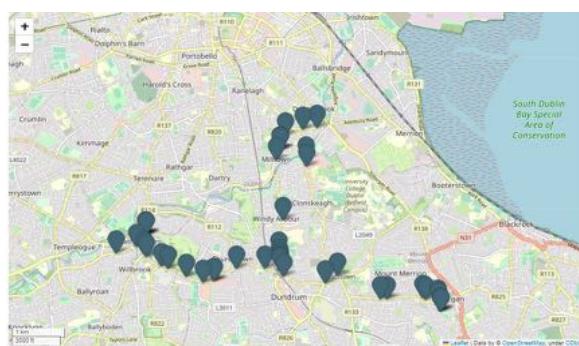


Figure 6: Map of area surveyed with pins depicting location of each potential vinyl banner detected by algorithm

<sup>1</sup>The journey can be viewed online on Mapillary itself

imagery the tool presented in the research detected 87.5% of property signs, 56 in total, and 81.1% of vinyl banners, 30 in total, in the area, which highlights the potential scale of the tool.

This tool allows citizens to have a considerable impact on the protection of the visual beauty of their city by reporting unauthorised advertising at scale. Before this research, the only way to report unauthorised advertising was to take a photo of any potentially unauthorised advertising witnessed, and report it in an email, manually recording the address of the advertising and the time the photo was taken. Consequently, reporting instances at a large scale is a very labour intensive process which is very time-consuming and impractical to execute. The automated workflow presented in this research ensures that the manual effort involved in the reporting of an instance has decreased significantly, due to automatic geotagging, detection and timestamping.

## References

- [Dev, 2022] (2022). Dublin City Development Plan 2022-2028 Volume 2: Appendices.
- [Abadi et al., 2016] Abadi, M. et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467*.
- [Alom et al., 2018] Alom, M. Z., Taha, T., Yakopcic, C., Westberg, S., Hasan, M., Esesn, B., Awwal, A., and Asari, V. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.
- [Bhargava and Bansal, 2021] Bhargava, A. and Bansal, A. (2021). Fruits and vegetables quality evaluation using computer vision: A review. *Journal of King Saud University-Computer and Information Sciences*, 33(3):243–257.
- [Bochkarev and Smirnov, 2019] Bochkarev, K. and Smirnov, E. (2019). Detecting advertising on building façades with computer vision. volume 156, pages 338–346. Elsevier B.V.
- [Cao et al., 2021] Cao, J., Song, C., Song, S., Xiao, F., Zhang, X., Liu, Z., and Ang, M. H. (2021). Robust object tracking algorithm for autonomous vehicles in complex scenes. volume 13. MDPI AG.
- [Chen and Gupta, 2017] Chen, X. and Gupta, A. (2017). An Implementation of Faster RCNN with Study for Region Sampling.
- [Cuffe and Lynch, 2023] Cuffe, P. and Lynch, S. (2023). Scripts and data from “Towards the Use of Computer Vision Techniques on Streetscape Imagery to Empower Citizens in the Planning Enforcement Process”. *DOI: 10.6084/m9.figshare.23703423.v1*.
- [Howard et al., 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
- [Jana and De, 2015] Jana, M. K. and De, T. (2015). Visual pollution can have a deep degrading effect on urban and suburban community: a study in few places of Bengal, India, with special reference to unorganized billboards. *European Scientific Journal*.
- [Jensen et al., 2014] Jensen, C. U., Panduro, T. E., and Lundhede, T. H. (2014). The vindication of Don Quixote: The impact of noise and visual pollution from wind turbines. *Land economics*, 90(4):668–682.
- [Jiang et al., 2020] Jiang, X.-H., Feng, H.-L., and Dong, Y.-J. (2020). Application of Neural Network in Image Detection of Illegal Billboards.

- [Krishna et al., 2017] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- [Liu et al., 2015] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2015). SSD: Single Shot MultiBox Detector.
- [Mahabir et al., 2020] Mahabir, R., Schuchard, R., Crooks, A., Croitoru, A., and Stefanidis, A. (2020). Crowdsourcing street view imagery: A comparison of mapillary and openstreetcam. *ISPRS International Journal of Geo-Information*, 9(6):341.
- [Moeslund and Granum, 2001] Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81(3):231–268.
- [Moffett et al., ] Moffett, C., Abunku, P., Zhu, J., Chaturvedi, I., Chen, G., and Dobler, G. Capstone Project Report CUSP NYU 2018 Automated Detection of Street-Level Tobacco Advertising Displays.
- [Neuhold et al., 2017] Neuhold, G., Ollmann, T., Bulo, S. R., and Kortschieder, P. (2017). The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. volume 2017-October, pages 5000–5009. Institute of Electrical and Electronics Engineers Inc.
- [Rahmat et al., 2019] Rahmat, R. F., Dennis, Sitompul, O. S., Purnamawati, S., and Budiarto, R. (2019). Advertisement billboard detection and geotagging system with inductive transfer learning in deep convolutional neural network. *Telkomnika (Telecommunication Computing Electronics and Control)*, 17:2659–2666.
- [Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- [Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- [Smith, 2007] Smith, R. (2007). An Overview of the Tesseract OCR Engine.
- [Tzutalin, 2015] Tzutalin, D. (2015). LabelImg. *GitHub repository*, 6.
- [Voronych, 2013] Voronych, Y. (2013). Visual pollution of urban space in lviv. *Przestrzeń I Forma*, (20):309–314.
- [Watve and Sural, 2008] Watve, A. and Sural, S. (2008). Soccer video processing for the detection of advertisement billboards. *Pattern Recognition Letters*, 29:994–1006.
- [Yu and Deng, 2016] Yu, D. and Deng, L. (2016). *Automatic speech recognition*, volume 1. Springer.
- [Zhou et al., 2017] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., and Liang, J. (2017). EAST: An Efficient and Accurate Scene Text Detector.
- [Ángel Morera et al., 2020] Ángel Morera, Ángel Sánchez, Moreno, A. B., Ángel D. Sappa, and Vélez, J. F. (2020). SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities. *Sensors (Switzerland)*, 20:1–23.

# Feature Based Approaches for Homography Estimation

Samuel Venezia<sup>1</sup>, Sonya Coleman<sup>1</sup>, Dermot Kerr<sup>1</sup>, and John Fegan<sup>2</sup>

<sup>1</sup> School of Computing, Engineering and Intelligent Systems, Ulster University, Londonderry

<sup>2</sup> Metro Surveillance Group LTD, Cookstown

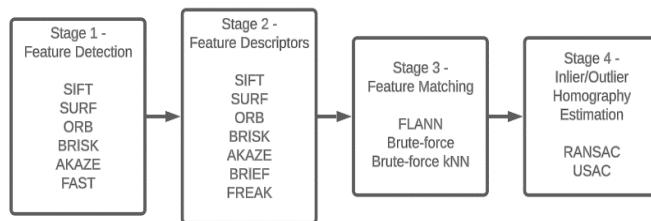
## Abstract

Image stitching is a method of producing a wider field of view by combining several overlapping images. With four main stages in the image stitching process, the algorithms used at each stage can have a dramatic impact on the success of stitching an image. For each stage, there are a wide range of algorithms to choose from and it can be a challenge to identify a stitching pipeline that will produce the best results. In this paper, we study the approaches involved in each of the four stages of image stitching. A real-world dataset is utilised to evaluate each algorithm, where images are transformed to different perspectives. The similarities of these images are compared to a warped perspective image obtained using the homographies provided by the dataset. The pipelines tested were limited to producing accurate results up to and including a 50° perspective change. Pipelines utilising BRISK's feature detector, FREAK, and Brute Force produced significant results. However, pipelines incorporating ORB, FAST, or BRIEF produce poor results when compared to other feature detection and feature description algorithms. Generally, the ratio test hindered the matched pairs process, although there were exceptions. Finally, the inlier/outlier detection algorithms, USAC and RANSAC, had similar performances with no definitive data to suggest that, in general, one outperforms the other.

**Keywords:** Image stitching, Perspective Warping, Feature Mapping, Homography estimation

## 1 Introduction

Computer Vision (CV) applications such as autonomous vehicles, security systems, and sports analytics benefit from an image with a wide field of view (FOV) and high resolution. However, obtaining an image with a wide FOV and high resolution can be a challenge. A wide-angle lens can replace a standard lens to fix this, however this distorts the image, reducing its quality. Moreover, the image is still restricted to the original resolution of the capture device. Image stitching is a technique that produces panoramic images without the use of a wide-angle lens and can be done with either moving or static cameras. Merging overlapping images of a scene through image stitching provides a high-resolution image with a broad field of view without the distortions of wide-angle lenses. This provides a viable method for creating a wider FOV [Wang and Yang, 2020]. Feature-based approaches to image stitching include feature detection, feature description, feature matching, outlier removal, homography estimation, and transformation of pixels to the stitched image [Jakubović and Velagić, 2018]. This paper evaluates multiple algorithms across four stages of feature-based image stitching, as listed in Figure 1, determining the techniques that produce the highest quality results.



**Figure 1: Image Warping Pipeline and Algorithms**

## 1.1 Feature Detectors and Descriptors

Feature detectors search for distinctive visual components within an image. Figure 1's Stage 1 lists several feature detector algorithms. When matching features appear in multiple images of the same scene, they can be used to locate overlapping areas and stitch the images together. Feature descriptors represent the detected feature's local pixel neighbourhood. A unique description of a feature is provided by descriptors, enabling feature matching across different images. Stage 2 of Figure 1 outlines several feature descriptor algorithms.

A ubiquitous feature detection and descriptor approach is the Scale Invariant Feature Transform (SIFT) [Lowe, 2004]. The SIFT feature detector utilises an image pyramid representation to provide invariance to scale changes [Leutenegger et al., 2011]. SIFT identifies features by using the Difference of Gaussians (DoG) approach to detect gradient changes and the Hessian to reject edge points. The SIFT descriptor utilises a histogram of local and global gradients to describe the neighbourhood of pixels around each detected feature. While SIFT has been considered one of the benchmark approaches to feature detection and description it has a disadvantage in that it requires significant computation.

As an alternative the Speeded-Up Robust Features (SURF) approach [Bay et al., 2006] uses a different image representation technique to reduce the computation. Rather than using an image pyramid and applying DoG, SURF uses Haar wavelet-based image filters which are applied at different scales across the image. SURF also includes a feature descriptor utilising a Haar wavelet distribution.

KAZE is an alternative feature detector and descriptor which takes inspiration from SIFT and SURF [Leutenegger et al., 2011]. It utilises a similar process for detecting features as SIFT while using a homogeneous process for computing descriptors as SURF. A variant of KAZE, Accelerated-KAZE (AKAZE), reduces the time needed to detect the features within the pyramidal framework by implementing the Fast Explicit Diffusion framework [Alcantarilla et al., 2013].

The Features from Accelerated Segment Test (FAST) corner detector [Trajković and Hedley, 1998] is a fast but stable corner detection algorithm able to identify corners with high accuracy. The Binary Robust invariant scalable keypoints (BRISK) feature detector improves upon FAST [Leutenegger et al., 2011], providing the efficiency of FAST whilst being invariant to scale and rotation due to the use of a pyramidal structure similar to that of SIFT. The BRISK feature descriptor uses a deterministic sampling pattern to ensure uniform density around the keypoint (an especially distinctive feature often invariant to image transformations) and retrieve its direction to maintain rotational invariance [Leutenegger et al., 2011]. By using fewer sampling points than bitwise comparisons, the feature descriptor lowers the number of comparisons and complexity [Leutenegger et al., 2011].

The Binary Robust Independent Elementary Features (BRIEF) is an alternative feature descriptor with an emphasis on reducing the memory consumption of the algorithm [Calonder et al., 2010]. Descriptors are computed from images by directly comparing intensities of point pairs using intensity difference tests.

The improved variation of FAST, orientated FAST (oFAST), incorporated an orientation operator using the intensity centroid approach [Rosin, 1999]. This enabled more information detailing the keypoint's orientation to be utilised by the descriptor. The oFAST and Rotated Brief (ORB) algorithm is based on FAST and the BRIEF descriptor [Calonder et al., 2010]. Rotated BRIEF (rBRIEF) was proposed [Rublee et al., 2011] as a modified version of BRIEF, aiming to reduce its sensitivity to in-plane rotations.

The conclusion drawn from real-world data was that ORB had better performance than SIFT and sometimes SURF when compared [Rublee et al., 2011]. Fast Retina Keypoint (FREAK) is an alternative feature descriptor inspired by the human visual system [Alahi et al., 2012]. Like BRISK, FREAK utilises a retinal circular sampling grid. Compared to BRISK, FREAK has a higher density of points that are closer to the centre [Alahi et al., 2012].

## 1.2 Feature Matchers

The feature matching methods, seen in Stage 3 of Figure 1, utilise the descriptors generated from the 2<sup>nd</sup> stage feature descriptors as seen in Figure 1. With descriptors acting as a unique signature of each feature point, enabling points across both images to be compared and matched. The Fast Library for Approximate Nearest Neighbours

(FLANN) and Brute-force (BF) are feature matchers, both utilising Euclidean distance to measure the similarity between feature points on each image [Noble, 2016, Muja and Lowe, 2014]. The BF approach compares descriptors of all features in one image with those in another image [Noble, 2016]. While FLANN utilises approximations comparing few features by utilising the k-Nearest Neighbours (kNN) algorithm to produce pairs faster than BF but at the cost of reduced accuracy [Muja and Lowe, 2014]. Previous research has suggested utilising BF with the kNN ratio test [Jakubović and Velagić, 2018], reduces the number of possible matches in the event of multiple points competing for a match.

### 1.3 Outlier Removal for Homography Estimation

Homography is a 2-D perspective transformation that aims to map the pixels in a source image to a destination image and as such has a direct application in image stitching. Algorithms, such as Random Sample Consensus (RANSAC), Universal RANSAC (USAC), and Progressive Sample Consensus (PROSAC), are applied in Stage 4 as seen in Figure 1 for the identification of inlying matched pairs and the removal of outliers [Raguram et al., 2013]. RANSAC has been commonly employed for classifying between inliers and outliers [Caparas, 2020, Jakubović and Velagić, 2018, Tong et al., 2021]. USAC, an alternative to RANSAC, produced results that are more in line with the ground truth when compared to RANSAC and PROSAC, a substitute algorithm that favours speed over accuracy [Raguram et al., 2013]. Inliers are used to estimate homography, which generates a matrix for mapping an image to a different perspective [Sharma and Jain, 2020].

### 1.4 State-of-the-Art Approaches

AKAZE is considered a state-of-the-art feature detector and descriptor algorithm, outperforming SIFT, SURF, ORB, and BRISK using a variety of datasets [Sharma and Jain, 2020, Tong et al., 2021]. While FREAK is a state-of-the-art feature descriptor, outperforming the feature descriptors SIFT, SURF, and BRISK [Alahi et al., 2012]. Meanwhile, within the matching stage, BF is state-of-the-art, producing more accurate descriptor pairings than FLANN, while incorporating kNN increases BF's effectiveness [Noble, 2016, Caparas, 2020]. USAC is a state-of-the-art algorithm for removing outlying matched pairs, as it produced more accurate results compared with RANSAC and PROSAC [Raguram et al., 2013].

Multiple standard lens cameras with overlapping FOVs can be used for applications like sports analytics to create a single-perspective view and enhance resolution through image stitching. Fast image stitching is better achieved with feature-based methods instead of end-to-end deep learning approaches. The latter requires re-running the entire stitching process for each frame, due to its architecture [Yi et al., 2016].

## 2 Methodology

Our overall approach is based on defining an image stitching pipeline with multiple algorithms available for each stage. Our aim is to compare algorithms for each stage of image stitching to establish the most effective pipeline. Figure 1 shows this pipeline, with each stage having multiple algorithms. By utilising the feature detectors of Stage 1 such as SIFT, ORB, BRISK, AKAZE, and FAST, we can determine the most effective algorithm for detecting keypoints in each image. Due to licensing restrictions, testing SURF will not be possible, so it will be left out of the method and results. Comparing corner detectors, such as FAST with its variants (ORB and BRISK) to pyramidal-based algorithms like SIFT or AKAZE, will determine the better feature detector.

The feature descriptors of Stage 2 are compared against each other to determine the best method for creating descriptors given the keypoints identified in Stage 1. The feature descriptors of SIFT, ORB, BRISK, and AKAZE are compared alongside the stand-alone feature descriptors of BRIEF and FREAK. For example, one pipeline would

use the SIFT feature detector and FREAK descriptor, while another pipeline would use the SIFT feature detector and descriptor.

Stage 3 produces pairs of corresponding keypoints that are within the overlapping regions of the images to be stitched. Because FLANN performed poorly in the past, it was excluded from the data. The algorithms of Stage 3 comprise of BF and BF kNN. These two algorithms are compared against each other, to determine what influence the matching algorithm has on the success of a pipeline. When implementing BF kNN, the threshold is set to 0.75, testing the ratio of distance, enabling additional flexibility when determining the keypoints for matching in each image.

Stage 4 assesses two outlier elimination algorithms. More specifically, a comparison between RANSAC and USAC is being made. The resulting inlying matched pairs will be utilised to estimate the homography.

Due to the focus on perspective transformation, the Oxford Affine Covariant Regions Viewpoint Graffiti (OACRVG) dataset [Mikolajczyk and Schmid, 2005] is utilised to assess the performance of the different algorithms which make up the stitching pipeline. OACRVG is a standard test dataset utilised in multiple image stitching experiments focusing on perspective warping [Calonder et al., 2010, Leutenegger et al., 2011]. The dataset contains six images, each of size 800x640 pixels and includes homographies to map the first image with the other five images, treating image one as a pair with each subsequent image [Mikolajczyk and Schmid, 2005]. Additionally, there is a viewing angle change of 40° between images one and three, with an increase of 10° of change for each subsequent image [Mikolajczyk and Schmid, 2004].

Using four of the best inlier matched pairs from Stage 4 of the pipeline, the homography is estimated for each image in the OACRVG dataset being mapped onto image one. Estimating the homography requires the eight degrees of freedom afforded by the four matched pairs to compute the homography. The two images are mapped onto a 2D plane, to enable a common viewpoint between the images. Once the homography is estimated, every image is warped onto image one through a perspective transformation, as illustrated in Figure 2. The homographies provided by the OACRVG dataset were used to create the warped images, which can be seen in Figure 3. We aim to compare the pipelines mentioned and shown in Figure 1 to identify the ones that produce warped images of comparable quality to those produced by the homographies available in the dataset. Processing pipelines and algorithms are implemented using OpenCV 4.6.0 [Bradski, 2000] and Numpy 1.24.2 [Van Der Walt et al., 2011].

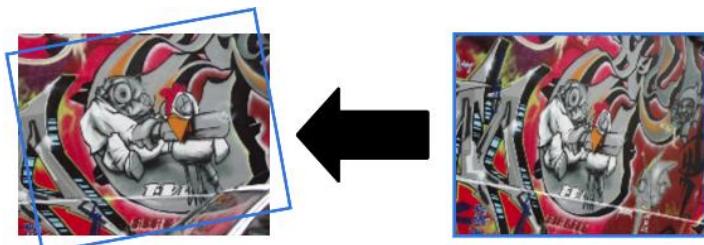


Figure 2: Perspective Transformation of Image 3 to 1.

Image #	Actual Image	Warped Image
2		
3		
4		
5		
6		

Figure 3: Actual and Expected Warped Images, mapping images 2, 3, 4, 5 and 6 onto image 1 generated using the inverted homographies provided by the OACRVG dataset.

### 3 Results

The Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Feature Similarity Index (FSIM) have previously been utilised to measure performance and concluded that while each method produced consistent results throughout, SSIM and FSIM are normalised and easier to interpret when compared with the absolute errors of MSE and PSNR [Sara et al., 2019]. PSNR and FSIM were chosen here to evaluate the warped images utilising the Image Similarity Measures toolbox [Müller et al., 2020].

Table 1 shows the mean FSIM and PSNR results for each pipeline on each image. FSIM is normalised between 0 and 1 where the higher the value, the more similar the images are [Sharma and Jain, 2020]. FSIM distinguishes complex parts of an image by detecting changes in light and gradient [Sara et al., 2019]. PSNR is the ratio calculated between the highest signal power and the power of the distorted noise [Sara et al., 2019]. Seven pipelines are among the top ten FSIM and PSNR from Table 1. Four of the seven pipelines that performed the best used BRISK, SIFT, and AKAZE detectors and descriptors, and three used BRISK and AKAZE detectors with a FREAK descriptor. Five of the pipelines employ BF, while two incorporate BF kNN as a matcher. RANSAC was utilised in three of the seven pipelines while USAC was utilised in four of the pipelines. While the best performing pipeline utilised USAC, RANSAC performed admirably, taking 2nd and 3rd place in Table 1 (ranked by PSNR).

#	Pipeline	FSIM ↑	PSNR ↑	#	Pipeline	FSIM	PSNR
1	BRISK – FREAK – BF – USAC	<b>0.604</b>	<b>93.605</b>	24	BRISK - BRIEF - BF - USAC	0.472	87.681
2	SIFT – SIFT – BF kNN – RANSAC	<b>0.581</b>	<b>93.326</b>	25	SIFT – FREAK – BF kNN – USAC	0.430	87.086
3	BRISK – FREAK – BF – RANSAC	<b>0.596</b>	<b>93.018</b>	26	SIFT – FREAK – BF – USAC	0.451	87.083
4	AKAZE – AKAZE – BF – USAC	<b>0.560</b>	<b>92.224</b>	26	ORB – BRIEF – BF – USAC	0.445	86.972
5	BRISK – BRISK – BF – RANSAC	<b>0.563</b>	<b>91.749</b>	28	SIFT – FREAK – BF kNN – RANSAC	0.426	86.970
6	AKAZE – FREAK – BF – RANSAC	0.512	<b>91.685</b>	29	BRISK – BRIEF – BF – RANSAC	0.421	86.392
7	SIFT – SIFT – BF kNN – USAC	<b>0.571</b>	<b>91.637</b>	30	AKAZE – BRIEF – BF – RANSAC	0.406	86.279
8	AKAZE – FREAK – BF – USAC	<b>0.556</b>	<b>91.194</b>	31	ORB – BRIEF – BF – RANSAC	0.416	85.945
9	BRISK – FREAK – BF kNN – RANSAC	0.532	<b>90.956</b>	32	SIFT – BRIEF – BF – USAC	0.382	85.405
10	AKAZE – AKAZE – BF – RANSAC	0.519	<b>90.891</b>	33	SIFT – BRIEF – BF – RANSAC	0.384	85.056
11	BRISK – FREAK – BF kNN – USAC	<b>0.547</b>	90.643	34	FAST – FREAK – BF – RANSAC	0.259	82.350
12	SIFT – SIFT – BF – USAC	<b>0.559</b>	90.640	35	BRISK – BRIEF – BF kNN – RANSAC	0.300	81.930
13	AKAZE – FREAK – BF kNN – RANSAC	0.498	90.232	36	FAST – BRIEF – BF – RANSAC	0.288	81.742
14	SIFT – SIFT – BF – RANSAC	<b>0.539</b>	90.208	37	FAST – FREAK – BF kNN – RANSAC	0.267	81.691
15	BRISK – BRISK – BF kNN – USAC	0.515	90.096	38	AKAZE – BRIEF – BF kNN – USAC	0.286	81.515
16	AKAZE – FREAK – BF kNN – USAC	0.504	89.717	39	BRISK – BRIEF – BF kNN – USAC	0.293	81.251
17	ORB – ORB – BF – USAC	0.481	89.459	40	SIFT – BRIEF – BF kNN – RANSAC	0.279	81.146
18	SIFT – FREAK – BF – RANSAC	0.489	88.821	41	AKAZE – BRIEF – BF kNN – RANSAC	0.276	81.085
19	ORB – FREAK – BF – USAC	0.504	88.745	42	FAST – FREAK – BF – USAC	0.264	81.037
20	ORB – ORB – BF – RANSAC	0.467	88.182	43	FAST – FREAK – BF kNN – USAC	0.259	80.427
21	BRISK – BF kNN – RANSAC	0.456	88.165	44	FAST – BRIEF – BF kNN – USAC	0.252	80.296
22	ORB – FREAK – BF – RANSAC	0.485	88.046	45	SIFT – BRIEF – BF kNN – USAC	0.254	80.188
23	AKAZE – BRIEF – BF – USAC	0.464	87.910	46	FAST – BRIEF – BF kNN – RANSAC	0.228	79.720

**Table 1: Combined FSIM and PSNR mean results of the pipelines ranked by PSNR metric.**

As shown in Table 2, the best pipeline from Table 1 was unable to produce a warped image to beyond image 4 (50° rotation), suggesting that, even though the pipeline scored the highest, there could potentially be a restriction in their ability to handle significant perspective transformations. In accordance with the conclusion from [Rublee et al., 2011], the simplicity of FAST hampers its performance. FAST is sub-par, producing the worst results, as it could not determine the orientation of a keypoint. ORB was the least successful out of the comprehensive feature detector and descriptors, with only two warped images generated using BF. This supports the conclusions of [Alahi et al., 2012], while disagreeing with [Rublee et al., 2011], with SIFT outperforming or at least performing at the same level as ORB. The pipelines incorporating BRISK for both feature detection and descriptors performed at the same level as ones that incorporated SIFT and AKAZE for both stages, with BRISK – BRISK – BF – RANSAC, falling within the top five of Table 1. This is in line with the conclusions of [Leutenegger et al., 2011], with BRISK as a feature descriptor performing competitively against SIFT. The BRISK feature detector approach outperforms other FAST corner detection-based algorithms in terms of PSNR and FSIM scores due to the scaling invariance BRISK provides. Four pipelines that use this version of FAST are among the top ten for FSIM and PSNR. This implies that BRISK's feature detector is the most skilled at generating superior keypoints, resulting in a broader image viewpoint.

Image #	2	3	4	5	6
Result					
PSNR	105.263	101.760	100.116	80.386	80.501
FSIM	0.868355	0.822749	0.800981	0.24983	0.276183
Pipeline	BRISK – FREAK – BF – USAC				

**Table 2: Comparison of the best method from Table 1 for each of the images based on the PSNR metric.**

As most of the pipelines involving SIFT and BRISK outperformed their AKAZE counterparts, these findings contrast with [Tong et al., 2021, Sharma and Jain, 2020]. Even though AKAZE yielded higher PSNR and FSIM scores compared to SIFT when BF and RANSAC were used in Stages 3 and 4, it was generally less effective than SIFT or BRISK-based pipelines, indicating that this is not the norm. When comparing BRIEF and FREAK, the dedicated feature descriptors, FREAK, produced higher PSNR and FSIM scores. FREAK is utilised in five of the top ten PSNR mean results, and also utilised in four of the FSIM results. BRIEF's poor performances can be traced to its inadequate performance with in-plane rotations. This result concurs with [Alahi et al., 2012], finding that FREAK outperformed the feature descriptors of BRISK, SIFT, AKAZE, and ORB.

Within the top ten FSIM and PSNR results in Table 1, seven utilised BF, compared with three using BF kNN. Finding fewer but more precisely matched pairs can reduce the possibility that a homography matrix is estimated. However, if there are over four matched pairs, but they are not matched correctly, the homography will be incorrectly calculated, producing an inaccurately warped image. This is not the case for all instances of kNN, with SIFT – SIFT – BF kNN – RANSAC producing a higher PSNR and FSIM score than its BF counterpart, with the former placing 2nd; while the latter performed worse placing 14th in Table 1. This suggests that kNN is preferable when enough features, with corresponding descriptors are detected.

Producing the highest similarity of FSIM and PSNR, BRISK – FREAK – BF – USAC suggests USAC minimised the number of matched pair outliers for estimating the homography, while its RANSAC equivalent pipeline placed 3rd with a slight decrease in overall performance. However, some pipelines utilising RANSAC outperformed their USAC counterparts, as shown with SIFT - SIFT - BF kNN – RANSAC and SIFT - SIFT - BF

kNN – RANSAC, with PSNR scores of 93.326 and 91.637, respectively. This disparity suggests neither algorithm outperforms the other, suggesting that they are of equal standing. The key factors affecting the performances of a warped image are the feature detectors, descriptors, and matching algorithms used in Stages 1, 2, and 3. Thus, it is recommended that pipelines with different algorithms be employed.

## 4 Conclusion

This paper compared the performances of image warping pipelines with 4 stages. The pipelines that used BRISK or SIFT in Stage 1 showed the best results for various perspectives using the OACRVG dataset. Using SIFT, BRISK, and AKAZE in Stages 1 and 2 yielded high PSNR and FSIM scores. Meanwhile, pipelines utilising FREAK perform similarly to or outperformed the other feature descriptors. Concluding that FREAK is a state-of-the-art feature descriptor, supporting the conclusion of [Alahi et al., 2012]. Pipelines with a ratio test (kNN) performed worse than matchers without, although exceptions exist. Furthermore, there is no definitive data to suggest that there is a significant performance increase in USAC compared to RANSAC that can be fully explained by the outlier removal approach, to the contrary of [Raguram et al., 2013]. Notwithstanding the accomplishments of the pipelines, the findings show they can only handle a maximum perspective change of 50°. This could be due to factors such as the dataset's resolution with a smaller overlapping region affecting the ability to warp the image. We intend to develop this research further by applying it to sports analytics. Further testing of the pipelines should involve the use of multiple datasets with different resolutions and environments, such as sports fields. Finally, DL approaches producing a homography matrix may provide an alternative solution to the pipelines discussed in this paper.

## Acknowledgements

This work was funded by a DfE CAST scholarship in collaboration with Metro Surveillance Group LTD.

## References

- [Alahi et al., 2012] Alahi, A., Ortiz, R., and Vandergheynst, P. (2012). FREAK: Fast retina keypoint.
- [Alcantarilla et al., 2013] Alcantarilla, P. F., Nuevo, J., & Bartoli, A. (2013). Fast explicit diffusion for accelerated features in nonlinear scale spaces.
- [Bay et al., 2006] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). SURF: Speeded up robust features. Volume 3951 LNCS.
- [Bradski, 2000] Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.
- [Calonder et al., 2010] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). BRIEF: Binary robust independent elementary features. Volume 6314 LNCS.
- [Caparas, 2020] Caparas, A. (2020). Feature-based Automatic Image Stitching Using SIFT, KNN and RANSAC. International Journal of Advanced Trends in Computer Science and Engineering, 9(1.1 S I).
- [Jakubović and Velagić, 2018] Jakubović, A., and Velagić, J. (2018). Image feature matching and object detection using brute-force matchers. Volume 2018-September.
- [Leutenegger et al., 2011] Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). BRISK: Binary Robust invariant scalable keypoints.

- [Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2).
- [Mikolajczyk and Schmid, 2004] Mikolajczyk, K., and Schmid, C. (2004). Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1).
- [Mikolajczyk and Schmid, 2005] Mikolajczyk, K., and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10).
- [Muja and Lowe, 2014] Muja, M., and Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11).
- [Müller et al., 2020] Müller, M. U., Ekhtiari, N., Almeida, R. M., and Rieke, C. (2020). Super-Resolution of Multispectral Satellite Images Using Convolutional Neural Networks. volume 5.
- [Noble, 2016] Noble, F. K. (2016). Comparison of OpenCV's feature detectors and feature matchers. In 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP), pages 1-6. IEEE.
- [Raguram et al., 2013] Raguram, R., Chum, O., Pollefeys, M., Matas, J., and Frahm, J. (2013). USAC: A Universal Framework for Random Sample Consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):2022-2038.
- [Rosin, 1999] Rosin, P. L. (1999). Measuring Corner Properties. *Computer Vision and Image Understanding*, 73(2).
- [Rublee et al., 2011] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF.
- [Sara et al., 2019] Sara, U., Akter, M., and Uddin, M. S. (2019). Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. *Journal of Computer and Communications*, 07(03).
- [Sharma and Jain, 2020] Sharma, S. K., and Jain, K. (2020). Image Stitching using AKAZE Features. *Journal of the Indian Society of Remote Sensing*, 48(10).
- [Tong et al., 2021] Tong, C., Jianfeng, G., Xueli, X., and Jianxiang, X. (2021). High-precision Image Mosaic Algorithm Based on Adaptive Homography Transform. Volume 2021-July.
- [Trajković and Hedley, 1998] Trajković, M., and Hedley, M. (1998). Fast corner detection. *Image and Vision Computing*, 16[2], 75-87. ID:271526
- [Van Der Walt et al., 2011] Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22-30.
- [Wang and Yang, 2020] Wang, Z., and Yang, Z. (2020). Review on image-stitching techniques. *Multimedia Systems*, 26(4).
- [Yi et al., 2016] Yi, K. M., Trulls, E., Lepetit, V., and Fua, P. (2016). LIFT: Learned invariant feature transform. Volume 9910 LNCS.

# Uncompromising Operator Safety: A Standalone Device Approach for Threat Immunity and Malfunction Prevention through Visual Cognition

Mihai Penica, Eoin O'Connell, Reenu Mohandas, William O'Brien, Martin Hayes

*University of Limerick*

## Abstract

The emergence of Industry 4.0 has generated significant attention regarding the mitigation of security breaches and ensuring operator safety within industrial settings. This scholarly article addresses the concept of cognitive interaction within Industry 4.0 frameworks and elucidates the pivotal role of a stand-alone system deployed in industrial environments. The system in question leverages an edge standalone architecture, incorporating a visual cognition AI model to proactively safeguard operators working in conjunction with robots. Moreover, this article delineates an end-to-end integration methodology encompassing visual detection, hardware integration, and the overarching objective of safeguarding operators in the event of a system compromise.

**Keywords:** Machine Vision, Interoperability, IoT, Cybersecurity, Industry 4.0

## 1 Introduction

The oversight of security considerations within the field of robotics frequently results in a state of heightened vulnerability for robots. This prevailing trend can be attributed to a multitude of factors. Primarily, the extant defensive security mechanisms designed for robots are still in the early stages of their development, lacking comprehensive coverage of the complete spectrum of potential threats. Consequently, these mechanisms prove to be inadequate in effectively safeguarding robots against malicious attacks. The intrinsic complexity inherent in robotic systems presents substantial challenges when attempting to implement robust security measures.

The intricate nature of these systems exacerbates both the technical and economic costs associated with ensuring their protection. As a consequence, manufacturers frequently adopt suboptimal security practices due to the considerable expenses and technical intricacies entailed in fortifying these advanced systems [Mayoral-Vilches, 2022]. Against this backdrop, this research paper introduces an autonomous system that leverages cost-effective components to mitigate the risks associated with robot compromise and potential harm resulting from malfunctions. The proposed system operates as a stand-alone entity shielded from external attacks while concurrently safeguarding the operator from potential dangers in scenarios involving robot hacking or malfunctions. Through the incorporation of visual cognition models, the system strives to ensure the operator's well-being and prevent injuries arising from compromised robot operations. This paper advocates for the implementation of an autonomous system fortified with visual cognition models, underscoring the importance of addressing security concerns and prioritizing the safety of operators engaged in robot-related scenarios. Figure 1 showcases a selection of robots on which our system has undergone testing and evaluation. The development of the proposed autonomous system aims to facilitate several key attributes,



Figure 1: Equipment used in the experiment

highlighting the importance of addressing security concerns and prioritizing the safety of operators engaged in robot-related scenarios. These attributes include:

- **Enhanced Security:** The system emphasizes the implementation of robust security measures, leveraging visual cognition models to detect potential threats and safeguard against likelihood of injuries or accidents.
- **Autonomous Functionality:** The system is designed to operate autonomously, reducing the reliance on manual intervention, and minimizing the potential for human error. This autonomous functionality enhances the efficiency and effectiveness of robot operations, while also mitigating the risks associated with operator involvement.
- **Real-time Threat Detection:** Through the integration of visual cognition models, the system is capable of real-time threat detection and response. By continuously monitoring the environment and identifying potential risks, the system can promptly react on anomalous activities, ensuring timely and appropriate actions are taken.
- **Operator Safety:** The system places a strong emphasis on operator safety, aiming to protect individuals from potential harm arising from robot compromise or malfunctions. By implementing proactive security measures and prompt response mechanisms, the system acts as a safety net, reducing the likelihood of injuries or accidents during robot operations.

## 2 Literature review

Cyber Physical systems in the Engineering perspective can be explained as the monitoring and dynamic control of physical components of any environment using sensor data and actuators that are part of a distributed computing system [Nunes, 2015]. These Cyber Physical Systems involves controlled interactions among robotics, Internet of things, Wireless sensor networks, Edge Processing and Cloud computing technologies. Humans have been an essential part of these cyber physical systems especially in the central control, but for any cyber physical system, human is an external and unpredictable element in the environment. This generated the idea of human-in-the-loop concepts which consider the human input through sensory data available and most importantly safety and security of the human involved in the control loop. Safety of the human in human-in-the-loop environment is the focus of work in this paper.

Current research in the human-in-the-loop front is about the interaction between machines and humans for data collection for interactive machine learning, called human-in-the-loop machine learning [Mosqueira-Rey, 2023]. Hence, ensuring the safety of the human operator at all stages on the factory floor has become ever more important and urgent. With advanced sensors and camera systems, video surveillance has been deployed in various environments, public infrastructures, commercial buildings, manufacturing settings and factory floors. Real-time automated analysis and inference from the image or video sequences generated by these camera systems are enabled through Deep Learning. Different applications of real-time person detection using deep learning has been developed over the recent years, access/egress, people counting, person identification and tracking, and most importantly in threat detection and emergency response[Ahmed I. and Jeon, 2021]. Although it's been widely used, person detection is a challenging task in machine vision, with problem of occlusion when person and object overlap each other. Adding further to the complexity is the diverse human postures, gestures and actions [Bouafia, Y., Guezouli, L. and Lakhlef, 2022]. MobileNet-V1 is the model that is used as a backbone in the detection system, which was proposed by Google researchers in 2017 using Depthwise Separate Convolution making it very efficient and less energy consuming [Bouafia, Y., Guezouli, L. and Lakhlef, 2022]. In 2019, MobileNet-V2 was announced with Bottleneck Residual Block instead of Depthwise Separable Convolution blocks.

The model is trained using INRIA person dataset and transfer learning has been used to combat the data-hungry nature of deep learning model when training from scratch. Transfer learning are of two types: Transfer learning via feature extraction and Transfer Learning via fine-tuning. In the former case, the pre-trained network is used to

extract features which could then be further used for classification or detailed data parameterization. In the latter case, the pre-trained network can be used fully by adding further layers and retraining the newly connected layers so that the currently stored weights could also be used to achieve good results.

Transfer learning is the best solution when the available dataset is small, compared to 14 million images in more than 21k categories in the ImageNet[J. Deng, 2009] dataset.

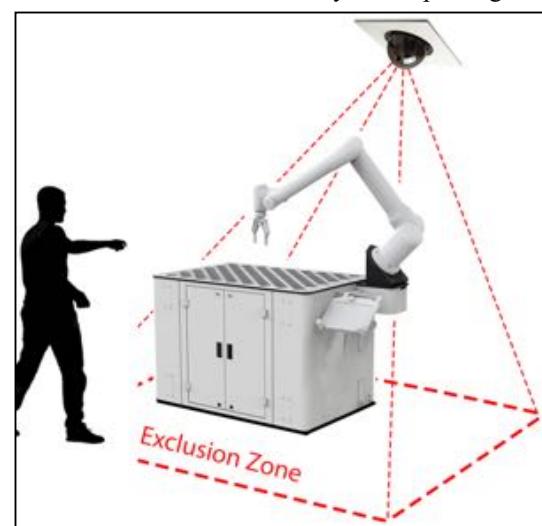
### 3 Proposed Solution

The methodology employed in this study adopts a qualitative research design, with a specific focus on qualitative content analysis. The primary objective of this research is to address real-world challenges related to operator safety in robot-related scenarios, particularly through the utilization of low-cost devices that are impervious to hacking. By emphasizing practicality and application, the study aims to identify and overcome the identified challenges in the field. The approach employed in this study involves practical experimentation and iterative development, allowing for the generation of tangible outcomes with practical applications that can be implemented in real-world settings. Through these iterative processes, the study aims to refine and enhance the proposed autonomous system, ensuring its effectiveness and reliability when addressing security concerns and prioritizing operator safety. By focusing on these key attributes, the system aims to mitigate potential risks and optimize the performance of robots in various operational contexts. By adopting a qualitative research methodology, this study seeks to provide a comprehensive understanding of the challenges and requirements associated with the development of the proposed autonomous system. Practical experimentation and iterative development serve as means to generate tangible outcomes that can be practically applied, thereby making notable contributions to the field of robotics and effectively addressing the identified challenges.

To effectively tackle the challenges associated with safety in robotic and industrial machinery operations, the proposed solution integrates AI-trained models and camera systems to establish a robust and real-time virtual safety fence. By implementing the proposed approach, the system establishes a virtual boundary encompassing the designated robot work area. In the event of a breach where the operator crosses this virtual boundary, the system triggers the fail-safe switch promptly and automatically. This instantaneous response ensures the immediate termination of potentially hazardous operations, effectively safeguarding the operators involved. Figure 2 visually illustrates the pivotal role played by the designed system in ensuring the safety of the surrounding environment and operators during robotic operations.

Because the system operates independently of any network connectivity, making it impervious to cyberattacks. The fail-safe switch, being a physical switch, cannot be overridden by any software intervention. This characteristic enhances the system's resilience and security, significantly reducing the risk of cyberattacks that could potentially harm the operators. The combination of an independent operation and a physical fail-safe switch enhances the system's suitability in preventing safety risks associated with cyber threats.

In summary, the proposed system creates a virtual boundary to delineate the robot work area, activating the fail-safe switch upon breaching this boundary. Figure 2 visually exemplifies the system's role in ensuring operator safety. Moreover, the system's independent operation and utilization of a physical fail-safe switch contribute to its robustness against cyberattacks, ultimately mitigating potential harm to the operators.



**Figure 2** Virtual boundary

Figure 3 provides a visual representation of the sequential steps involved in the development and construction of the system proposed in this paper. The depicted process encompasses several key stages, including the design of the hardware solution, the implementation and testing of the hardware components, the development of the software, the integration of the software and hardware components, the design and training of deep learning networks, and the final stage of integrating and testing the software and hardware components together.

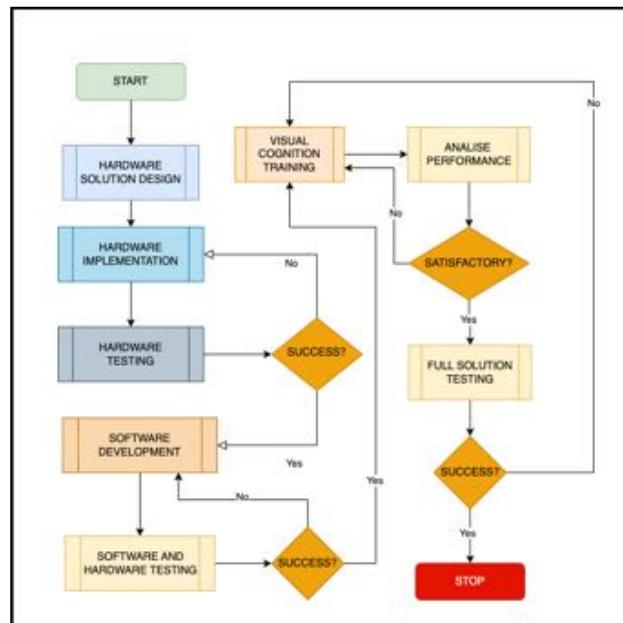
The first stage, labelled as the hardware solution design, involves the conceptualization and design of the hardware components that constitute the system. This stage encompasses determining the necessary hardware elements and their specifications to meet the objectives of the proposed system. The hardware solution undergoes implementation and testing to ensure its functionality and compatibility with the intended objectives. Rigorous testing and evaluation are conducted to verify the performance and reliability of the hardware components. Parallel to the hardware development, the software components of the system are created. This entails the development of software modules and algorithms that enable the desired functionality of the system. The software development process includes coding, debugging, and optimization to ensure efficient and effective operation. Once the software and hardware components are individually developed, the integration process takes place. This

phase involves combining the software and hardware elements to create a unified system. The integration process ensures seamless communication and coordination between the software and hardware components. The design and training of deep learning networks constitute a critical stage in the development process. Deep learning models are designed and trained to enable the system to perform complex tasks such as visual recognition or decision-making. This stage involves data collection, pre-processing, model design, training, and evaluation to achieve optimal performance.

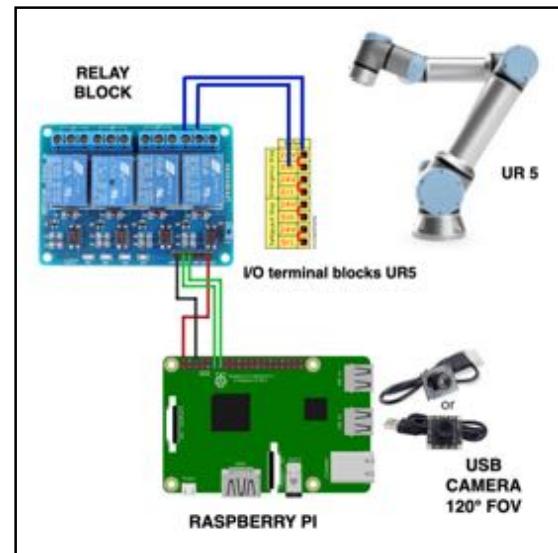
Finally, the integrated software and hardware components undergo comprehensive testing to validate the functionality, performance, and reliability of the system as a whole. This testing phase ensures that the system operates as intended and meets the specified requirements.

### 3.1 Hardware design

The hardware employed in this prototype encompasses the utilization of a Raspberry Pi 4 Model B/4G, renowned for its advanced capabilities. This device is equipped with an ARM Cortex-A72 CPU, which provides ample computational power necessary to sustain the execution of a dedicated mode within the system. This mode is specifically designed to facilitate deep learning detection, enabling the system to discern and identify various visual patterns or cues. Moreover, this powerful CPU facilitates the execution of subsequent actions required to activate safety procedures, ensuring



**Figure 3 Solution development**



**Figure 4 Hardware components**

the well-being and protection of operators in robot-related scenarios.

The Raspberry Pi 4 Model B/4G is purposefully employed in an autonomous manner, operating independently and without reliance on any external network connections. This autonomous functionality serves a crucial role in the system, allowing it to operate seamlessly and self-contained, irrespective of any external network dependencies. By leveraging the processing capabilities of the Raspberry Pi, the system can effectively execute visual cognition models, which are fundamental in detecting the presence of operators and evaluating potential risks associated with their interactions with the robot. The system is able to promptly trigger the robot's kill switch mechanism, ensuring a swift and appropriate response to mitigate any potential dangers or hazardous situations. The hardware connections and configurations are visually represented in Figure 3. This figure provides a comprehensive illustration of the interconnections and arrangements of the hardware components within the system.

### 3.2 Software design

The software utilized in this prototype, which runs on the Raspberry Pi, was developed using Python, a prominent object-oriented high-level programming language. Python is renowned for its emphasis on code readability and simplicity, which has contributed to its extensive usage across diverse domains such as web development, data analysis, scientific computing, and artificial intelligence, among others. Python's support for object-oriented programming principles empowers developers to create and define objects that encapsulate both data and behaviour. This approach facilitates the construction of modular and reusable code, enhancing the overall organization and structure of software applications. By leveraging Python's object-oriented capabilities, developers can design software components that accurately model real-world entities and interactions, resulting in more efficient and maintainable codebases. The utilization of Python in this prototype ensures that the software implementation is coherent, understandable, and facilitates the integration of various functionalities required for the successful operation of the system. By harnessing Python's versatility and object-oriented paradigm, the software developed for the Raspberry Pi effectively accomplishes the intended goals and seamlessly interacts with the hardware components[Penica M., 2021]. To facilitate control over the relay responsible for triggering the robot's kill switch, the RPi.GPIO library was incorporated into the software. This library provides convenient and straightforward control over the relay, enabling the system to effectively manage the relay's activation. By utilizing the RPi.GPIO library, the software gains the ability to manipulate the relay with ease, ensuring precise control and timely response when necessary. The integration of the RPi.GPIO library into the software streamlines the process of controlling the relay, enhancing the overall functionality and reliability of the system. With this library's assistance, the system achieves efficient and precise triggering of the relay, thus enabling swift and accurate activation of the robot's kill switch mechanism as required for operator safety and protection.

### 3.3 Visual cognition

The person detection is achieved using TensorFlow object detection model with an artificial Neural Network as the backbone and Single Shot Detection as the detection method. The deep learning model is trained using publicly available benchmark dataset for person detection, the INRIA persons dataset [Dalal and Triggs 2005]. Transfer Learning is used to achieve higher accuracy in the detection, the model pretrained on ImageNet has been downloaded from the TensorFlow repository and fine-tuned for detection process. GPU GeForce RTX 2080 Ti is used to complete the model training and the TensorFlow-Lite graph is exported onto the Raspberry Pi.

Edge processing is vital in this application since the detection system will be component of(mounted on) a robotic arm. Added to that, the detection is real-time from the Raspberry Pi. SSD-MobileNetV2 is the model used in this prototype, achieving consistently high confidence score of detections, > 95% within detection time of <1ms, in all the scenarios tested. The SSD-MobileNetV2 framework is specifically designed for mobile and Edge applications, with Depthwise separable convolution modules making detections faster thereby reducing reaction time after the virtual fence has been breached.

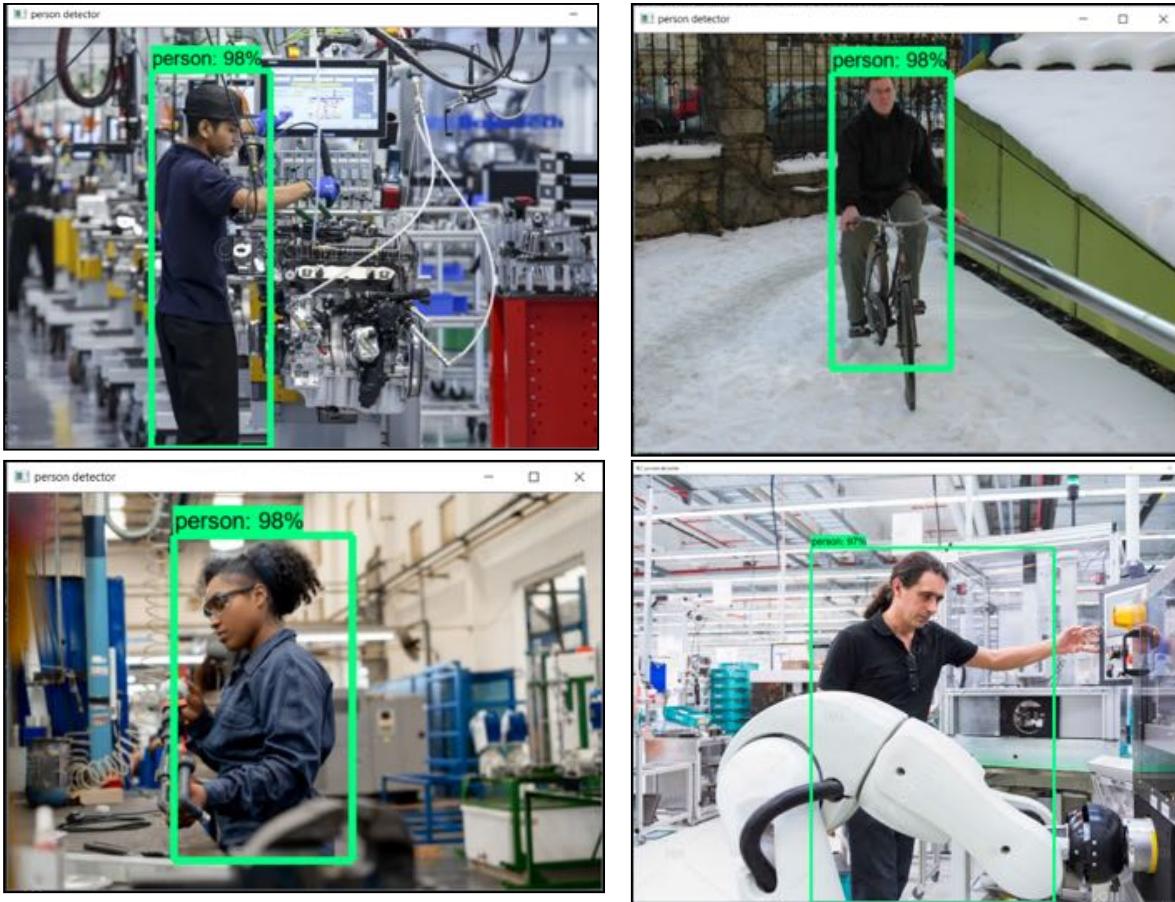


Figure 5: The above images shows high accuracy in person detection in factory settings and one of them in a random environment.

## 4 Results

In order to evaluate the performance of the system in detecting individuals and ensuring operator safety, a series of tests were conducted using the developed prototype and following the methodologies outlined in this paper. The tests involved ten volunteers with diverse appearances, who participated in ten different scenarios while wearing various accessories such as glasses. The objective of these tests was twofold: first, to determine the distance at which the system would effectively detect the presence of a person, and second, to measure the response time required to ensure the operator would be in a safe environment in the event of a malfunction or intentional hack aimed at causing harm to the operator. During the testing scenario, volunteers approached the robot while wearing different accessories and performed various actions. The system's performance was evaluated based on its ability to accurately detect the presence of the volunteers within a predefined distance and the timeliness of its response in ensuring operator safety. The conducted tests yielded valuable insights regarding the system's performance in real-world scenarios. Based on the collected data, optimal distances for detection and virtual fence placement were determined. For instance, it was found that the optimal distance for person detection was 3 meters away. Additionally, the optimal distance for placing the virtual fence varied depending on the type of robot used in the experiment.

Table 1 below presents the minimum distances and the reaction time required for the robots to come to a stop based on the specific robot types used, ensuring the operator is in a safe environment:

Robot Type	Minimum Safe Distance	Reaction time after the virtual fence has been breached(ms)
UR 3	43 cm	0.321ms
UR5	83.5 cm	0.191 ms
UR 10	122 cm	0.201 ms

**Table 1: Minimum safe distance**

Table 2 below presents the evaluation results for the person detection module used in this prototype. With finetuning only with INRIA dataset, the model has been able to achieve an IOU value > 80%. Intersection Over Union (IOU) value is used to determine the prediction accuracy of a model, it is the ratio of area of intersection of the predicted bounding box vs area of intersection of the ground truth bounding box. Precision value of a model represents the ratio of True Positives to All Positive Detections. Higher value of Precision value shows that the detections are very accurate and near negligible false detections. Recall is another metric to measure the number of false detections, it's the ratio of True positives against all ground truth instances, means, the number of objects detected from all the instances present in the scene. Higher value of recall indicates the system is highly sensitive to the trained object and detects all instances in range.

Detection Time(ms)	IOU value	Precision	Recall
0.813	0.8134	1	1

**Table 2: Evaluation Metrics for Operator Detection**

These findings provide valuable guidance for implementing the system in practical applications, aiding in the effective detection of individuals and the establishment of appropriate safety measures based on the specific robot being used. In summary, the tests conducted using the developed prototype and following the methodologies outlined in this paper provided insights into the system's performance in detecting individuals and ensuring operator safety. Optimal distances for person detection and virtual fence placement were determined, taking into account the type of robot utilized in the experiments. These findings contribute to the successful implementation of the system in real-world scenarios, promoting enhanced safety for operators engaged in robotic operations.

## 5 Conclusions

This paper provides compelling evidence for the practicality of constructing an affordable device capable of running deep learning models at the edge. By proficiently detecting operators, the device effectively safeguards against malfunctions and external malevolent attacks, ensuring operator safety. The detection capability of deep learning model could be further improved by adding more training instances from manufacturing settings which presents the model with more variable instances of human operators working with robots. The research findings underscore the efficacy of employing cost-effective devices to enforce operator safety. The proposed approach emphasizes the device's standalone nature, offering an independent solution that is impervious to hacking or tampering. The study demonstrates that the utilization of a low-power ARM device yields optimal error rates and cognition model performance, enabling fast decision-making capabilities. Furthermore, the research validates the feasibility of implementing lightweight models on embedded devices. In conclusion, this paper conclusively establishes the feasibility of developing a cost-effective device capable of executing deep learning models at the edge. Through precise operator detection, the device effectively guarantees operator safety, even in the presence

of malfunctions or external threats. The study affirms the independent operation of the proposed approach and showcases the exceptional performance of the cognition model when utilizing a low-power ARM device. Additionally, the successful implementation of lightweight models on embedded devices further strengthens the paper's contributions to the field.

## Acknowledgements

The authors would like to thank the staff of the Electronic and Computer Engineering Department at the University of Limerick for their assistance. The authors would also like to acknowledge resources made available from Science Foundation Ireland through the grant award (16/RC/3918) to Confirm Centre for Smart Manufacturing.

## References

- [Mayoral-Vilches, 2022] Mayoral-Vilches, V., 2022. Robot hacking manual (rhm). *arXiv preprint arXiv:2203.04765*.
- [Nunes, 2015] Nunes, D.S., Zhang, P. and Silva, J.S., 2015. A survey on human-in-the-loop applications towards an internet of all. *IEEE Communications Surveys & Tutorials*, 17(2), pp.944-965.
- [Mosqueira-Rey, 2023] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. and Fernández-Leal, Á., 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4), pp.3005-3054.
- [Ahmed I. and Jeon, 2021] Ahmed, I. and Jeon, G., 2021. A real-time person tracking system based on SiamMask network for intelligent video surveillance. *Journal of Real-Time Image Processing*, 18, pp.1803-1814.
- [Bouafia, Y., Guezouli, L. and Lakhlef, 2022] Bouafia, Y., Guezouli, L. and Lakhlef, H., 2022. Human Detection in Surveillance Videos Based on Fine-Tuned MobileNetV2 for Effective Human Classification. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 46(4), pp.971-988.
- [J. Deng, 2009] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.
- [Penica M., 2021] Penica, M., Mohandas, R., Bhattacharya, M., Vancamp, K., Hayes, M. and O'Connell, E., 2021, June. A Covid-19 viral transmission prevention system for embedded devices utilising deep learning. In *2021 32nd Irish Signals and Systems Conference (ISSC)* (pp. 1-8). IEEE.
- [Dalal and Triggs, 2005] Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. *Comput Vis Pattern Recogn CVPR 2005* IEEE 1:886–893

# YOLOatr : Deep Learning Based Automatic Target Detection and Localization in Thermal Infrared Imagery

Aon Safdar<sup>1</sup>, Usman Akram<sup>1</sup>, Waseem Anwar<sup>2</sup>, Basit Malik<sup>1</sup>, Mian Ibad Ali<sup>3</sup>

<sup>1</sup>*National University of Sciences and Technology, Islamabad, Pakistan*

<sup>2</sup>*Mälardalen University, Västerås, Sweden*

<sup>3</sup>*University of Galway, Ireland*

## Abstract

Automatic Target Detection (ATD) and Recognition (ATR) from Thermal Infrared (TI) imagery in the defense and surveillance domain is a challenging computer vision (CV) task in comparison to the commercial autonomous vehicle perception domain. Limited datasets, peculiar domain-specific and TI modality-specific challenges i.e., limited hardware, scale invariance issues due to greater distances, deliberate occlusion by tactical vehicles, lower sensor resolution and resultant lack of structural information in targets, effects of weather, temperature, and time of day variations and varying target to clutter ratios all result in increased intra-class variability and higher inter-class similarity making accurate real-time ATR a challenging CV task. Resultantly, contemporary state-of-the-art (SOTA) deep learning architecture under-perform in the ATR domain. We propose a modified anchor-based single-stage detector called YOLOatr, based on a modified YOLOv5s, with optimum modifications to detection heads, feature-fusion in the neck, and a custom augmentation profile. We evaluate the performance of our proposed model on a comprehensive DSIAC MWIR dataset for real-time ATR over both correlated and decorrelated testing protocols. The results demonstrate that our proposed model achieves state-of-the-art ATR performance of up to 99.6%.

**Keywords:** Automatic Target Detection, Automatic Target Recognition, Yolov5, Deep Learning, Computer Vision.

## 1 Introduction

Object detection, localization, classification, and tracking are the key components of modern computer-vision-based applications. Automatic Target Detection (ATD) encompasses the detection and localization of the objects in an image while Automatic Target Detection (ATR) involves further classification of the detected objects into relevant classes. Both terms (ATD and ATR) are usually associated with object detection/recognition in surveillance and defense domains where robust target recognition and tracking are considered critical. Conventional imaging modalities in these domains include Infrared (IR) and synthetic aperture radar (SAR) with Thermal IR (TIR) being the predominant modality for ground-based tactical platforms with proven benefits [Zhao et al., 2022]. TIR includes shortwave infrared (SWIR 1-3  $\mu\text{m}$ ) and mediumwave infrared (MWIR, 3-5  $\mu\text{m}$ ) bands of the infrared wavelength where most captured radiations are emitted and not reflected from the targets [Berg et al., 2015]. TIR thus provides the ability to see targets both in the presence or absence of light. It is a frequently used imaging modality for ground based tactical/surveillance vehicles.

Despite the numerous advantages offered by TIR images, object detection/recognition using IR modality in tactical/surveillance applications suffers from peculiar domain challenges that affect reliability and robustness of an ATR system [Berg et al., 2015]. Ground-based tactical platforms generally operate in extreme weather and temperature conditions. This affects the image sensor output thus degrading heat signature differential between the target and the background [Liang et al., 2022]. The occlusion in defense scenario is deliberate as the target tries to camouflage or conceal itself and attempts to be occluded from view. Scale variations are high in case of ATR as the sensor-to-target distances and much greater. Low resolutions at far distances make discerning the defining target features difficult resulting in high inter-class similarities. Moreover, the environment is highly cluttered that makes detection/recognition a daunting task [Arif & Mahalanobis, 2020a]. Contextually, ATR has to be performed on a live video and not on still images. In videos, motion blur and occlusion are more vital than in still images. Moreover, fast inference speeds for real-time performance becomes critical. ATD/R has to be performed in varying practical scenarios with respect to movement of source and target. Consequently, there is high viewpoint, occlusion and scale

variability in practical scenarios. Furthermore, IR data captured by the sensor varies significantly with metrological conditions, signal-to-clutter noise ratio (SCNR), time of day, sensor calibration and target viewpoint variations that increase the intra-class variabilities [Batchuluun et al., 2020]. For defence and surveillance applications, a significant problem is non-availability of comprehensive IR datasets that can be used for training and evaluation of ATR algorithms. The lack of suitable large-scale labeled IR dataset for military / surveillance objects is attributable to confidentiality. Owing to these reasons highlighted above, latest state of the art (SOTA) algorithms not only remain under-evaluated for TIR-based ATR, they also under-perform due to peculiarities specific to the IR modality and to the tactical domain. In this paper, we investigate the application of a state-of-the-art Deep Convolutional Neural Network (DCNN) based detector to address target detection, recognition and localization in thermal infrared video imagery for defence domain. Specifically, we explore different learning approaches, architectural changes and data augmentation profiles with a single-stage, single-frame object detector called You Only Look Once – Version 5 (YOLOv5) and propose the optimum learning approach, architectural changes and data augmentations to improve accuracy and robustness. We evaluate our approach and proposed model on the Defense Systems Information Analysis Center (DSIAC) benchmark ATR dataset.

This paper is organized as: Literature review is provided in Section 2. The YOLOatr development methodology is explained in Section 3. Result Evaluation forms part of Section 4. The discussion and limitations are given in Section 5. Finally, the article is concluded in Section 6.

## 2 Literature Review

The learning-based target detection/recognition solutions can be categorized into traditional machine learning-based methodologies and deep learning-based methodologies. The prominent and pertinent dataset used for vehicle detection algorithm development in the tactical domain is the DSIAC ATR dataset. Deep neural networks applied to this dataset can be divided into two categories based on the testing range: correlated and decorrelated datasets. Correlated datasets involve testing on the same range as training, while decorrelated datasets involve testing on higher ranges.

[Mahalanobis & McIntosh, 2019] compare the performance of Faster R-CNN network with a quadratic correlation filter (QCF) based algorithm using the DSIAC MWIR dataset. Millikan et al., 2018 proposes QCF filters in the first layer of a CNN for target classification but does not discuss ATR at longer ranges with high clutter. A different approach [d'Acremont et al., 2019] addresses the issue of inadequate datasets for training CNNs by using a compact and fully convolutional neural network with global average pooling, along with a simulated/synthetic dataset. This approach shows improved recognition performance and robustness to scale and viewpoint variations. Another study [Arif & Mahalanobis, 2020] employs a CNN-based auto-encoder to generate unseen views for the DSIAC dataset, achieving a test accuracy of 68% for a single vehicle class at a 1000m range. Similarly, an autoencoder and Siamese network are used [Arif & Mahalanobis, 2021] to generate realistic images from the DSIAC dataset. Another experiment [Chen et al., 2021] compares three variants of a single-stage detector with varying backbone structure for feature extraction, achieving high mean Average Precision (mAP) for ATR using four target classes.

For decorrelated datasets, various studies [McIntosh et al., 2020a/b; Jiban et al., 2021; Cuellar & Mahalanobis, 2021] argue that training and testing over the same ranges introduce potential biases, as clutter diversity, scale, and resolution issues affect detector performance. These representative studies split the DSIAC dataset for training at lower and testing at higher ranges. One study [McIntosh et al. 2020] proposes TCRNet, a custom network utilizing analytically derived eigen filters to maximize the Target-to-Clutter (TCR) metric. Another study [Jiban et al., 2021] improves upon TCRNet by processing target and clutter information in parallel channels before combining for TCR metric optimization. However, these studies focus on target detection (ATD) rather than target classification (ATR).

A few studies [Millikan et al., 2018 ; Liang et al., 2022] use multi-frame techniques based on background consistency and target sparsity hypotheses for moving target detection in the DSIAC datasets. However, these hypotheses are violated in real-time scenarios. In such cases, single-frame detection techniques are considered more practical

**Research Gap:** Based on literature review, the identified research gaps include:1) SOTA detectors largely remain under-evaluated and also under-perform in tactical setting due to peculiar IR and domain challenges. 2) There is a dearth of TIR based datasets for the ATR community. Moreover, annotating and preparing large video-based

datasets (such as DSIAC dataset) requires tremendous amount of time and human effort and is required to be automated. 3) A robust single-frame, single-stage, anchor-based detector is required for high-accuracy, high-speed inference using limited hardware onboard tactical vehicles with generalization ability over decorrelated ranges, at multiple-combined distances and in varied illumination/SCNR (Day/Night) conditions. To the best of the author's knowledge there are no studies to date that have improved SOTA single stage detector, specifically YOLOv5, for ATR task on the DSIAC MWIR dataset at both correlated and decorrelated ranges with combined distances and time (both day and night). Our goal is to improve YOLOv5 for a robust solution to reliably detect small IR targets in high clutter at multiple ranges simultaneously in all weather, illumination, range and, viewpoint variations.

### 3 YOLOatr Development Methodology

Our YOLOatr development methodology (Figure 1) is divided into two parts. Firstly, we preprocess the DSIAC dataset for our experiments. Secondly, we follow an experimental approach and perform an ablation study over various model variants of the base model (YOLOv5s) to find the optimum modifications in terms of training methodology, architecture and data augmentation. Finally, we use the optimum model (YOLOatr) and evaluate its performance and generalization ability. We finally perform a comparative analysis of results achieved by YOLOatr with other SOTA detectors used in various studies.

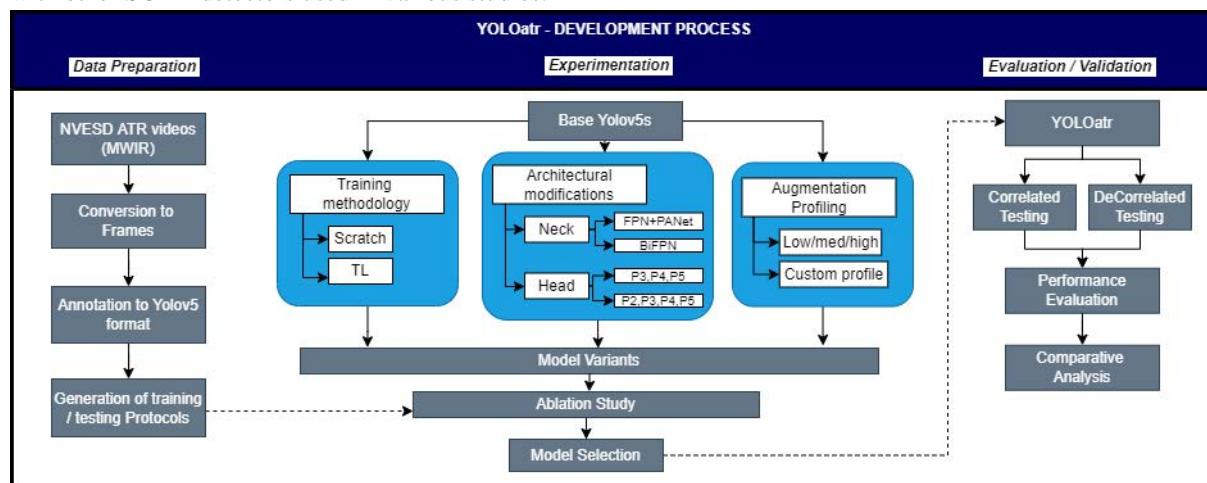


Figure 1 : Methodology for development of YOLOatr

#### 3.1 Dataset and Partitioning Protocols

**DSIAC MWIR Dataset:** We have utilized the US Army Night Vision and Electronic Sensors Directorate's (NVESD) ATR Algorithm Development Image Database which is the largest and most comprehensive publicly available TIR dataset for ground-based tactical platforms. The dataset was released by the U.S Department of Defence (DoD) for ATR algorithm development and contains visible and mid-wave infrared (MWIR) thermal

	MWIR	Visible
Source Camera	“cegr”	“ilco”
Size	207 GB	106 GB
Aspect Angles	72 - obtained by driving in circle of 100m @ 10mph	
Distances	1000m to 5000m @ increments of 500m for both day and night	
Files	186 Files in ARF format viewable using ImageJ software via a plugin, ground truth in ATG format	
Length	1 min videos @ 30fps (1800 frames per video)	
Targets	10 Vehicles (2 civilian and 8 tactical vehicles) 2 Human (slow- and fast-moving pedestrians)	
Samples (one target)	1800 frames x 18 videos = 32400 frames (Vehicle) 1800 frames x 12 videos = 21600 frames (Human)	
Total Dataset size	(32400 x 10) + (21600 x 2) = 367,200 approx.	

Table 1 : The DSIAC dataset

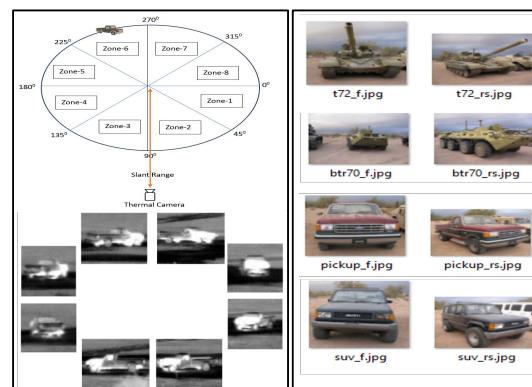


Figure 2 : Image Acquisition (Left) and Selected Targets (Right)

images of tactical vehicles over long ranges in the desert environment. It contains (Table 1) images of 13 different

tactical and civilian vehicles moved from 1000m to 5000m with increments of 500m. Data is collected for both day and night situations. 72 aspect angles for each vehicle type are obtained by driving in a circle of 100m at a speed of 10mph. Minute-long videos @ 30fps for each vehicle during both day and nighttime were converted into frames and annotated for ground truth. Four different vehicles (T72, BRDM2, Pickup, SUV) were used for our experiments (Figure 2). A total of 3600 images were used for each vehicle type with 70% data for training, 20% for validation, and 10% for testing.

**Training and Testing Protocols:** To ascertain the performance and generalization ability of our model, we use two testing protocols. The ‘correlated dataset partitioning’ protocol (T1) where training and testing images are from the same range and the ‘decorrelated dataset partitioning’ protocol (T2) where training is done at lower ranges and testing at higher ranges. The details of partitioning protocols are given in Table 2.

### 3.2 YOLOv5s Architectural Modifications

ATR for TIR requires high accuracy and real-time inference speed. This dictates the choice of selecting an appropriate Object Detector that strikes a balance between both speed and accuracy. YOLO (You Only Look Once) is a cutting-edge single-stage, single frame object detector that can achieve both goals (i.e., speed and accuracy). The YOLO family has many variants (e.g., YOLOv1, YOLOv2, YOLOv5, YOLOX, YOLOR etc.). For our base model we selected YOLOv5s, which is the small variant of the Yolov5 series [Jocher, 2022]. We experimented with architectural modifications to the original YOLOv5s neck and head. As our dataset contains great concentration of small objects with only few pixels on target and because YOLOv5 is known to struggle with small object detection, we introduced an extra small detection head (P2) in order improve small object detection performance. Moreover, we replaced the original PANet neck with BiFPN (Weighted Bi-directional Feature Pyramid) Network neck to promote better feature fusion and propagation. Our modified YOLOv5s architecture is shown in Figure 3.

### 3.3 Learning Methodology

We experimented with two learning methodologies two learning methodologies i.e, learning from scratch and transfer learning. We tested both approaches initially with baseline YOLOv5s. For learning from scratch, we trained the model using 100 epochs with randomly initialized weights. We used the training images as per established protocol for training from scratch. For transfer learning we used pre-trained weights (Imagenet) and trained with the backbone layers frozen for 30 epochs, followed by 10 epochs for fine-tuning in which all layers were unfrozen. The results with training from scratch were slightly better than from transfer learning. This may be attributed to the fact that the weights learned from RGB images do not comply with features present in the TIR objects. Using weights/knowledge from RGB domain for the TIR domain did not offer any advantage except for reduced training time/ epochs.

### 3.4 Augmentation Profiling

YOLOv5s uses a rich set of data augmentation techniques at training time to improve model accuracy by diversifying input images. Considering the nature of available images and ATR task, we created a custom augmentation profile (CAP) for model training as highlighted in Table.3.

Dataset	Training (Day + Night)			Testing (Day + Night)					
	Range in Km	Images		Correlated (T1)			Decorrelated (T2)		
		One Target	Four Targets	Range in Km	One Target	Four Targets	Range in Km	One Target	Four Targets
DS1	1.0, 1.5, 2.0, 2.5	10080	8064	1.0, 1.5, 2.0, 2.5	2880	2304	3.0	3600	2880
DS2	3.0, 3.5, 4.0, 4.5	10080	8064	3.0, 3.5, 4.0, 4.5	2880	2304	5.0	3600	2880

Table 2 : Dataset partitioning protocols

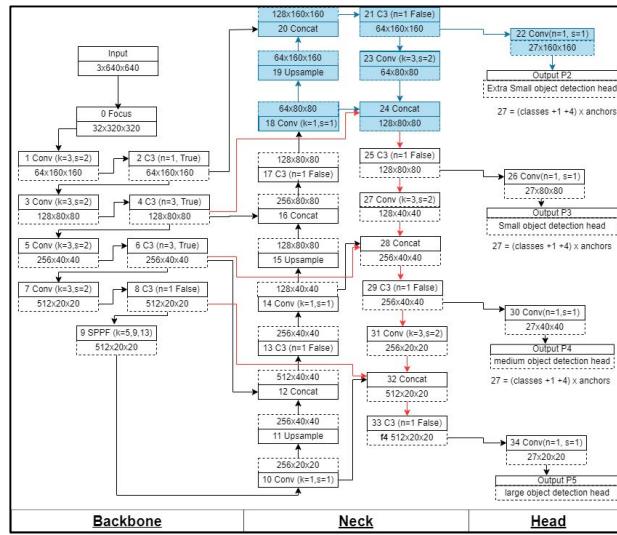
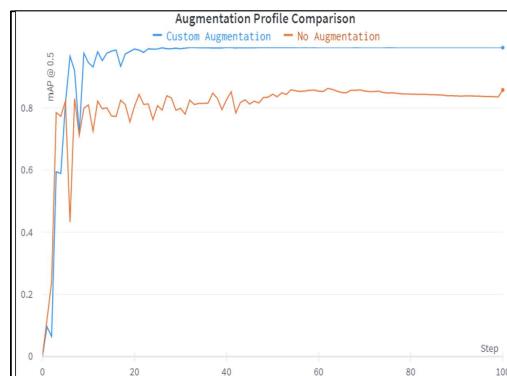


Figure 3 : Architectural modifications to YOLOv5

Augmentation type	Custom value
<i>fl_gamma</i> - Focal Loss Gamma	0.3
<i>hsv_h</i> - Hue	0.015
<i>hsv_s</i> - Saturation	0.7
<i>hsv_v</i> - Value	0.4
<i>degrees</i> - Rotation (+/- degree)	3
<i>translate</i> - Translation (+/- fraction)	0.1
<i>Scale</i> - Image Scale (+/- gain)	0.3
<i>shear</i> - Image Shear (+/- deg)	0.0
<i>perspective</i> (+/- fraction)	0.0005
<i>flipud</i> - flip up/down	0.1
<i>filpr</i> - flip left/right	0.5
<i>mosaic</i>	0.1
<i>mixup</i>	0.4
<i>copy_paste</i>	0.5

**Table 3 : Custom Augmentation Profile****Figure 4 : Model Training Accuracy Comparison of Default vs CAP**

Specifically, the brightness and contrast changes were kept as they are in line with varied illumination conditions caused by weather and time of day/night in practical scenarios. Similarly, the translation, rotation, scaling, flipping, and perspective changes were kept at medium level as ample variations in translation and aspect angles are available in the original dataset. The mosaic/ tiling data augmentations helps model with small object detection. It was kept low as the targets in original images at long ranges occupy only a few pixels and reducing their size further would be counterproductive. Similarly, as the structural information in low resolution original images at long ranges is already very low, the shear augmentation was turned off. Mix-up and copy-paste augmentations cross-stitch samples to increase the richness/ diversity of data and were kept high for that reason. YOLOv5s also provides albumentations integration for added augmentation of blurring etc. However, albumentations being additive to data augmentations were turned off so as not to overwhelm the model training process. Table 3 details the changes incorporated through our custom augmentation profile. Figure 4 shows how the CAP improves model training and validation accuracy in comparison with no/low data augmentation.

## 4 Evaluation and Validation

As the YOLOv5s model has not been previously used with the DSIAC dataset, we first performed experiments with the default YOLOv5s model to ascertain baseline performance of this cutting-edge detector. We then we incorporated our proposed architectural changes to neck and head, along with our custom data augmentation to train YOLOatr with our selected learning methodology and compared performance with the baseline model.

### 4.1 Experimental Setup and Evaluation Metrics

All experiments were performed on Google Colab pro platform using Nvidia Tesla P100-PCIE (16 GB) GPU and 32 GB High-speed RAM. The input image size was kept to 640x640. SGD optimizer was used with default learning rate and momentum instead of Adam because SGD generalizes better while Adam converges faster. The batch size was kept 32 and models were trained for 100 epochs. The commonly used evaluation metrics of precision (1), recall (2) and mean Average Precision (mAP) (3) at an Intersection over Union (IoU) threshold of 0.5 was used for ascertaining performance. The tactical/ATR domain equivalent of Precision is called 'Probability of target declaration (Pdc)', while Recall is analogous to 'Probability of detection (Pd)'. However, we use the commonly used terms of precision and recall in all our results.

$$Precision = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP}{All\ Detection} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{All\ Ground\ Truths} \quad (2)$$

$$mAP = \frac{1}{|Classes|} \sum_{i=1}^{|Classes|} AP^i \quad (3)$$

### 4.3 Experimental Results

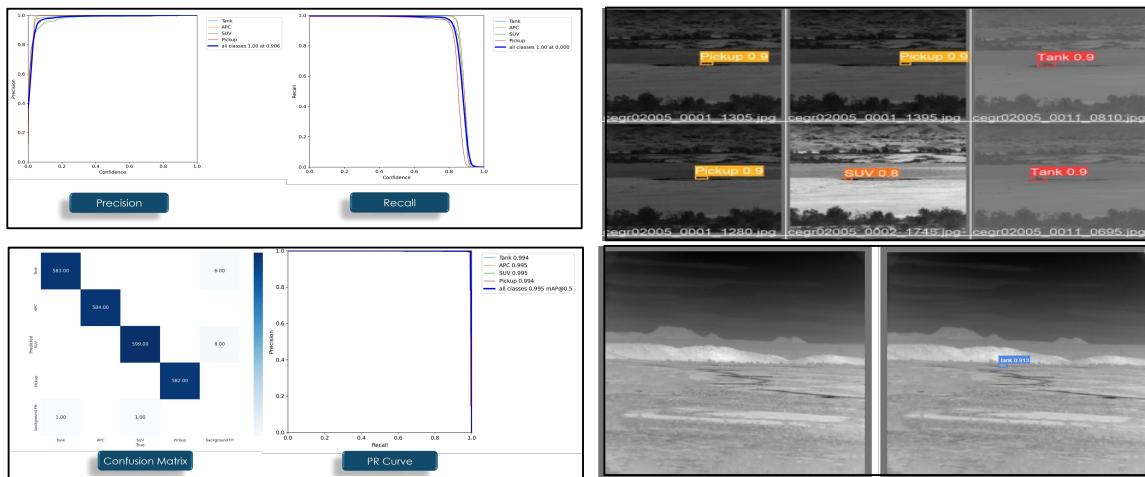
We ascertained the performance with both the original YOLOv5s and proposed YOLOatr on DS1 subset of the DSIAC ATR dataset. Both models were trained from scratch for 100 epochs with SGD optimizer. The results are shown in Table 4 (Left) showing that YOLOv5s performed commendably for the correlated test range (T1) for all target types achieving a mAP score of 99.4% while YOLOatr achieved a mAP of 99.6% with a 0.02% performance gain over YOLOv5s. The results establish the dominance of single-frame, single-stage, anchor-based DCNN in achieving high accuracy with a lean structure.

Model	Training Range	Time	Testing Results (mAP)		Parameters	Inference time (ms)	GFLOPs
			T1 (1.0-2.5 km)	T2 (3.0 km)			
YOLOv5s	1.0-2.5	Day and Night	0.994	0.263	7020913	4.0	16.2
YOLOatr			0.996	0.377	7086449	4.5	16.4
Performance Gain			+0.02%	+11.4%			

**Table 4 : Performance of Yolov5 and YOLOatr (Left), YOLOatr Target-wise performance comparison (Right)**

However, the model accuracy of YOLOv5s declined significantly when tested over the decorrelated range (T2) achieving a mAP of 27.1%. The results highlight that although the YOLOv5s performs well over correlated ranges, it struggles to generalize over the decorrelated ranges. The probable reasons maybe that as range increases the scale becomes smaller and the model struggles to find small targets. Moreover, the clutter changes with decorrelated range and resolution, structure of the target also degrades at higher range. Here, YOLOatr outperformed YOLOv5s with a performance gain of 11.4% achieving mAP of 37.7%. Assessing the target-wise performance of YOLOatr in Table 4 (Right), the most affected target was Pickup truck with a mAP of 16.3% while Tank T-72 achieved highest accuracy at T2 with 62.2% score. This may be attributed to size and thermal signature which is more for the tactical vehicle as compared to the commercial vehicles.

Overall, YOLOatr shows a performance gain of 0.02% for ATR at decorrelated ranges and 11.4% for decorrelated ranges against YOLOv5s. The results achieved by YOLOatr are SOTA to date with near-perfect accuracy of 99.6% for small objects over long ranges in highly cluttered IR images. Performance of YOLOatr is highlighted in the Precision and Recall curves, confusion matrix and mAP (PR Curve) for unseen data in Figure 5 (Left). The results highlight that YOLOatr achieves SOTA performance for the ATR task.



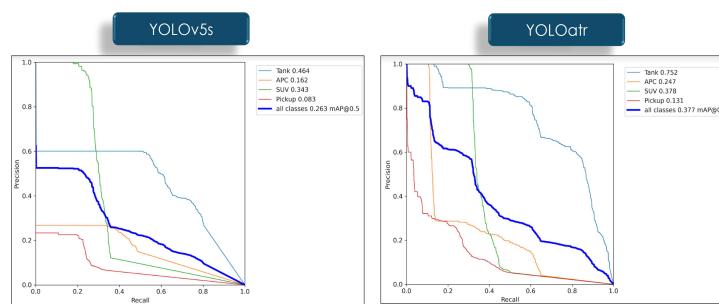
**Figure 5 : YOLOatr performance on Test data (Left), Detection of different targets (Top Right), Target detection at 5000m (Bottom Right), Detection at 5000m range (Bottom Right)**

Visual inference results for different targets are shown in Figure 5 (Right). YOLOatr detects target in highly cluttered background with a high confidence which is a prohibitive task for human vision. Considering the inherent challenges of IR modality and tactical domain, the performance of YOLOatr is phenomenal.

## 5 Discussion and Limitations

In this paper we have attempted to solve the problem of ATR in TIR for the tactical and surveillance domain. Specifically, we have developed, trained and tested YOLOatr, a lean, accurate, fast, and robust model for recognizing targets in challenging TIR images in real time. We have demonstrated that our model is able to detect robust classification features achieving up to 99.6% accuracy at distances up to 5000m. This is a prohibitive task for human vision but YOLOatr can both reliably detect and recognize different targets. The reliability of our ATR algorithm to detect and recognize each individual vehicle from as far as 5,000 meters is SOTA. Moreover, as compared to YOLOv5, YOLOatr improves the generalization ability over decorrelated ranges.

An analysis of YOLOv5s performance over the difficult problem of decorrelated testing and improvements shown by YOLOatr is shown in Figure 6. It can be seen that YOLOatr improves results over decorrelated ranges and generalizes better than original YOLOv5s. A few other studies attempt to solve the problem with different techniques. However, mostly, the studies target ATD which only includes target detection and not recognition. Moreover, some studies only use nighttime data for result evaluation while most studies only use testing at correlated ranges only. A comparative analysis of YOLOatr with leading studies is drawn in Table 6. It can be seen that significant improvement was achieved in terms of accuracy and speed.



**Figure 6 : Performance improvement of YOLOatr over YOLOv5s**

Ser	Publication	ATR/D	Classes	D/N	Range (km)		Performance		Performance YOLOatr	
					Tr	Te	mAP	Speed (fps)	mAP	Speed
1.	Chen et al., 2021	ATR	4	N	4.0	98.0%	92.6-99.3*	92.6-99.3*	99.4%	110@
				N	5.0	99.5%			99.5%	
				D	5.0	85.5%			88.3%	
				N	4.0-5.0	99%			99.4%	
2.	Ceullar et al., 2021	ATD	1	D+N	4.0-4.5	95.87%	110@	110@	99.6%	110@
					4.0-4.5	5.0			73.3%	
					4.0-4.5	5.0			73.3%	
					4.0-4.5	5.0			99.6%	
					4.0-4.5	5.0			73.3%	
3.	Maliha et al., 2021	ATR	10	D+N	1.0-2.0		88%	@Using NVIDIA PC100	98%	@Using NVIDIA PC100
4.	Acremont et al., 2019	ATR	10		1.0		70%		99%	
5.	McIntosh et al., 2020	ATR	10		1.0-2.0	2.5-3.5	83%		85.5%	

\* Using NVIDIA Quadro RTX5000 @480x480 image size  
Other models using Geforce GTX Titan X @ 480x480 image size:

- Yolov2 - 59 fps
- Faster-RCNN with VGG16 - 7 fps
- Faster-RCNN with Resnet - 5 fps

@Using NVIDIA PC100  
Underlined: denotes Accuracy

**Table 5 : Performance comparison of YOLOatr with SOTA models on DSIC dataset**

YOLOatr proves to be a robust single-stage detector that performs exceedingly well for both ATR and ATD cases to reliably detect targets in TIR images. It is a lean architecture with very fast inference speed of 110 fps making it a viable detector with high accuracy and less computational requirements. YOLOatr has few limitations. Firstly, optimum hyperparameter selection has not been explored via genetic algorithm available with yolov5. Secondly, comparative analysis of model accuracy with False Alarm Rate (FAR) has not been performed as done by few other studies. Finally, modifications to the head structure are not explored that may improve feature extraction. Performance over recently proposed variants of YOLO (Yolov7 Yolov8, yolo NAS etc.) is under experimentation and will form part of our research in future.

## 6 Conclusion

In this paper, we aim to achieve robust Automatic Target Detection / Recognition of small IR objects in cluttered background. We strive to do so over long distances, in cluttered background, and in varying illumination conditions. We focus on improving detection and classification accuracy with fast inference speed using a leaner architecture

compatible with limited computational power available on-board ground-based tactical and surveillance vehicles. We propose YOLOatr with optimum training approach, augmentation profile and structural modifications to Yolov5s and tailored for the ATR case in challenging tactical / surveillance TIR domain. We demonstrate improvement in detection accuracy while preserving implementation potential. Additionally, we demonstrate better generalization ability for improved robustness over different range, viewpoint, scale, clutter and illumination variations.

## References

- [Zhao et al., 2022] Zhao, M., Li, W., Li, L., Hu, J., Ma, P., & Tao, R. (2022). Single-frame infrared small-target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine*.
- [Berg et al., 2015] Berg, A., Ahlberg, J., & Felsberg, M. (2015, August). A thermal object tracking benchmark. In 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6).
- [Liang et al., 2022] Liang, X., Liu, L., Luo, M., Yan, Z., & Xin, Y. (2022). Robust infrared small target detection using Hough line suppression and rank-hierarchy in complex backgrounds. *Infrared Physics & Technology*, 120, 103893.
- [Arif & Mahalanobis, 2020] Arif, M., & Mahalanobis, A. (2020). View prediction using manifold learning in non-linear feature subspace. In MIPPR 2019: Pattern Recognition and Computer Vision (Vol. 11430, pp. 316–323). SPIE.
- [Batchuluun et al., 2020] Batchuluun, G., Kang, J. K., Nguyen, D. T., Pham, T. D., Arsalan, M., & Park, K. R. (2020). Deep learning-based thermal image reconstruction and object detection. *IEEE Access*, 9, 5951-5971.
- [Mahalanobis & McIntosh, 2019] Mahalanobis, A., & McIntosh, B. (2019). A comparison of target detection algorithms using DSAC ATR algorithm development data set (Vol. 1098808, No. May 2019, p. 3). doi: 10.1117/12.2517423.
- [Millikan et al., 2018] url: <https://dsiac.org/databases/> Accessed (Online) Jun 2023.
- [Millikan et al., 2018] Millikan, B., Foroosh, H., & Sun, Q. (2018). Deep convolutional neural networks with integrated quadratic correlation filters for automatic target recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1222-1229).
- [d'Acremont et al., 2019] d'Acremont, A., Fablet, R., Baussard, A., & Quin, G. (2019). CNN-based target recognition and identification for infrared imaging in defense systems. *Sensors*, 19(9), 2040.
- [Arif & Mahalanobis, 2020] Arif, M., & Mahalanobis, A. (2020). Multiple view generation and classification of mid-wave infrared images using deep learning. *arXiv preprint arXiv:2008.07714*.
- [Arif & Mahalanobis, 2021] Arif, M., & Mahalanobis, A. (2021). Infrared target recognition using realistic training images generated by modifying latent features of an encoder-decoder network. *IEEE Transactions on Aerospace and Electronic Systems*, 57(6), 4448-4456.
- [Chen et al., 2021] Chen, H. W., Gross, N., Kapadia, R., Cheah, J., & Gharbieh, M. (2021, March). Advanced Automatic Target Recognition (ATR) with Infrared (IR) Sensors. In 2021 IEEE Aerospace Conference (50100) (pp. 1-13). IEEE.
- [McIntosh et al., 2020a] McIntosh, B., Venkataramanan, S., & Mahalanobis, A. (2020). Infrared target detection in cluttered environments by maximization of a target to clutter ratio (TCR) metric using a convolutional neural network. *IEEE Transactions on Aerospace and Electronic Systems*, 57(1), 485-496.
- [McIntosh et al., 2020b] McIntosh, B., Venkataramanan, S., & Mahalanobis, A. (2020, October). Target Detection in Cluttered Environments Using Infra-Red Images. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 2026-2030). IEEE.
- [Jiban et al., 2021] Jiban, M. J. H., Hassan, S., & Mahalanobis, A. (2021, September). Two-Stream Boosted TCRNet for Range-Tolerant Infra-Red Target Detection. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 1049-1053). IEEE.
- [Cuellar & Mahalanobis, 2021] Cuellar, A., & Mahalanobis, A. (2021, September). Detection of Small Moving Ground Vehicles in Cluttered Terrain Using Infrared Video Imagery. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 1099-1103). IEEE.
- [Baili, 2020] Baili, N. (2020). Automatic target recognition with convolutional neural networks.
- [Jocher, 2022] Jocher, G. (2022). ultralytics/yolov5. Retrieved from <https://github.com/ultralytics/yolov5>

# Decisive Data using Multi-Modality Optical Sensors for Advanced Vehicular Systems

Muhammad Ali Farooq<sup>1</sup>, Waseem Shariff<sup>1</sup>, Mehdi Sefidgar Dilmaghani<sup>1</sup>, Wang Yao<sup>1</sup>, Moazam Soomro<sup>2</sup>, and Peter Corcoran<sup>1</sup>

<sup>1</sup>*School of Engineering, University of Galway, Ireland*

<sup>2</sup>*College of Engineering and Computer Science, University of Central Florida, USA*

## Abstract

Optical sensors have played pivotal role in acquiring real world data for critical applications. This data when integrated with advanced machine learning algorithms provides meaningful information thus enhancing human vision. This paper focuses on various optical technologies for design and development of state-of-the-art out-cabin forward vision systems and in-cabin driver monitoring systems. The focused optical sensors include Long Wave Thermal Imaging (LWIR) cameras, Near Infrared (NIR), Neuromorphic/ event cameras, Visible CMOS cameras and Depth cameras. Further the paper discusses different potential applications which can be employed using the unique strengths of each these optical modalities in real time environment.

**Keywords:** LWIR, NIR, Event Cameras, Imaging, Image Processing, Machine Vision.

## 1 Introduction

Advanced vehicular systems rely on various optical modalities and hardware sensors to enhance driver safety, vehicles efficiency, and performance. Most common key technologies which are directly associated with next generation autonomous vehicles [Automotive, 2022] includes.

- Camera Vision Systems: Camera-based vision systems capture visual information from inside the vehicle and the vehicle's surroundings. They are used for in cabin applications such as adjusting car internal temperature adjustment according to driver and occupant requirements, lane departure warning, traffic sign recognition, pedestrian detection, and object tracking [Farooq et al., 2021], [Jia et al., 2008].
- Head-Up Display (HUD) and Augmented Reality (AR): HUD projects important information, such as speed, navigation instructions, and alerts, onto the windshield or a dedicated display unit. It makes easy for the driver to access vital information without taking their eyes off the road, enhancing safety and situational awareness. [Charassis and Papanastasiou, 2010]. Similarly AR can be used in vehicle windshields to overlay digital information, such as navigation cues, traffic data, and safety warnings, onto the real-world view. This technology enhances driver assistance and provides a more intuitive driving experience [Wang et al., 2020].
- Vehicle-to-Everything (V2X) Communication: V2X communication enables vehicles to exchange information with other vehicles, infrastructure, and pedestrians. Optical communication systems, such as visible light communication (VLC), is used to transmit data between vehicles or from traffic signals, enabling real-time cooperative driving, collision avoidance, and traffic management [Wang et al., 2019].

Keeping this in view in this study we will mainly discuss various camera technologies along with their respective strengths and weakness to elaborate their importance in challenging environmental and atmospheric conditions for advanced driver assistance systems (ADAS) and driver monitoring systems (DMS). Further we have highlighted the importance of machine learning algorithms to extract meaningful and decisive information that can aid drivers in real time driving experience.

## 2 Optical Sensor Technologies and Applications for Vehicular Systems

In this work we have mainly focused camera vision systems that can effectively employed and integrated with existing vehicular automation systems.

### 2.1 Visible CMOS Cameras

CMOS (complementary metal oxide semiconductor) camera is a digital camera that uses a CMOS image sensor, which is a device that converts light into an electrical signal. Generally, CMOS consists of the following parts: microlenses, color filters, metal lines, photodiodes, and substrates. CMOS cameras have several advantages compared with traditional CCD (charge-coupled device) cameras, including low power consumption, fast readout speeds, high integration, and low cost. These features allow CMOS cameras to be used in a variety of applications as follows.

- **Vehicle Detection and Tracking:** CMOS cameras are used to detect and position vehicles [He and Zhou, 2021] on the road. The captured images or video streams are analyzed by image processing and computer vision algorithms to identify and track vehicles on the road [Do and Yoo, 2019] for traffic monitoring, vehicle counting, and vehicle speed measurement.
- **Advanced Driver Assistance Systems (ADAS):** CMOS cameras are integral components of ADAS, which is used to warn drivers of any possible hazards on the road including lane departure warnings, forward collision warnings, pedestrian detection, and traffic sign recognition. These cameras capture the road ahead and provide real-time data to ADAS algorithms [Kuo et al., 2011], enabling the system to detect and provide assistance to the driver, which is important for improving traffic safety and preventing accidents.
- **Autonomous Vehicle Navigation:** CMOS cameras play a crucial role in autonomous vehicle navigation [Liu et al., 2008]. They provide visual input to the autonomous driving system, allowing the vehicle to understand its surroundings, detect and track other vehicles, and make decisions based on the detected objects and their trajectories.
- **Traffic Signal Control:** CMOS cameras can be used in traffic signal control systems to automatically adjust signal timing and dispensing by monitoring vehicle flow [Liu et al., 2012] and traffic conditions on the road in real-time to optimize traffic flow and reduce congestion.

Thus, CMOS cameras play a vital role in vehicular systems, providing efficient and accurate image acquisition and processing capabilities, supporting traffic monitoring, safety management, traffic flow analysis, and other applications, and contributing to traffic management and road safety.

### 2.2 LWIR Thermal Camera

LWIR (Long-Wave Infrared) cameras, also known as thermal infrared camera s, are designed to capture thermal radiation emitted by objects. They operate in the long-wave infrared spectrum covering the wavelengths ranging from  $8\mu\text{m}$  to  $14\mu\text{m}$  (8,000 to 14,000nm). Since LWIR cameras works by detecting the thermal infrared radiations also referred to as heat patterns, these cameras do not rely on external lighting conditions which makes them ideal for low lighting scenarios and even zero lightning conditions. The key applications of these camera are as follows.

- Night Vision: LWIR cameras can capture images in total darkness or low-light conditions. This feature makes them perfect choice for external roadside environmental monitoring by integrating these cameras with deep learning-based algorithms. We can find various published studies elaborating its extensive usage for external object detection [Farooq et al., 2021], object tracking, and lane detection [Farooq et al., 2023b].
- Vision in Adverse Conditions: LWIR cameras are less affected by factors like smoke, fog, dust, or poor visibility caused by environmental conditions. They can penetrate certain materials and provide visibility in situations where conventional CMOS based visible cameras are unable to produce robust results. This feature allows the drivers to get comprehensive roadside thermal perception even in harsh weather conditions.
- Temperature Measurement: LWIR cameras enable non-contact temperature measurement. They can accurately measure the temperature of objects, even in challenging environments or from a distance. This capability is valuable feature for in cabin application such as driver and occupant temperature monitoring thus adjusting internal temperature of the car. Moreover, this feature can be useful for drivers' drowsiness and fatigue detection thus generating timely alerts to avoid traffic collisions.

Figure 1 shows the out-cabin object detection results on various thermal frames acquired from publicly available thermal datasets [Farooq et al., 2023a].



Figure 1: Thermal Imaging Frames with Out-cabin Object Detection Results

### 2.3 NIR Camera

NIR (Near-Infrared) thermal cameras, also referred to as SWIR (Short-Wave Infrared) thermal cameras, can acquire thermal radiation in the near-infrared spectrum, typically ranging from  $0.9 \mu\text{m}$  to  $1.7 \mu\text{m}$  wavelength spectrum. Likewise, LWIR thermal cameras NIR cameras offer non-contact temperature measurement, and vision through obscuring elements. These features make them optimal choice for different in-cabin and out-cabin vehicular applications such as monitoring driver's attentiveness, detect signs of drowsiness or distraction, and issue alerts and timely warnings. Moreover, NIR cameras can monitor the interior environment of the vehicle as shown in the figure 2. They can detect the presence of occupants, their seating positions, and activity levels. This information can be utilized for personalized climate control, optimizing airbag deployment, or triggering alerts in case of unattended passenger such as children.



Figure 2: NIR output for monitoring the driver and passenger

### 2.4 Neuromorphic Sensors

Neuromorphic sensors, also known as event cameras, have a unique way of capturing scenes. Unlike traditional RGB cameras that capture multiple frames per second, event cameras have independent pixels that only respond

to changes in light, resulting in a stream of events rather than continuous images. Each event includes a timestamp, the pixel's position, and a binary polarity indicating an increase or decrease in light. This type of camera offers advantages such as speed, low memory usage, fewer computations, and privacy preservation, making it ideal for autonomous driving [Graça et al., 2023, Chen et al., 2020]. This research focuses on discussing the in-cabin and out-of-cabin vision applications of event cameras in the automotive industry.

- In-cabin applications: The event camera's high temporal resolution makes it an excellent choice for applications such as monitoring driver drowsiness in vehicles. Traditional cameras struggle to track subtle facial expressions, blink counting, head pose/gaze estimation, and saccadic movement analysis, which are crucial for assessing drowsiness levels. By utilizing event cameras, vehicles can move closer to achieving full autonomy. Researchers propose new algorithms to monitor drivers, analyze key factors in drowsiness, and address challenges in event camera-based monitoring systems [Ryan et al., 2021, Dilmaghani et al., 2022]. One approach involves using a LSTM-based system that detects driver distraction by creating 3D tensors from event camera output streams [Yang et al., 2022]. Another algorithm focuses on detecting yawning, an important indicator of drowsiness, using frames generated from event streams [Kielty et al., 2023].
- Out-cabin applications: Pedestrian detection is crucial for autonomous driving and various approaches have been proposed. In [Wan et al., 2021], a YOLO-v3 based architecture is suggested to detect pedestrians from frames generated by integrating event streams. Another study [Ojeda et al., 2020] introduces a hardware-efficient architecture using event streams, consisting of denoising the input stream and a simple neural network for pedestrian detection. Object detection is also important in next generation automotives, with [Wzorek and Kryjak, 2022] using YOLO-v4 to detect traffic signs, [Shariff et al., 2022] presenting a proof of concept for YOLO-based forward perception systems, and [Mentasti et al., 2022] proposing two event-based solutions for object detection and tracking in traffic monitoring, one based on geometrical schemes and the other on neural networks.



Figure 3: Event camera output samples: a) in-cabin driver monitoring [Ryan et al., 2021], b) out-cabin road side object detection [Shariff et al., 2022]

## 2.5 Depth Cameras

Depth cameras, also known as 3D cameras or depth sensing cameras, are a type of sensor used in automotive applications to capture depth information of the surrounding environment. These cameras operate by emitting infrared (IR) or laser light and measuring the time it takes for the light to bounce back after hitting objects in the scene. By analysing the stereo vision, time-of-flight or structured light patterns, depth cameras can create a three-dimensional representation of the scene [Khan et al., 2022, Khan et al., 2023, Kumar et al., 2020]. In the context of automotive, depth cameras offer several key features and benefits:

- Object Detection with Depth Sensing: Depth cameras provide accurate depth information, allowing for precise detection of moving objects. This enables advanced driver-assistance systems (ADAS) and autonomous vehicles to identify and monitor the position, size, and movement of pedestrians, vehicles, and

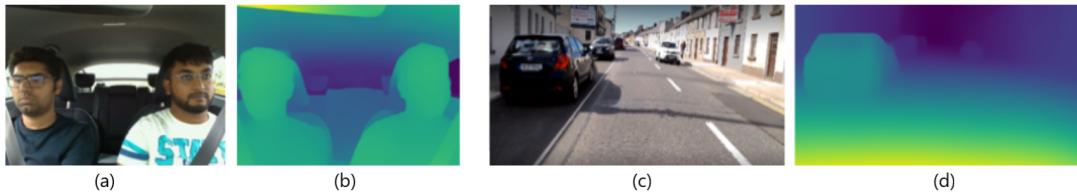


Figure 4: Depth Camera Sensor Outputs. a) In-cabin RGB camera b) Corresponding in-cabin depth output c) Out-cabin RGB camera d) Corresponding out-cabin depth output.

other obstacles on the road. This information is crucial for making real-time decisions and ensuring safe driving [Zhou et al., 2022].

- **Distance Measurement:** Depth cameras enable accurate measurement of distances between the vehicle and objects in the surrounding environment. This information is essential for adaptive cruise control systems, automatic braking systems, and parking assistance, as it helps determine the appropriate response or manoeuvre required to maintain a safe distance from other vehicles or objects [Kumar et al., 2020].
- **Scene Understanding and Mapping:** Depth cameras provide a detailed representation of the scene, allowing vehicles to understand the geometry and structure of the environment. This helps in creating high-definition maps, path planning, and navigation, especially in complex urban environments or during adverse weather conditions where traditional sensors may struggle [Mehrzed, 2023].
- **Gesture and Occupant Recognition:** Depth cameras can be utilized for driver monitoring and gesture recognition inside the vehicle. By analysing the depth information, these cameras can track the driver's gaze, head position, and hand movements, enabling features such as driver drowsiness detection, distraction monitoring, and intuitive in-car controls through gestures [Leu et al., 2011].
- **Enhanced Safety and Collision Avoidance:** The depth information captured by these cameras enhances overall safety by providing a more comprehensive understanding of the environment. This enables the development of collision avoidance systems that can detect potential hazards and take proactive measures to prevent accidents [Ding and Su, 2023].
- **Low-Light Performance:** Like LWIR thermal cameras, some depth cameras are designed to work in low-light conditions. They use active illumination, such as infrared light, to capture depth information, making them suitable for driving scenarios with limited visibility or at night [Saxena et al., 2008].

By combining depth information with other sensor data, such as radar and LiDAR, depth cameras contribute to a comprehensive perception system for autonomous vehicles. They enable accurate and robust scene understanding, enhancing the safety, efficiency, and overall driving experience in automotive applications.

### 3 Discussion

This section will summarize the potential advantages of different optical modalities along with their weaknesses in automotive scenarios.

- The visible CMOS sensor provides high-resolution images in good external lighting conditions however the performance of these cameras is severely effected in low-light conditions. The LWIR thermal camera excels in low-light and even in zero-lighting situations by detecting thermal signatures. The NIR camera enhances visibility in low-light, while thermal cameras outperform NIR cameras in complete darkness. The event camera aids in real-time object tracking with low-latency imaging, while the depth camera focuses on depth sensing and not optimized for night vision.

- For object detection, the visible CMOS sensor enables precise object recognition and classification with its high-resolution color imagery. In challenging lighting conditions, the LWIR thermal camera surpasses visible light cameras by detecting objects based on their thermal emissions. While the NIR camera is capable of object detection, thermal cameras generally outperform NIR cameras due to their ability to sense thermal signatures. The event camera, although primarily designed for real-time tracking, can contribute to overall perception systems by providing valuable visual information. Additionally, the depth camera can offer additional in-depth information, assisting in the identification of objects.
- In terms of facial recognition, the visible CMOS sensor may struggle in low-light conditions, while the LWIR thermal camera primarily captures thermal information rather than detailed facial features. The NIR camera is well-suited for facial recognition in low-light scenarios, capturing detailed facial features not easily visible to visible light cameras. The event camera's low-latency imaging capabilities can aid in real-time tracking of facial movements, and the depth camera can provide additional depth information for more accurate analysis of facial characteristics.
- For drowsiness and fatigue detection, the visible CMOS sensor has limited capability, while LWIR thermal cameras primarily capture thermal information rather than eye movements or facial features. NIR cameras are well-suited for monitoring driver drowsiness and fatigue due to their ability to analyze facial characteristics and detect eye movements even in low-light conditions. Event cameras have the potential to contribute to drowsiness and fatigue detection with their temporal information feature, but research in this area is limited. Additionally, depth cameras can provide additional depth information for a more comprehensive analysis of the driver's condition.
- In the domain of gesture recognition, the visible CMOS sensor can capture gestures but may require proper lighting conditions. LWIR thermal cameras excel in capturing thermal information but are not commonly used for gesture recognition. NIR cameras are well-suited for detailed hand movement capture, even in low-light conditions. Event cameras, with their low-latency imaging capabilities, contribute to real-time tracking of hand movements. Depth cameras, specifically designed for depth sensing, enable precise tracking of hand gestures for gesture recognition.
- When it comes to 3D mapping, the visible CMOS sensor has limited capability as it primarily captures color imagery without depth information. LWIR thermal cameras focus on capturing thermal information rather than depth and are not typically used for 3D mapping. Similarly, NIR cameras mainly capture color imagery without depth information and are not primarily used for 3D mapping. Event cameras provide valuable visual information for environment perception, while depth cameras, specifically designed for depth sensing, are ideal for creating accurate 3D maps of the vehicle's surroundings.

## 4 Conclusion and Future Work

In summary, the choice of optical sensor depends on the specific requirements for desired application in real-time environment. While visible CMOS sensors offer high-resolution colour imagery, thermal cameras excel in low-light and adverse weather conditions. NIR cameras provide enhanced visibility in low-light situations, event cameras offer real-time tracking capabilities, and depth cameras enable accurate 3D mapping. To make the best choice, it's essential to consider factors such as cost, performance requirements, and the specific needs of the application. By carefully evaluating these factors, one can determine the most appropriate optical sensor that strikes the right balance between functionality and affordability. Ultimately, selecting the right sensor can significantly impact the effectiveness and efficiency of a given system or application. As the possible future directions the array of multi modality optical sensors can be deployed in parallel with conventional sensors which includes LIDAR and RADAR for achieving maximum benefits for in-cabin as well as out-cabin applications.

## References

- [Automotive, 2022] Automotive, S. (2022). Synopsys automotive. <https://www.synopsys.com/automotive/autonomous-driving-levels.html> [Accessed: 31/05/2023].
- [Charissis and Papanastasiou, 2010] Charissis, V. and Papanastasiou, S. (2010). Human–machine collaboration through vehicle head up display interface. *Cognition, Technology & Work*, 12:41–50.
- [Chen et al., 2020] Chen, G., Cao, H., Conradt, J., Tang, H., Rohrbein, F., and Knoll, A. (2020). Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine*, 37(4):34–49.
- [Dilmaghani et al., 2022] Dilmaghani, M. S., Shariff, W., Ryan, C., Lemley, J., and Corcoran, P. (2022). Control and evaluation of event cameras output sharpness via bias. *arXiv preprint arXiv:2210.13929*.
- [Ding and Su, 2023] Ding, I.-J. and Su, J.-L. (2023). Designs of human–robot interaction using depth sensor-based hand gesture communication for smart material-handling robot operations. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 237(3):392–413.
- [Do and Yoo, 2019] Do, T.-H. and Yoo, M. (2019). Visible light communication-based vehicle-to-vehicle tracking using cmos camera. *IEEE Access*, 7:7218–7227.
- [Farooq et al., 2021] Farooq, M. A., Corcoran, P., Rotariu, C., and Shariff, W. (2021). Object detection in thermal spectrum for advanced driver-assistance systems (adas). *IEEE Access*, 9:156465–156481.
- [Farooq et al., 2023a] Farooq, M. A., Shariff, W., and Corcoran, P. (2023a). Evaluation of thermal imaging on embedded gpu platforms for application in vehicular assistance systems. *IEEE Transactions on Intelligent Vehicles*, 8(2):1130–1144.
- [Farooq et al., 2023b] Farooq, M. A., Shariff, W., O’callaghan, D., Merla, A., and Corcoran, P. (2023b). On the role of thermal imaging in automotive applications: A critical review. *IEEE Access*, 11:25152–25173.
- [Graça et al., 2023] Graça, R., McReynolds, B., and Delbruck, T. (2023). Shining light on the dvs pixel: A tutorial and discussion about biasing and optimization. *arXiv preprint arXiv:2304.04706*.
- [He and Zhou, 2021] He, J. and Zhou, B. (2021). Vehicle positioning scheme based on visible light communication using a cmos camera. *Optics express*, 29(17):27278–27290.
- [Jia et al., 2008] Jia, Z., Balasuriya, A., and Challa, S. (2008). Vision based data fusion for autonomous vehicles target tracking using interacting multiple dynamic models. *Computer vision and image understanding*, 109(1):1–21.
- [Khan et al., 2022] Khan, F., Farooq, M. A., Shariff, W., Basak, S., and Corcoran, P. (2022). Towards monocular neural facial depth estimation: Past, present, and future. *IEEE Access*, 10:29589–29611.
- [Khan et al., 2023] Khan, F., Shariff, W., Farooq, M. A., Basak, S., and Corcoran, P. (2023). A robust light-weight fused-feature encoder-decoder model for monocular facial depth estimation from single images trained on synthetic data. *IEEE Access*.
- [Kiely et al., 2023] Kiely, P., Dilmaghani, M. S., Ryan, C., Lemley, J., and Corcoran, P. (2023). Neuromorphic sensing for yawn detection in driver drowsiness. *arXiv preprint arXiv:2305.02888*.
- [Kumar et al., 2020] Kumar, G. A., Lee, J. H., Hwang, J., Park, J., Youn, S. H., and Kwon, S. (2020). Lidar and camera fusion approach for object distance estimation in self-driving vehicles. *Symmetry*, 12(2):324.

- [Kuo et al., 2011] Kuo, Y.-C., Pai, N.-S., and Li, Y.-F. (2011). Vision-based vehicle detection for a driver assistance system. *Computers & Mathematics with Applications*, 61(8):2096–2100.
- [Leu et al., 2011] Leu, A., Aiteanu, D., and Gräser, A. (2011). A novel stereo camera based collision warning system for automotive applications. In *2011 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 409–414. IEEE.
- [Liu et al., 2008] Liu, C., Chen, J., Xu, Y., and Luo, F. (2008). Intelligent vehicle road recognition based on the cmos camera. In *2008 IEEE Vehicle Power and Propulsion Conference*, pages 1–5. IEEE.
- [Liu et al., 2012] Liu, Z., Lin, S., Li, K., and Dong, A. (2012). Traffic flow video detection system based on line scan cmos sensor. *Advanced Science Letters*, 7(1):478–483.
- [Mehrzed, 2023] Mehrzed, S. (2023). Optimization of a simultaneous localization and mapping (slam) system for an autonomous vehicle using a 2-dimensional light detection and ranging sensor (lidar) by sensor fusion.
- [Mentasti et al., 2022] Mentasti, S., Kambale, A. W., and Matteucci, M. (2022). Event-based object detection and tracking-a traffic monitoring use case. In *Image Analysis and Processing. ICIAP 2022 Workshops: ICIAP International Workshops, Lecce, Italy, May 23–27, 2022, Revised Selected Papers, Part II*, pages 95–106. Springer.
- [Ojeda et al., 2020] Ojeda, F. C., Bisulco, A., Kepple, D., Isler, V., and Lee, D. D. (2020). On-device event filtering with binary neural networks for pedestrian detection using neuromorphic vision sensors. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3084–3088. IEEE.
- [Ryan et al., 2021] Ryan, C., O’Sullivan, B., Elrasad, A., Cahill, A., Lemley, J., Kielty, P., Posch, C., and Perot, E. (2021). Real-time face & eye tracking and blink detection using event cameras. *Neural Networks*, 141:87–97.
- [Saxena et al., 2008] Saxena, A., Chung, S. H., and Ng, A. Y. (2008). 3-d depth reconstruction from a single still image. *International journal of computer vision*, 76:53–69.
- [Shariff et al., 2022] Shariff, W., Farooq, M. A., Lemley, J., and Corcoran, P. (2022). Event-based yolo object detection: Proof of concept for forward perception system. *arXiv preprint arXiv:2212.07181*.
- [Wan et al., 2021] Wan, J., Xia, M., Huang, Z., Tian, L., Zheng, X., Chang, V., Zhu, Y., and Wang, H. (2021). Event-based pedestrian detection using dynamic vision sensors. *Electronics*, 10(8):888.
- [Wang et al., 2019] Wang, J., Shao, Y., Ge, Y., and Yu, R. (2019). A survey of vehicle to everything (v2x) testing. *Sensors*, 19(2):334.
- [Wang et al., 2020] Wang, Z., Han, K., and Tiwari, P. (2020). Augmented reality-based advanced driver-assistance system for connected vehicles. In *2020 ieee international conference on systems, man, and cybernetics (SMC)*, pages 752–759. IEEE.
- [Wzorek and Kryjak, 2022] Wzorek, P. and Kryjak, T. (2022). Traffic sign detection with event cameras and dcnn. In *2022 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 86–91. IEEE.
- [Yang et al., 2022] Yang, C., Liu, P., Chen, G., Liu, Z., Wu, Y., and Knoll, A. (2022). Event-based driver distraction detection and action recognition. In *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 1–7. IEEE.
- [Zhou et al., 2022] Zhou, D., Song, X., Fang, J., Dai, Y., Li, H., and Zhang, L. (2022). Context-aware 3d object detection from a single image in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18568–18580.

# Navigating Uncertainty: The Role of Short-Term Trajectory Prediction in Autonomous Vehicle Safety

Sushil Sharma<sup>1,2</sup>, Ganesh Sistu<sup>2</sup>, Lucie Yahiaoui<sup>2</sup>, Arindam Das<sup>1,3</sup>,  
Mark Halton<sup>1</sup> and Ciarán Eising<sup>1</sup>

<sup>1</sup>*firstname.lastname@ul.ie* and <sup>2,3</sup>*firstname.lastname@valeo.com*

<sup>1</sup>*University of Limerick, Ireland*, <sup>2</sup>*Valeo Vision Systems, Ireland*, <sup>3</sup>*DSW, Valeo India*

## Abstract

Autonomous vehicles require accurate and reliable short-term trajectory predictions for safe and efficient driving. While most commercial automated vehicles currently use state machine-based algorithms for trajectory forecasting, recent efforts have focused on end-to-end data-driven systems. Often, the design of these models is limited by the availability of datasets, which are typically restricted to generic scenarios. To address this limitation, we have developed a synthetic dataset for short-term trajectory prediction tasks using the CARLA simulator. This dataset is extensive and incorporates what is considered complex scenarios - pedestrians crossing the road, vehicles overtaking - and comprises 6000 perspective view images with corresponding IMU and odometry information for each frame. Furthermore, an end-to-end short-term trajectory prediction model using convolutional neural networks (CNN) and long short-term memory (LSTM) networks has also been developed. This model can handle corner cases, such as slowing down near zebra crossings and stopping when pedestrians cross the road, without the need for explicit encoding of the surrounding environment. In an effort to accelerate this research and assist others, we are releasing our dataset and model to the research community. Our datasets are publicly available on <https://github.com/sharmasushil/Navigating-Uncertainty-Trajectory-Prediction>.

**Keywords:** Trajectory Prediction, CNN-LSTM and CARLA simulator

## 1 Introduction

Autonomous vehicles are revolutionizing the transportation industry with a core focus on improved safety and an enhanced driver experience. At the same time, ensuring the safe manoeuvring of autonomous vehicles in real-world scenarios remains very challenging. One of the key components of autonomous vehicle safety is the ability to accurately predict short-term trajectories that allow the host vehicle to navigate through uncertain and dynamic environments [Zhao and Malikopoulos, 2020]. Short-term trajectory prediction refers to the estimation of the future position and movement within a limited time frame of a vehicle. By accurately perceiving the ego vehicle trajectory, an autonomous vehicle can anticipate potential hazards and proactively plan its actions to avoid collisions [Dixit et al., 2021] or respond to risky situations[Botello et al., 2019]. Significant progress has been made in recent years in developing trajectory prediction models for autonomous vehicles [Venkatesh et al., 2023]. To make predictions about the future movements of surrounding entities, these models employ diverse data sources, including sensor data like LiDAR, radar, and cameras. Machine learning techniques, including deep neural networks [Fayyad et al., 2020], have proven to be effective in capturing complex spatiotemporal patterns [Chen et al., 2023] and improving trajectory prediction accuracy. By analyzing and understanding previous data, researchers and engineers can identify common patterns, critical factors, and potential risks associated with trajectory prediction [Cui et al., 2019]. This knowledge can guide the development of more reliable and effective prediction algorithms, leading to enhanced safety measures

[Atakishiyev et al., 2021] and increased public trust in autonomous vehicles. In this study, we aim to investigate the role of short-term trajectory prediction in ensuring the safety of autonomous vehicles. The main contributions of this paper are as follows:

1. Short-term trajectory prediction of the vehicle from only perspective view images with no explicit knowledge encoding.
2. A novel dataset\* to encourage the research community to pursue the direction of end-to-end implicit vehicle trajectory prediction learning methods.

## 2 Related Work

Several studies have focused on the crucial role of short-term trajectory prediction in ensuring the safety of autonomous vehicles[Yalamanchi et al., 2020]. One notable work is the research conducted by [Zhu et al., 2019]. The authors proposed the use of Recurrent Neural Networks (RNNs) to accurately forecast the future trajectories of surrounding objects in complex driving environments. By training their model on real-world driving datasets, they achieved impressive trajectory prediction results, allowing autonomous vehicles to navigate with improved safety and awareness. Another relevant study by [Hegde et al., 2020], explored the application of Generative Adversarial Networks (GANs) for probabilistic trajectory prediction. By leveraging GANs, the researchers were able to generate multiple plausible future trajectories for vehicles, incorporating uncertainty into the prediction process. This probabilistic approach provides valuable information for autonomous vehicles to assess potential risks and make informed decisions in dynamic traffic situations. Furthermore, the work of [Li et al., 2019], emphasized the significance of considering interaction information between vehicles and pedestrians. The researchers proposed a novel interaction-aware trajectory prediction model [Krüger et al., 2020] that effectively captured the mutual influence and dependencies between different road users. By incorporating interaction information, the model achieved improved accuracy [Deo and Trivedi, 2018] and reliability in trajectory prediction, contributing to enhanced safety in autonomous driving scenarios. These studies collectively highlight the importance of short-term trajectory prediction in autonomous vehicle safety. Our study aims to address the limitations encountered in previous studies by incorporating frequently occurring critical actors - pedestrians at the crosswalks, keeping the volume of the dataset minimal and balanced while ensuring all critical cases are covered. To address these challenges, we have created a novel dataset using the Carla simulation platform. By leveraging this synthetic dataset, we expect our model to achieve enhanced performance. This approach enables us to overcome the constraints related to data collection and processing, thereby augmenting the results of our study.

## 3 Methodology

### 3.1 Architecture Topology

The network developed in this work is designed to predict the future trajectories of a vehicle based on a sequence of perspective-view images. It comprises two primary components: a Convolutional Neural Network (CNN) and a Long-Short Term Memory Network (LSTM). CNN plays a crucial role in extracting essential features from the input image sequence using convolution, a mathematical operation that filters the data to capture specific features. These deep features obtained from the CNN serve as inputs to the LSTM, which specializes in temporal prediction tasks by capturing long-term dependencies and maintaining memory over time. The LSTM network learns to infer future positions of the vehicle within the predicted trajectory based on the extracted deep features and the input image sequence.

Overall, the architecture of the network is shown in Figure 1, which provides a visual summary of how the CNN and LSTM components are connected. The architecture topology diagram illustrates the CNN used

\*Dataset: <https://drive.google.com/drive/folders/1JPb64bGV88ymZkJrUBaKQg12tToZVF7T?usp=sharing>

for trajectory estimation from a sequence of  $n$  input images. The left side of the diagram shows the input layer of the network where the image sequence is fed into the CNN. The images are then processed through a series of convolutional layers, pooling layers, and activation functions to extract features from the images. The extracted features are then passed through one or more fully connected layers to estimate the trajectory, which is displayed on the right side of the diagram. We use a custom encoder here since this work aims to provide a solution that can be deployed on a low-resource device within less than 1 TOPS. However, given no constraints on the device, any state-of-the-art encoder as a backbone can be integrated into the feature extraction stage.

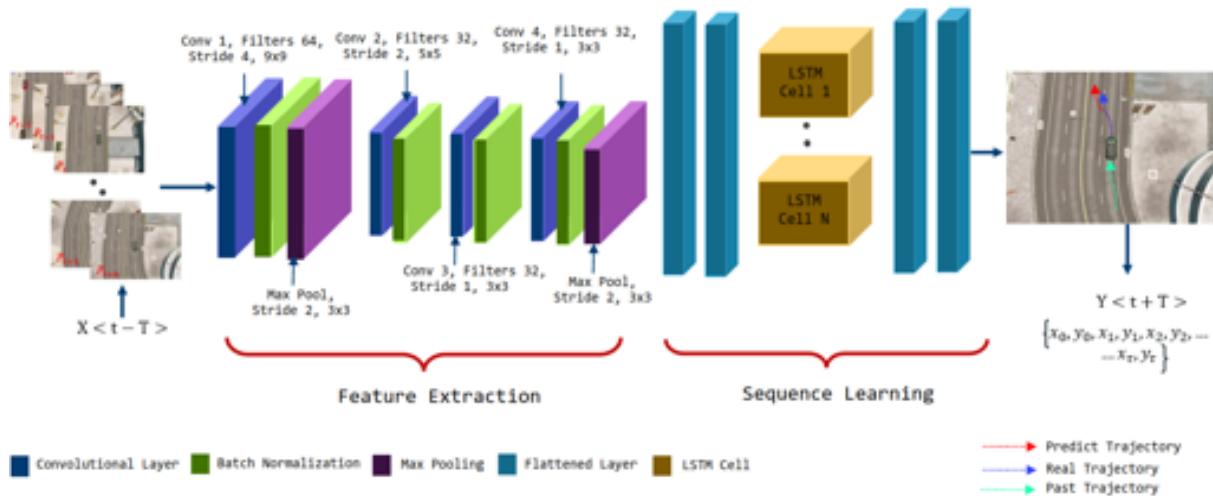


Figure 1: Proposed CNN architecture topology diagram for trajectory estimation from image sequences, showcasing the flow of information through convolutional and fully connected layers.

### 3.2 Trajectory prediction approach

We define a sequence as a sequence of  $n$  images as  $X^{<t-n>}$  at time  $t$ . The purpose of  $X^{<t-n>}$  is to anticipate future trajectory position.

$$Y^{<t+n>} = \{x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3, \dots, x_n, y_n\} \quad (1)$$

where  $x$  and  $y$  depict the location of the ego vehicle. The distance feedback  $d_1^{<t-n>}$  measures the distance between the current pose of the ego vehicle  $P_{ego}$  and the final pose in the sequence  $P_{dest}$  and expressed in (2). This objective's goal is to shorten the ego vehicle's local travel path.

$$d_1^{<t+n>} = \sum_{i=1}^{n_0} \|P_{ego}^{<t+i>} - P_{dest}^{<t+n>}\|_2^2 \quad (2)$$

(3) demonstrates how to derive the lateral velocity  $V_{lat}$  from the angular velocity of the ego vehicle  $v_\delta$ . The main objective in terms of trajectory prediction is to anticipate the future path of the ego vehicle based on its current state and inputs. By understanding the vehicle's lateral velocity  $V_{lat}$  and angular velocity  $v_\delta$ , we can estimate its trajectory and predict its future position and orientation.

$$V_{lat}^{<t+n>} = \sum_{i=1}^{n_0} v_\delta^{<t+n>} \quad (3)$$

(4) defines the  $y$  direction velocity component, which is used to determine the longitudinal velocity as  $V_{long}$ . The speed is set at 30 kph. The main objective is to calculate the future longitudinal velocity denoted as  $V_{long}^{<t+n>}$ ,

based on the sum of forward velocities, represented as  $v_f^{}$ , over a specific time horizon.

$$V_{long}^{} = \sum_{i=1}^{n_0} v_f^{} \quad (4)$$

Root mean squared error (RMSE) is also being tested as an objective, which is represented by the Euclidean distance between the predicted position of the ego vehicle  $\hat{P}_{ego}$  for a given trajectory at a specific time-step and the actual position of the ego vehicle  $P_{ego}$  at that time-step, as seen in equation (5).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{P}_{ego} - P_{ego})^2}{n}} \quad (5)$$

### 3.3 Why the CARLA simulation platform?

We have chosen to utilize the CARLA simulation platform [Dosovitskiy et al., 2017] for creating our dataset for a number of reasons. It is a dedicated open-source simulator for autonomous driving research. When compared to other simulation platforms such as Carsim [Dupuy et al., 2001], MATLAB [Herrera et al., 2016], and Gazebo [Mengacci et al., 2021], the CARLA simulation stands out for its comprehensive feature set, realistic simulation, customization options, and support for multi-sensor data generation.

A comparison of each of the aforementioned simulation platforms is given in Table 1.

Requirement	Simulation			
	CARLA	CarSim	MATLAB	Gazebo
Perception: sensor models supported	✓	✓	✓	✓
Perception: support for various weather conditions	✓	✗	✗	✗
Camera Calibration	✓	✗	✓	✓
3D Virtual Environment	✓	✓	✗	✓
Path planning	✓	✓	✓	✓
Traffic scenarios simulation	✓	✓	✓	✗
Scalability via a server multi-client architecture	✓	✗	✗	✗
Portability such as window and Linux	✓	✓	✓	✓
Flexible API	✓	✓	✓	✓
Open Source	✓	✗	✗	✓

Table 1: Comparison of various simulators.

## 4 Experimental Setup

### 4.1 Implementation Details

A CNN-LSTM model for trajectory prediction is implemented using the TensorFlow framework and trained on perspective view images from the CARLA simulation. The first step is preprocessing, which involves resizing, normalizing, and splitting the dataset into a ratio of 60% : 20%: 20% for training, validation and testing respectively. The next step involves a CNN extracting spatial features, which are then fed into an LSTM to model for temporal dependencies. The training process involves utilizing suitable loss functions to train the model effectively. In the case of trajectory prediction, one commonly used loss function is the Mean Squared Error (MSE) loss. This particular loss function calculates the average of the squared differences between the predicted trajectories and the ground truth trajectories. By doing so, it penalizes larger deviations between the predicted and actual trajectories, we specifically employed the Adam optimization algorithm [Liu et al., 2023]. Table 2 outlines the hyperparameters employed to train the proposed CNN-LSTM networks. The optimal values of these hyperparameters are obtained experimentally to maximize the performance and increase the generalization capabilities of the model.

S.No	Parameter Name	Optimal Values
1	Batch Size	(50,75,100)
2	Epochs	(10,20,30,40)
3	Loss Function	(Mean square error (MSE))
4	Momentum	(0.8,0.85,0.9)
5	Optimizer	(Nadam, Adam)
6	LSTM cells	(1,2,3,4)
7	LSTM Dropout	(0.25,0.3,0.35,0.4)
8	Hidden Units	(100,125,150,175,200)
9	CNN Flattened 1	(256,512,768,1024)
10	CNN Flattened 2	(256,512,768,1024)
11	LSTM Flattened 1	(64,128,256,512)
12	LSTM Flattened 2	(64,128,256,512)
13	Flattened Dropout	(0.05,0.1,0.15,0.2,0.25)

Table 2: The table provides a list of hyperparameters that include batch size, loss function, optimiser, and others for our proposed training setup.

## 4.2 Dataset Generation

The specifics of how the datasets were created using the CARLA simulator are described in this section. In the simulator, we adjusted the camera position to achieve a top-down view, enabling us to capture comprehensive 360° and Bird's Eye View (BEV) perspectives of each scene. Note that the utilization of the top-view image approach for trajectory prediction in autonomous vehicles provides a comprehensive and detailed comprehension of the surrounding environment, which in turn aids in making accurate decisions. Each image has dimensions of 800 pixels width and 600 pixels height. The field of view (FOV) for the camera is set to 90°. The CARLA camera position and orientation are defined as `cam_rotation = (-90, 0, -90)` and `cam_location = (0, 0, 15)`. If we consider a camera positioned 15 meters above the host vehicle, it would imply that the camera is mounted at a height of 15 meters from the ground level. This positioning suggests that the camera is elevated significantly above the vehicle, capturing a top-down or bird's eye view perspective of the surrounding environment.

The initial dataset, referred to as "Level 1", consists of 1000 perspective view images. To achieve a more enhanced real-life simulation, ego vehicles and pedestrians were incorporated into each scene. In addition, we annotated each image with supplementary details such as speed and local coordinates (x,y and z). Similarly, the subsequent, more challenging dataset, referred to as "Level 2", consists of 5000 images. In order to generate more intricate and diverse scenarios, the number of vehicles and pedestrians was augmented in each scene. This deliberate augmentation aimed to ensure that our machine-learning models could effectively handle a wide range of real-life scenarios. As done previously, all images in this dataset were annotated with supplementary information, as presented in Figure 2.

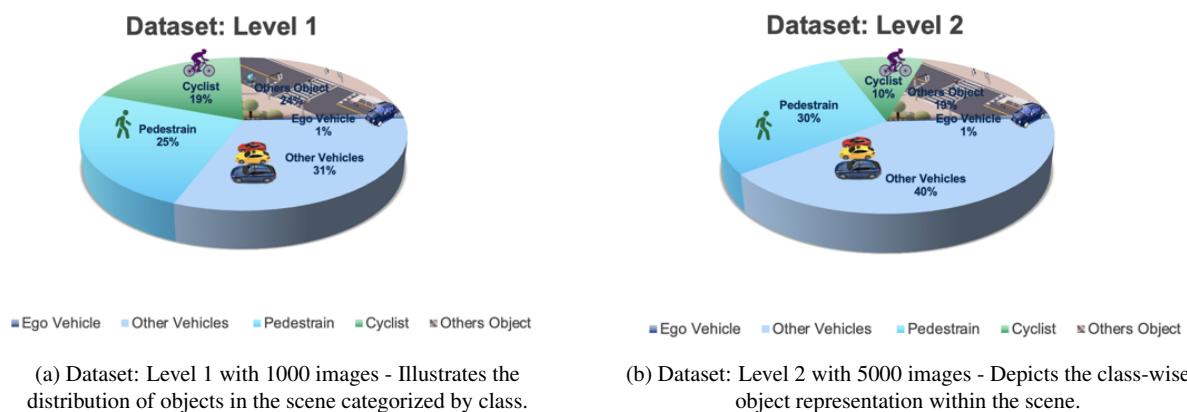


Figure 2: Data statistics for each class of our newly created dataset

## 5 Results

In the Carla simulation, we define an IMU sensor that captures linear acceleration ( $\text{m/s}^2$ ) and angular velocity data (rad/sec). The IMU sensor records the frame number at which each measurement occurs, enabling us to determine the simulation start time accurately. By utilizing a sensor, we can collect real-time measurements of linear acceleration and angular velocity throughout the simulation. The time elapsed from when the simulation started to when the measurement was taken was also recorded. The GNSS sensor provides the position and rotation of the sensor at the time of measurement, relative to the vehicle coordinates. The position is typically represented in meters, while the rotation is expressed in degrees. A comparison between the ground truth trajectory and our predicted trajectory is shown in Figure 3(a). In Figure 3(b), we focus on a specific critical scenario where a pedestrian enters the road, leading to vehicles coming to a halt. For more insight, a demo video on this real-life scenario can be checked in detail at <https://youtu.be/DZDqGbInko?t=31>

An ablation study on the number of LSTM cells ( $\alpha=1$ ,  $\beta=2$ ,  $\gamma=3$ ,  $\delta=4$ ) is conducted on our CNN-LSTM model. This comparison was performed using the CARLA dataset for the two specified levels, Level 1 and Level 2 respectively. For this analysis, three evaluation metrics - RMSE, MAPE, and AED. A summary of the results is shown in Table 3.

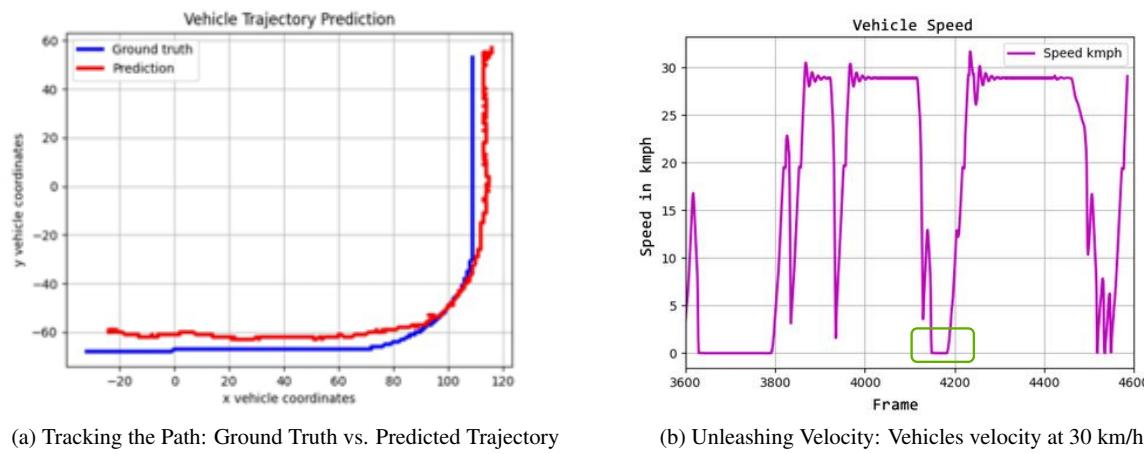


Figure 3: CARLA simulation results

CARLA: Dataset	Model	ARMSE	AMAPE	AED
<b>Dataset:</b> Level 1	CNN-LSTM ( $\alpha$ )	0.0046	0.0056	0.0050
	CNN-LSTM ( $\beta$ )	0.0034	0.0043	0.0039
	CNN-LSTM ( $\gamma$ )	0.0028	0.0038	0.0032
	<b>CNN-LSTM (<math>\delta</math>)</b>	<b>0.0024</b>	<b>0.0033</b>	<b>0.0028</b>
<b>Dataset:</b> Level 2	CNN-LSTM ( $\alpha$ )	0.0126	0.0172	0.0154
	CNN-LSTM ( $\beta$ )	0.0097	0.0133	0.0127
	CNN-LSTM ( $\gamma$ )	0.0082	0.0119	0.0107
	<b>CNN-LSTM (<math>\delta</math>)</b>	<b>0.0065</b>	<b>0.0107</b>	<b>0.0079</b>

Table 3: Analysis of ARMSE, AMAPE, and AED

One corner case is presented in Figure 4. For visual purposes, a blue bounding box represents the position of the object as per the ground truth and a red bounding box is used to highlight the prediction of the same object from our proposed model. Moreover, our model demonstrates exceptional performance in predicting the frame at time  $t + 5$ , closely aligning with the ground truth  $t$ . Notably, it excels even in challenging scenarios, such as when a vehicle is navigating a bend and a pedestrian unexpectedly appears. The model effectively

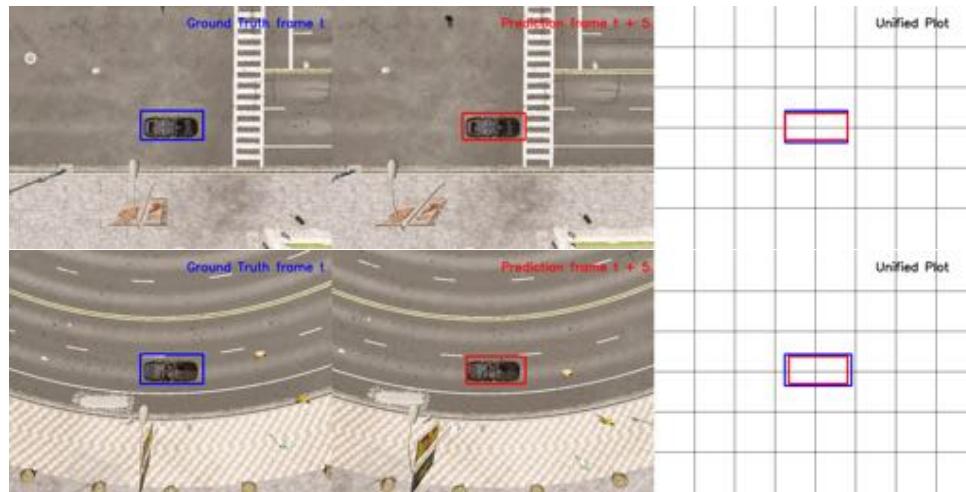


Figure 4: Demonstrating robustness of our model in handling corner cases, such as pedestrian crossings, without explicit encoding of surrounding knowledge. The comparison of Ground Truth (blue) and Predicted Bounding Boxes (red) for trajectory prediction

captures and comprehends the unpredictable behaviour of pedestrians when crossing the road, showcasing its remarkable trajectory prediction capabilities. Our model performance in these critical scenarios is demonstrated at <https://youtu.be/DZDqGbInko>

## 6 Conclusions

In this work, we have developed a novel single-stage end-to-end deep network proposal for short-term vehicle trajectory prediction. First, we introduce a CNN-LSTM network topology for trajectory prediction, leveraging its effectiveness in handling complex stochastic tasks. Then we generate a large synthetic dataset using the CARLA simulator, providing a valuable resource for training and evaluating trajectory prediction models in a supervised learning fashion with a focus on safety. The data-driven approach presented in this paper offers a scalable alternative to traditional rule-based optimization algorithms, paving the way for further advancements in the field. The provided synthetic dataset serves as a baseline for future research, encouraging the research community to compare their models against the proposed methodology. We hope this effort will foster innovation and drive improvements in trajectory prediction models.

## Acknowledgments

This article has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049.

## References

- [Atakishiyev et al., 2021] Atakishiyev, S., Salameh, M., Yao, H., and Goebel, R. (2021). Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*.
- [Botello et al., 2019] Botello, B., Buehler, R., Hankey, S., Mondschein, A., and Jiang, Z. (2019). Planning for walking and cycling in an autonomous-vehicle future. *Transportation Research Interdisciplinary Perspectives*, 1:100012.

- [Chen et al., 2023] Chen, H., Wang, J., Shao, K., Liu, F., Hao, J., Guan, C., Chen, G., and Heng, P.-A. (2023). Traj-mae: Masked autoencoders for trajectory prediction. *arXiv preprint arXiv:2303.06697*.
- [Cui et al., 2019] Cui, H., Radosavljevic, V., Chou, F.-C., Lin, T.-H., Nguyen, T., Huang, T.-K., Schneider, J., and Djuric, N. (2019). Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE.
- [Deo and Trivedi, 2018] Deo, N. and Trivedi, M. M. (2018). Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*.
- [Dixit et al., 2021] Dixit, A., Kumar Chidambaram, R., and Allam, Z. (2021). Safety and risk analysis of autonomous vehicles using computer vision and neural networks. *Vehicles*, 3(3):595–617.
- [Dosovitskiy et al., 2017] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*.
- [Dupuy et al., 2001] Dupuy, S., Egges, A., Legendre, V., and Nugues, P. (2001). Generating a 3d simulation of a car accident from a written description in natural language: The carsim system. *arXiv preprint cs/0105023*.
- [Fayyad et al., 2020] Fayyad, J., Jaradat, M. A., Gruyer, D., and Najjaran, H. (2020). Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, 20(15).
- [Hegde et al., 2020] Hegde, C., Dash, S., and Agarwal, P. (2020). Vehicle trajectory prediction using gan. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*.
- [Herrera et al., 2016] Herrera, A. M., Suhandri, H. F., Realini, E., Reguzzoni, M., and de Lacy, M. C. (2016). gogps: open-source matlab software. *GPS solutions*, 20:595–603.
- [Krüger et al., 2020] Krüger, M., Novo, A. S., Nattermann, T., and Bertram, T. (2020). Interaction-aware trajectory prediction based on a 3d spatio-temporal tensor representation using convolutional–recurrent neural networks. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1122–1127. IEEE.
- [Li et al., 2019] Li, X., Ying, X., and Chuah, M. C. (2019). Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving. *arXiv preprint arXiv:1907.07792*.
- [Liu et al., 2023] Liu, M., Yao, D., Liu, Z., Guo, J., Chen, J., et al. (2023). An improved adam optimization algorithm combining adaptive coefficients and composite gradients based on randomized block coordinate descent. *Computational Intelligence and Neuroscience*, 2023.
- [Mengacci et al., 2021] Mengacci, R., Zambella, G., Grioli, G., Caporale, D., Catalano, M. G., and Bicchi, A. (2021). An open-source ros-gazebo toolbox for simulating robots with compliant actuators. *Frontiers in Robotics and AI*, 8.
- [Venkatesh et al., 2023] Venkatesh, N., Le, V.-A., Dave, A., and Malikopoulos, A. A. (2023). Connected and automated vehicles in mixed-traffic: Learning human driver behavior for effective on-ramp merging. *arXiv preprint arXiv:2304.00397*.
- [Yalamanchi et al., 2020] Yalamanchi, S., Huang, T.-K., Haynes, G. C., and Djuric, N. (2020). Long-term prediction of vehicle behavior using short-term uncertainty-aware trajectories and high-definition maps. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6.
- [Zhao and Malikopoulos, 2020] Zhao, L. and Malikopoulos, A. A. (2020). Enhanced mobility with connectivity and automation: A review of shared autonomous vehicle systems. *IEEE Intelligent Transportation Systems Magazine*, 14(1):87–102.
- [Zhu et al., 2019] Zhu, J., Qin, S., Wang, W., and Zhao, D. (2019). Probabilistic trajectory prediction for autonomous vehicles with attentive recurrent neural process. *arXiv preprint arXiv:1910.08102*.

# Aerially Determined Dynamic Environment Mapping for Enhanced Road Vehicle Awareness

Brendan Halligan<sup>2</sup>, Dara Molloy<sup>1</sup>, Edward Jones<sup>1</sup>, Brian Deegan<sup>1</sup>, Martin Glavin<sup>1</sup> and Liam Kilmartin<sup>2</sup>

<sup>1</sup>*Connaught Automotive Research (CAR) Group*

<sup>2</sup>*College of Engineering and Science*

*University of Galway*

*Ireland*

## Abstract

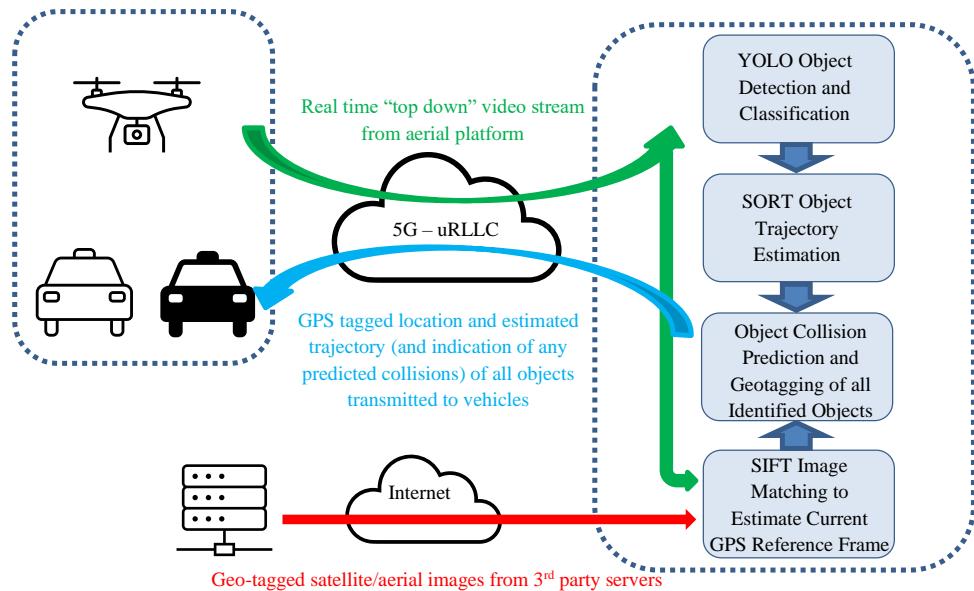
Autonomous vehicles require real-time data from several onboard sensors and cameras to navigate their environment. Aerial dynamic environment mapping information can provide additional detailed representations of a local road environment allowing for more robust operation of the algorithms utilised to control autonomous road vehicles. This paper proposes an OpenCV-based framework that utilises a YOLOv7 object detection model, trained upon a custom dataset achieving a mean Average Precision (mAP) @0.5 score of 95.6%, to derive dynamic environment mapping information from an aerial video stream. The framework detects and tracks cars, cyclists, and pedestrians, determining their instantaneous trajectories with an accuracy of 96%, 92% and 78% respectively. Real-time analysis of these trajectories allows for the accurate identification of potential collisions. The framework also accurately estimates each tracked object's geographic coordinates by comparing the aerial video stream with available 3<sup>rd</sup> party geo-tagged imagery.

**Keywords:** Aerial Environment Mapping Information, Location Mapping, Autonomous Vehicle, Machine Vision

## 1 Introduction

The reliable performance of autonomous road vehicles and driver assistance technologies depends heavily on real-time sensor and camera data, which can be limited by weather conditions, occlusions, sensor failures, positioning and orientation of the sensors\cameras. Aerially determined (e.g. using cameras deployed on Unmanned Aerial Vehicles (UAVs) such as drones and dirigibles) dynamic environment mapping (ADEM) information can provide an alternative and potentially more accurate and reliable representation of the local driving environment thus aiding autonomous vehicle software in making more robust decisions. Such a system, where cameras are deployed on a UAV to provide a “top down” view of a road scene, could be used to enhance or harden local environment mapping information provided by systems using infrastructure and vehicle mounted cameras/sensors. In addition, such an aerial based system could also provide a low cost and quickly deployable system for use in situations such as when temporary roadworks or traffic flow control restrictions are in place. ADEM could detect and track other road users and obstacles beyond the range of the vehicle’s sensors and incorporate relevant data such as geographic coordinates. This paper proposes an algorithmic framework which utilises YOLOv7 and OpenCV to determine such dynamic environment mapping information from a video stream captured from an aerial (i.e. drone) platform. Figure 1 provides both a high-level overview of the proposed framework’s architecture and a possible deployment scenario where all computationally demanding\energy intensive processing is completed on a cloud-based server with appropriate 5G network services (e.g. Ultra-Reliable Low Latency Communication) providing the required inter-connection between the system components. The proposed framework required the training and validating of a YOLOv7 model using a custom video dataset. An OpenCV-based algorithm was developed to detect and track these objects and to instantaneously predict potential collisions through the analysis of the trajectory of each tracked object. The framework also utilises Scale-Invariant Feature Transform (SIFT)

key-point matching in combination with Bilinear Interpolation to map each tracked object to real-world geographic coordinates.



**Figure 1 – Overview of Algorithm Framework and Potential System Deployment Model**

## 2 Related Work

The use of ADEM technology to enhance the awareness of road vehicles, particularly in the context of Intelligent Transport Systems (ITS), is a rapidly emerging area of research. Several studies have recently been conducted on object detection models for autonomous vehicle applications. YOLOv7 has shown promise in achieving high accuracy while maintaining low inference times, making it a suitable choice for such applications. YOLOv7 is an extension to the You Only Look Once (YOLO) family of object detection models and it has achieved state-of-the-art performance of several benchmark datasets. The YOLOv7 model improved upon previous versions of YOLO by incorporating a ResNet backbone, feature pyramid network, and enhanced anchor generation. It surpasses all known object detectors in both speed and accuracy in the range from 5 FPS to 160 FPS and has the highest accuracy 56.8% AP among reported real-time object detectors operating at 30 FPS or higher. [Wang et al., 2022]. The use of unmanned aerial vehicles (UAVs) in tandem with YOLO-based architectures for object detection in aerial imaging has been comprehensively documented in [Koay et al., 2021] [Luo et al., 2022]. UAVs offer the potential to collect a vast, encompassing quantum of detailed spatial information in real time at a relatively low cost [Fornace et al., 2014]. The deployment of UAVs for the derivation of ADEM in an automotive context is a relatively novel field. A localization technique using aerial imagery maps and LIDAR-based reflectivity for autonomous vehicles in urban environments [Vora et al., 2020] introduces the potential for environment mapping in an automotive setting. The framework proposed in this paper differs from that previous work in focussing on an object detection-based dynamic map, including mapping to geographic coordinates. A SIFT key-point matcher [Lowe, 2004] was utilised to perform this comparison, in a broadly similar manner to other work such as the use of SIFT for GPS location mapping at a street view level, in tandem with satellite imagery in [Yicong et al., 2017].

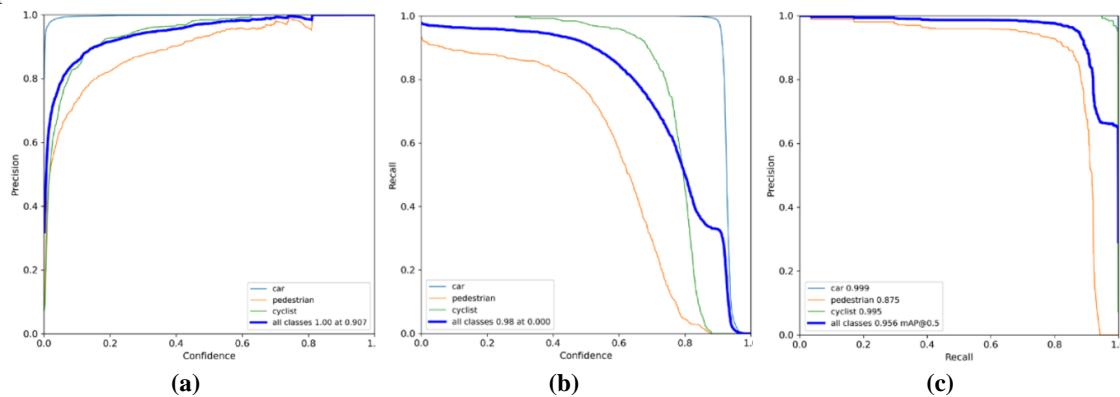
## 3 Algorithm Architecture and Implementation

This section details the steps taken to implement the core application features, including the acquisition and

training of a highly accurate YOLOv7 objects detection model. The design and development of an OpenCV-based application to facilitate object tracking, via the Simple Online and Realtime Tracking (SORT) algorithm, and instantaneous collision detection is outlined. The use of SIFT and bilinear interpolation to implement location mapping for tracked objects is also documented.

### 3.1 YOLOv7 Model Acquisition and Evaluation

Deep learning approaches use neural networks for feature detection of objects and their subsequent classification into one of a pre-determined set of classes or labels. YOLOv7 offers high accuracy and low latency compared to other object detection models, which makes it suitable for autonomous vehicle applications where lower inference times are particularly critical. For the application-at-hand, it was necessary to train a YOLOv7 model on a custom annotated dataset, as opposed to an open-source dataset such as the Microsoft Common Objects in Context (MS COCO) dataset. This was due to the need for accurate object detection from a “birds-eye” perspective, as opposed to the ground level view offered by MS COCO. A custom dataset was acquired through the annotation of 600 randomized images from our captured aerial video footage using the Computer Vision Annotation Tool (CVAT). A YOLOv7 model was then trained via Google Colaboratory for 200 epochs. The following metrics were obtained from the training process.



**Figure 2 – (a): YOLOv7 Model Precision Curve – (b): Recall Curve – (c): Precision-Recall Curve**

The precision curve shown in Figure 2(a) illustrates the proportion of true positive predictions out of all the positive predictions made by the model. The model is accurate with a precision of 90.7%. The recall curve in Figure 2(b) measures the proportion of true positive predictions out of all the actual positive instances in the data with the model achieving a 98% score for recall. The Precision-Recall curve graphs precision as a function of recall. The model recorded a mean mAP@0.5 score of 95.6%, as per Figure 2(c). For reference, MS COCO has a typical mAP@0.5 of approximately 60%. The confusion matrix shown in Table 1 summarizes the performance of the classification model by reporting the proportion of true and false positives across each class. This study focused on developing a framework to generate ADEM information relating to Car, Cyclist and Pedestrian classes of objects. The values on the diagonal of the matrix in Table 1 detail the correct classification rates for the three classes, highlighting that objects belonging to the “Car” class were detected with 100% accuracy, “Cyclists” were detected with 98% accuracy, whilst “Pedestrians” were detected with 89% accuracy. This performance metrics compare very favourably with reported recognition rates of other machine vision application in autonomous driving studies using COCO [Carranza-García et al., 2021].

### 3.2 Object Tracking using the SORT Algorithm

It is necessary to maintain a record of object speeds and angles of movement on a frame-by-frame basis to determine the instantaneous trajectory of identified Cars, Cyclists, and Pedestrians objects. The SORT

algorithm is an object tracking algorithm that uses a combination of a Kalman filter and a Hungarian algorithm to track multiple objects in video streams. Figure 3 provides a high-level overview of the SORT algorithm's operation.

	Car	Pedestrian	Cyclist	Background FP
Car	1	0	0	0.07
Pedestrian	0	0.89	0.02	0.81
Cyclist	0	0	0.98	0.12
Background FP	0	0.11	0	0

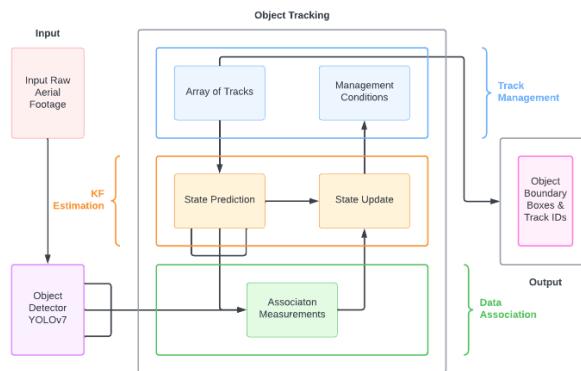
**Table 1: YOLOv7 Model Confusion Matrix**

when tracking multiple objects in a single frame. It boasts an average Intersection over Union (IoU) score of 0.6 and 0.8 and achieves a high Multiple Object Tracking Accuracy (MOTA) score, ranging from 60% to 80%.

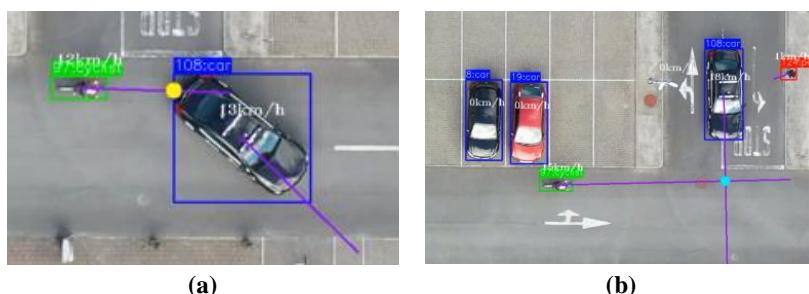
### 3.3 Instantaneous Collision Detection

In the context of an ADEM, a vehicle's control software may wish to utilise the provided local environment mapping data to identify as soon as possible any "obvious" collisions which might be likely to occur. Once a given object is tracked, its instantaneous trajectory may be estimated using basic trigonometric principles.

The algorithm works by detecting objects in each frame of a video stream and assigning them unique IDs. It then uses a Kalman filter to predict the position of the object in the next frame. Subsequently, it associates the predicted with the actual position of the objects on the next frame using the Hungarian algorithm, the Kalman filter helps to smooth out the motion of the objects and handle occlusions, while the Hungarian algorithm helps to maintain the correct ID assignments. The SORT algorithm was chosen due to its reported high accuracy, particularly



**Figure 3 - High-level Operation of the SORT Algorithm**



**Figure 4 – (a): Intersection Between a Trajectory and a Boundary Box – (b): Intersection Between Two Trajectories (R) Flagged as Collisions**

Each object's speed and angle of movement was stored in data structures keyed on a tracking ID. The speed and angle of movement of an object was calculated based on how far an object travelled between frames and in what direction that movement occurred. Once each tracked object had a calculated instantaneous trajectory, any collisions could be immediately identified as an intersection between its instantaneous trajectory and another tracked object's boundary box or an intersection between two or more mapped

instantaneous trajectories, as shown in Figure 4(a) and Figure 4(b) respectively. An OpenCV algorithm was developed, which factored in several edge cases, including multiple collisions that occur on the same boundary box. Clearly, the accuracy of a predicted collision is a directly dependent on the accuracy of both object's trajectories and their boundary boxes. An analysis of the accuracies of these components is detailed in section 4 of this paper.

### 3.4 Location Mapping

A key functionality of the proposed framework is the ability to tag identified objects with accurate “real-world” (i.e. GPS) coordinates. Whilst the aerial video stream may be geotagged (e.g. based on a GPS receiver on the aerial platform), this information may not be sufficient to provide an accurate regular grid structure needed for the bilinear interpolation process which estimated the GPS coordinates of objects in the video frames. In particular, inaccuracies in the estimated altitude of the drone may introduce significant errors or require complex aerial platform control functionality to address. In this framework, a SIFT Brute Force (BF) matcher was used to facilitate feature extraction and detection, as well as key-point matching. The performance of the matcher was enhanced via the use of a suitable Lowe ratio. A location can be accurately geo-tagged by comparing intermittent frames of the video stream to accurately geo-tagged images from a 3<sup>rd</sup> party online source (e.g. Google Earth). In the proposed framework, two images (one from the aerial platform’s video stream and the other from the online source) are determined to be of the same location if there are 50 or more identified key-points between the two images. This threshold was chosen using an iterative approach on a limited target dataset representing Google Earth images of the general geographic location where the drone video stream was captured. Figure 5 illustrates this key point matching process.



**Figure 5 - SIFT Key-point Matching Between Two Images of the Same Location at different times and from different camera positions**

Figure 6(a) illustrates how the GPS coordinate information from the geotagged matched image can be mapped to corners of the frames of the aerial video footage. The GPS coordinates of individual tracked objects within the video stream were then determined using a bilinear interpolation technique with Figure 6(b) illustrating the final output of the implemented framework where all objects are identified, their trajectory estimated and their location accurately geo-tagged.

## 4 Evaluation and Performance

### 4.3 Accuracy of a Tracked Object's Trajectory

The overall accuracy of a tracked object's trajectory may be considered as the probability that the magnitude and angle of the tracked object's trajectory are both correct. This probability may be calculated using (1).

$$P(T) = P(S) \times P(A) \quad (1)$$

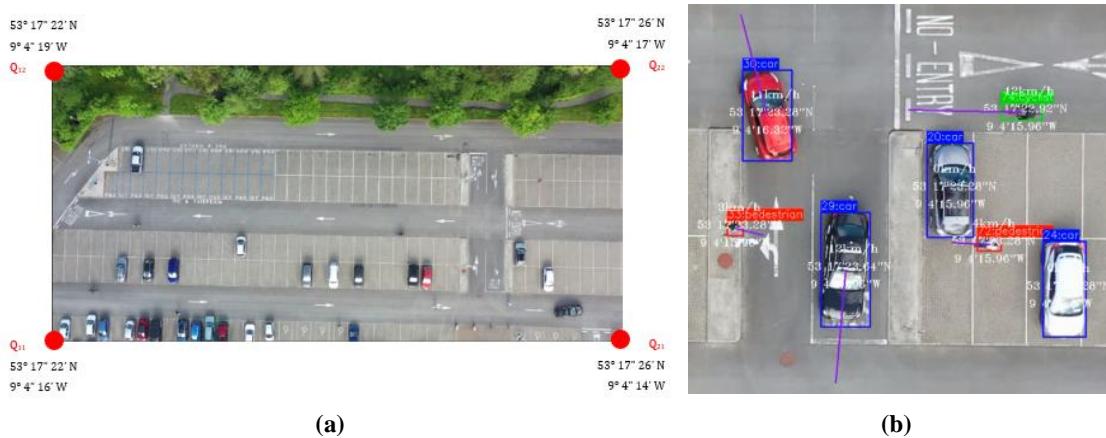
where:

P(T): The probability that the object's overall trajectory is correct

P(S): The probability that the object's speed is correct

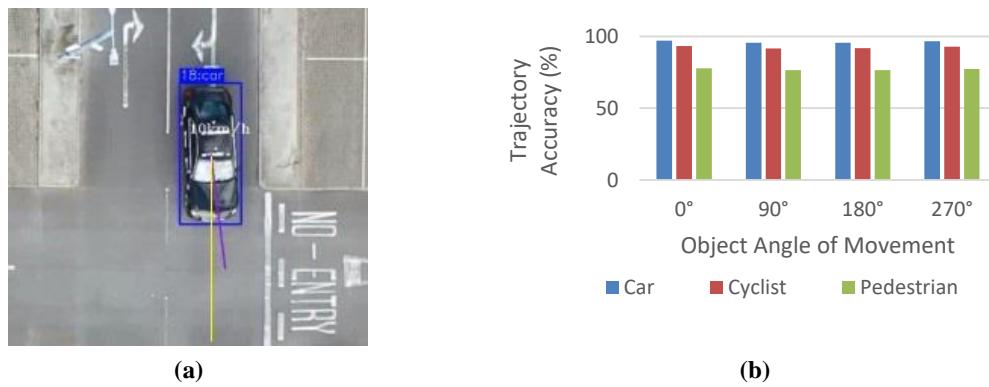
P(A): The probability that the object's trajectory angle is correct

The length of a given trajectory is determined by the speed at which the object is moving. The probability that the object's speed is correct may be derived from the confusion matrix shown in Table 1. By assuming the probability of a given object being detected in a frame is an independent event, the likelihood that the object's speed is correct is the square of the probability of the object being detected. The probability that the trajectory's angle is correct was estimated by comparing the calculated angle of movement to the ground truth angle of movement (see Figure 7(a) for example) with the disparity between the estimated and "ground truth" angles yield a margin of error. Repeating this process over several instances of object motion in the video stream (for different object classes) allowed a mean margin of error to be estimated.



**Figure 6 – (a): Corner Coordinates Mapped from Matched Geotagged Images – (b): Output ADEM**

Figure 7 (b) illustrates the overall accuracy of each class for periods where the objects were moving in one of four different directions (as measured relative to the vertical axis of the video frame).



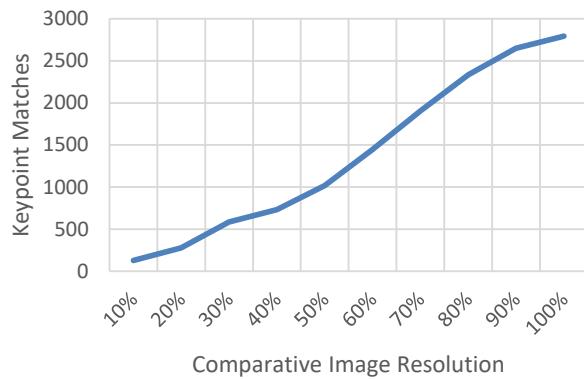
**Figure 7 – (a): Comparison of the Expected Angle of Movement (Yellow) to the Calculated Angle of Movement (Purple) – (b): Accuracy of Trajectories Angle for Each Class**

It shows that both the Car and Cyclist classes have extremely accurate trajectories. Pedestrians have slightly less accurate trajectories, primarily due to their lower detection probability. The results produced from this analysis suggests that the application has a valid claim to accuracy. Cars moving vertically or horizontally boast an average instantaneous trajectory accuracy of over 96%, with the accuracies of cyclists and pedestrians being approximately 92% and 78% respectively.

#### 4.4 SIFT Algorithm Robustness

As outlined in section 3.4 of this paper, accurate GPS coordinate information relating to objects in the video stream may be estimated by first comparing frames from the video stream to 3<sup>rd</sup> party geo-tagged satellite\ aerial imagery e.g. sourced from Google Earth. However, it is important to consider the impact that the satellite\ aerial image quality may have on the algorithm performance.

The accuracy of the SIFT matcher is a function of image resolution and, hence, it was important that the SIFT matcher performs well even when operating with quite low-resolution images from a 3<sup>rd</sup> party source. An analysis was undertaken to ensure the accuracy of the key-point matching algorithm across a range of satellite image resolutions (from Google Earth). Starting with a 4K image, the number of key points detected by the SIFT algorithm was recorded. By reducing the resolution of the image, the number of matches could be compared over a range of image quality levels. From Figure 8, it is evident that higher resolution images provided better accuracy, producing more key-point matches. In fact, a near linear relationship between accuracy and resolution was found to exist. Even at 10% of the original image resolution, the SIFT algorithm still produces more than 100 key-point matches (bearing in mind the previously stated condition that, in this framework, two images are considered to be of the same location if they have 50 or more common key-points). Therefore, this particular evaluation provides a valuable result, as it illustrates the robustness of the SIFT algorithm (and therefore the location mapping functionality), even when utilising lower quality images for 3<sup>rd</sup> parties.



**Figure 8: Robustness of the SIFT Algorithm over a Range of Image Qualities**

### 5 Conclusions and Future Work

This paper has proposed a framework to derive dynamic environment mapping information to support autonomous and assisted driving application streamed from an aerial platform. The framework utilises a highly accurate YOLOv7 object detection model, which is trained and validated using a custom dataset comprising annotated images of cars, cyclists, and pedestrians. The accuracy of the model was extensively quantified and analysed, with the model recording an mAP@0.95 score of 95.6%. An OpenCV-based application detects and tracks objects in the provided aerial footage, allowing for the calculation of a trajectory for each tracked object. Subsequently, any potential collisions based on the intersection of these trajectories could be detected. Analysis regarding the accuracy of these trajectories was also undertaken, with the calculated angle of movement determined to have less than a 5% margin of error with the ground truth. A system to map tracked objects to their real-world coordinates was also implemented using SIFT techniques coupled with Bilinear Interpolation. The coordinate mapping system was analysed for its performance when passed lower resolution satellite images, such as those which were retrieved from Google Earth. There is significant potential for future further development of this framework. Expanding the object detection model to support a wider range of objects would be required for deployment in diverse environments. There is also a need to analyse the performance of the model in non-ideal lighting and adverse weather conditions. Fully integrating the location mapping functionality with a wider aerial image source, such as Google Earth, and the deployment of a complete system possibly using an architecture similar to that suggested in Figure 1 would provide an excellent testbed for further investigations of the proposed framework. Further work could also investigate whether an edge-processing based architecture (with most

of the processing implemented either on the UAV or a road-side hub connected to the UAV using lower power radio links) may offer a better overall solution in terms of power consumption and UAV flight time. Lastly, a more in-depth review of data privacy considerations relating to the proposed approach should be completed though the processing of “top down” video streams arguably raise significantly less data privacy concerns compared to those associated with roadside and vehicle camera systems (due to the fact that it is more difficult to robustly identify and track specific individuals in a video stream from the UAV platform in our architecture).

## References

- [Wang et al., 2022] Wang, C.-Y., Bochkovskiy, A., and M. Liao, H.-Y. (2022). *YOLOv7: Trainable bag of freebies sets new state-of-the-art for real time object detectors*. ArXiv. <https://doi.org/10.48550/arXiv.2207.02696>
- [Koay et al., 2021] Koay, H.V., Chuah, J.H., Chow, C.-O., Chang, Y.-L., and Yong, K.K. (2021). *YOLO-RTUAV: Towards Real-Time Vehicle Detection through Aerial Images with Low-Cost Edge Devices*. Remote Sens. <https://doi.org/10.3390/rs13214196>
- [Luo et al., 2022] Luo, X., Wu, Y., and Wang, F. (2022). *Target Detection of UAV Aerial Imagery Based on Improved YOLOv5*. Remote Sens. <https://doi.org/10.3390/rs14195063>
- [Fornace, et al., 2014] Fornace, K., Drakeley, C., William, T., Espino, F., and Cox, J. (2014)., *Mapping infectious disease landscapes: Unmanned aerial vehicles and epidemiology*. Trends in Parasitology. <https://doi.org/10.1016/j.pt.2014.09.001>
- [Vora et al., 2020] Vora, A., Agarwal, S., Pandey, G., and McBride., J. (2020) *Aerial Imagery based LIDAR Localization for Autonomous Vehicles*. ArXiv. <https://doi.org/10.48550/arXiv.2003.11192>
- [Lowe, 2004] Lowe, D.G. (2004) *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer-Vision. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- [Yicong et al., 2017] Yicong, T., Chen, C., and Shah, M. (2017). *Cross-view Image Matching for Geolocation in Urban Environments*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). <https://doi.org/10.1109/CVPR.2017.216>
- [Carranza-García et al., 2021] Carranza-García, M., Torres-Mateo, J., Lara-Benítez, P., and García-Gutiérrez, J. (2021) *On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles using Camera Data*. Remote Sens. <https://doi.org/10.3390/rs13010089>

# Empowering Visually Impaired Individuals: A Novel Use of Apple Live Photos and Android Motion Photos

Seyedalireza Khoshirat      Chandra Kambhamettu

*Video/Image Modeling and Synthesis (VIMS) Lab, University of Delaware*

## Abstract

Numerous applications have been developed to assist visually impaired individuals that employ a machine learning unit to process visual input. However, a critical challenge with these applications is the sub-optimal quality of images captured by the users. Given the complexity of operating a camera for visually impaired individuals, we advocate for the use of Apple Live Photos and Android Motion Photos technologies. In this study, we introduce a straightforward methodology to evaluate and contrast the efficacy of Live/Motion Photos against traditional image-based approaches. Our findings reveal that both Live Photos and Motion Photos outperform single-frame images in common visual assisting tasks, specifically in object classification and VideoQA. We validate our results through extensive experiments on the ORBIT dataset, which consists of videos collected by visually impaired individuals. Furthermore, we conduct a series of ablation studies to delve deeper into the impact of deblurring and longer temporal crops.

**Keywords:** Live Photo, Motion Photo, Deep Learning, Visually Impaired

## 1 Introduction

*Live Photos* and *Motion Photos*, technologies from Apple and Android, allow a single photo to function as a still image and when activated, a short video with motion and sound. These technologies leverage a background feature that continuously captures images when the Camera app is opened, regardless of whether the shutter button is pressed. When a Live/Motion Photo is taken, the device records this continuous stream of photos, capturing moments before and after the shutter press. These images are stitched into a three-second animation, complemented by optional audio recorded during the same span. Live/Motion Photos surpass video clips due to their ease of capture and standardized format. Figure 1 depicts the main three components of a Live/Motion Photo, and Figure 5 shows screenshots of the Apple iOS environment for capturing and working with Live Photos.

People with visual impairments often rely on assistive devices that provide insights about their surroundings. For instance, people with low vision often rely on magnification tools to better observe the content of interest, or those with low vision and no vision rely on on-demand technologies [BeMyEyes, 2023, BeSpecular, 2023, Khoshirat and Kambhamettu, 2023] that deliver answers to submitted visual questions. Two fundamental computer vision tasks in these aids are object classification and video question answering (VideoQA). Object classification, though basic, is a key component of more advanced methods [Khoshirat and Kambhamettu, 2022]. In contrast, VideoQA accurately responds to inquiries about any video, empowering visually impaired people to access information about real-world or online videos [Hosseini et al., 2022].

A significant problem with the current visual assisting technologies is the limitation of the visually impaired people to capture the desired image for these technologies. The images taken by blind people have different quality flaws, such as blurriness, brightness, darkness, obstruction, and so on [Maserat et al., 2017]. Image quality issues may make it difficult for humans and machine learning systems to recognize image content, causing the system to provide set responses, such as “unanswerable”. Prior research has indicated that this can

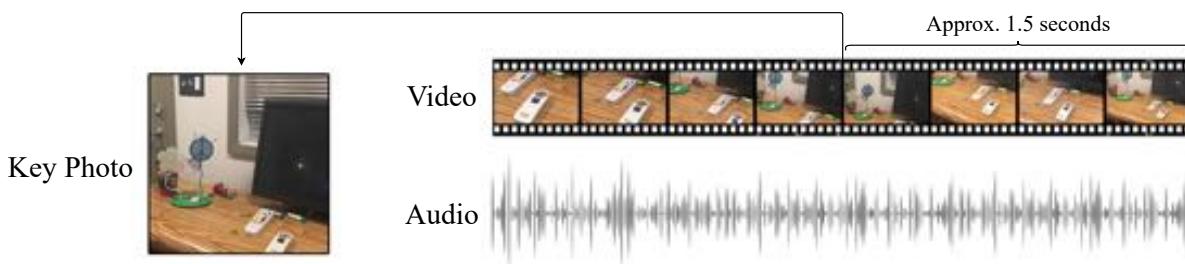


Figure 1: Apple Live Photo structure. A Live/Motion Photo consists of a key photo, a three-second-long video, and the optional corresponding audio. The key photo is the middle frame of the video, but it can be changed to another frame.

be frustrating for people with visual impairments using accessible applications, requiring extra time and effort to determine what is going wrong and get an answer [Bhattacharya et al., 2019]. Figure 2 shows a recorded video by a visually impaired user where half of the frames cover only a small portion of the object.

We posit that the additional contextual information provided by Live/Motion Photos can significantly enhance the ability of the assistance systems to accurately interpret and analyze the content of the images. Not only does this approach provide multiple frames for analysis, which could increase the chances of capturing a clear shot of the subject, but it also offers temporal information that can be critical for understanding dynamic scenarios. Through the course of this paper, we will present empirical evidence demonstrating how the use of Live/Motion Photos can mitigate the challenges faced by visually impaired individuals in capturing clear images.

Our contributions are as follows:

- We introduce a straightforward approach for comparing Live/Motion Photos to images.
- We evaluate state-of-the-art methods on Live/Motion Photos and images for object classification and VideoQA tasks.
- We conduct ablation studies on the impact of deblurring and varying temporal crop lengths.

## 2 Related Work

A plethora of commercial systems have been developed to empower individuals with visual impairments. These commercial systems are categorized into two distinct types: human-in-the-loop systems and end-to-end (E2E) automated systems. Human-in-the-loop systems are designed to bridge the gap between visually impaired individuals and sighted volunteers or staff members. Through these systems, users can make inquiries or seek assistance with visual tasks. Some notable examples of human-in-the-loop platforms are BeMyEyes, BeSpecular, and Aira [BeMyEyes, 2023, BeSpecular, 2023]. Contrary to human-in-the-loop systems, end-to-end systems rely on artificial intelligence and cloud computing to provide visual assistance to users. These systems do not involve human intermediaries. Examples of E2E systems include TapTapSee and Microsoft's Seeing AI.

A critical factor that determines the efficacy of these systems is the clarity and relevance of the content within the images that are sent for analysis. Given that visually impaired individuals might face challenges in capturing well-composed images, ensuring that the subject matter of the image is clear and discernible is not a trivial task. In this paper, we introduce an innovative approach to alleviate this challenge by utilizing Live Photos or Motion Photos.



Figure 2: A visually impaired user trying to record a video of a keyboard [Massiceti et al., 2021]. Adjusting the camera field of view to cover a whole object is a challenging task for blind users. The frames are uniformly sampled, and the total video length is five seconds.

### 3 Method

Studying Live/Motion Photos poses a significant challenge due to the absence of existing datasets. The process of creating a comprehensive dataset solely from visually impaired users is laborious and complex [Massiceti et al., 2021]. To address this issue, we leverage pre-existing video datasets collected by visually impaired individuals or those containing content relevant to the daily experiences of the blind. By extracting three-second temporal crops from these videos, we simulate Live/Motion Photos for tasks such as object classification and VideoQA. This enables us to evaluate and compare the effectiveness of different methods on both simulated Live/Motion Photos and standard images.

#### 3.1 Object Classification

To demonstrate the impact of Live/Motion Photos on object classification accuracy, we conduct experiments using the ORBIT dataset [Massiceti et al., 2021]. This dataset is a collection of videos recorded on cell phones by people who are blind or low-vision. The ORBIT dataset consists of 3,822 videos with 486 object categories recorded by 77 blind or low-vision people on their cell phones. Each video is meant to capture one main object, although the object may not be visible in all frames. The videos are captured in various lengths, from one second to two minutes.

To simulate Live/Motion Photos, we create short video clips with the same length as Live/Motion Photos from ORBIT and compare the performance of different image classifiers to video classifiers on these clips. To this aim, we train each image classifier on image frames of the videos and report the average classification accuracy of the frames. To evaluate the video classifiers, we train and test each method on random temporal crops of three seconds. We choose the top-performing image and video classifiers; specifically, ResNet [He et al., 2016], MViTv2 [Li et al., 2021], and EfficientNetV2 [Tan and Le, 2021] for image classification, and ViViT [Arnab et al., 2021] and MViTv2 [Li et al., 2021] for video classification. We use the same hyperparameters and setup as in the original implementations, and the input size is fixed across all the methods. Following [Massiceti et al., 2021], we use frame accuracy as the evaluation metric for the frame-by-frame classification and video accuracy for the holistic video classification. Frame accuracy is the average number of correct predictions per frame divided by the total number of frames in a video. Video accuracy is the number of correct video-level predictions divided by the total number of videos.

Table 1 reports the object classification accuracy. The highest accuracy using images is 70.9% and achieved by EfficientNetV2-L. The results show that video classification approaches outperform frame-by-frame classification. More specifically, for Live/Motion Photos (videos of three seconds long), MViTv2 achieves an accuracy of 77.1% which is an improvement of 6.2% over EfficientNetV2-L. Since MViTv2 is designed for both image and video classification, it exhibits the benefit of using video clips over images better than other methods. Similarly, ViViT reaches an accuracy of 74.9% which is higher than EfficientNetV2-L by a margin of 4.0%. This

Method	Accuracy
ResNet-152 [He et al., 2016]	69.2
MViTv2-B [Li et al., 2021]	70.7
EfficientNetV2-L [Tan and Le, 2021]	70.9
ViViT [Arnab et al., 2021]	74.9
MViTv2-B [Li et al., 2021]	77.1

Table 1: Comparison of frame-by-frame to holistic object classification methods on the ORBIT test set. The top three methods use images, and the bottom two use Live/Motion Photos.

Method	Accuracy
mPLUG [Li et al., 2022]	28.9
BEiT-3 [Wang et al., 2022]	30.1
Just Ask [Yang et al., 2021]	34.9
Singularity [Lei et al., 2022]	38.6

Table 2: Results of image-based and video-based methods for the VideoQA task on the ActivityNet-QA test set. mPLUG and BEiT-3 use images, while Just Ask and Singularity use Live/Motion Photos.

result strongly supports the effectiveness of Live/Motion Photos over single images.

### 3.2 Video Question Answering

We investigate the effectiveness of Live/Motion Photos in the VideoQA task. We compare the performance of multiple VQA methods on image frames to the performance of VideoQA methods on video clips with the same length as Live/Motion Photos. While there are numerous video question answering datasets, we choose the ActivityNet-QA dataset [Yu et al., 2019] since it contains video clips similar to the day-to-day life of people with visual impairments. The ActivityNet-QA dataset adds question-answer pairs to a subset videos of the ActivityNet dataset [Caba Heilbron et al., 2015]. The ActivityNet-QA dataset contains 5,800 videos with 58,000 human-annotated question-answer pairs divided as 3,200/1,800/800 videos for train/val/test splits. This dataset contains 200 different types of daily human activities, which is suitable for visual assisting applications.

We train image-based methods on randomly drawn frames with their corresponding question-answer pairs from the ActivityNet-QA dataset. Similarly, we train video-based methods on random temporal crops with the same length as Live/Motion Photos. We employ mPLUG [Li et al., 2022] and BEiT-3 [Wang et al., 2022] as the image-based methods and Just Ask [Yang et al., 2021] and Singularity [Lei et al., 2022] as the video-based methods for Live/Motion Photos. These methods achieve state-of-the-art accuracy in the VQA and VideoQA tasks, and their implementation code is publicly available. For each method, we re-use the original hyperparameters that achieve the best results.

As for the evaluation criteria, we use accuracy, a commonly used criterion to measure the performance of classification tasks. For the QA pairs in the test set with size  $N$ , given any testing question  $\mathbf{q}_i \in Q$  and its corresponding ground-truth answer  $\mathbf{y}_i \in Y$ , we denote the predicted answer from the model by  $\mathbf{a}_i$ .  $\mathbf{a}_i$  and  $\mathbf{y}_i$  correspond to a sentence that can be seen as a set of words. The accuracy measure is defined as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\mathbf{a}_i = \mathbf{y}_i] \quad (1)$$

where  $\mathbf{1}[\cdot]$  is an indicator function such that its output is one only if  $\mathbf{a}_i$  and  $\mathbf{y}_i$  are identical, and zero otherwise [Yu et al., 2019]. We follow previous evaluation protocols for open-ended settings [Yang et al., 2021, Yu et al., 2019, Lei et al., 2022] and use a fixed vocabulary of training answers.

Table 2 reveals the results of our experiments for the VideoQA task. The highest accuracy for image-based approaches is 30.1% and achieved by BEiT-3. Both VideoQA methods outperform the VQA methods. More specifically, using Live/Motion Photos, Singularity achieves the highest accuracy of 38.6%, which is more than 8% higher than BEiT-3 accuracy. Similarly, Just Ask reaches an accuracy of 34.9% which is 4.8% higher than BEiT-3.

The outcomes of our experiments in object classification and VideoQA confirm the benefit of using Live/Motion Photos over images.

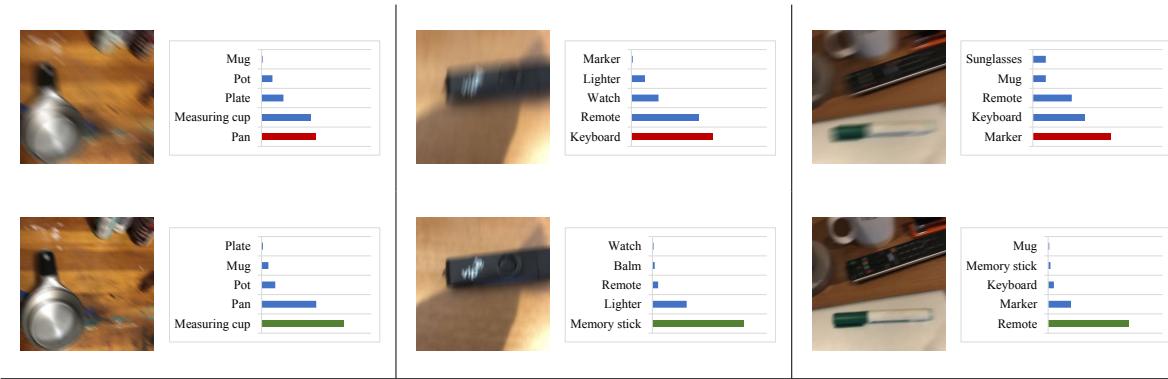


Figure 3: Sample video frames from ORBIT dataset with their corresponding model output. **Top:** Original frame. **Bottom:** After deblurring. Deblurring enhances the precision of model predictions.

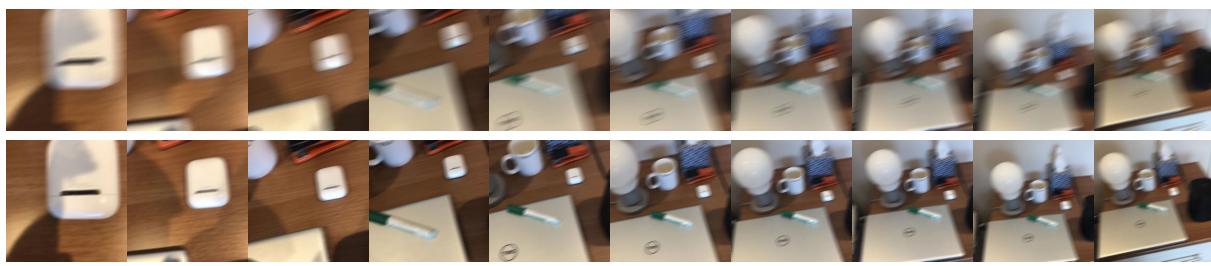


Figure 4: Ten uniformly sampled frames from a random video in ORBIT, before and after deblurring. **Top:** Original video. **Bottom:** After deblurring. Deblurring tends to provide greater benefits to frames containing smaller objects.

## 4 Deblurring Impact

Blurring is a prevalent issue in images and videos captured by individuals with visual impairments [Chiu et al., 2020], and this issue can adversely affect the efficacy of assistive technologies. In this section, we undertake a systematic investigation to discern the potential benefits of deblurring on the accuracy of object classification and VideoQA. For the deblurring process, we employ the FGST method [Lin et al., 2022], a state-of-the-art video deblurring algorithm that amalgamates Transformer modules with optical flow estimation. We then proceed to apply FGST on two datasets, namely ORBIT and ActivityNet-QA, to deblur the visual content. With the deblurred datasets, we replicate the experiments as outlined in Section 3.1 and Section 3.2.

The outcomes of this investigation are tabulated in Table 3. The table segregates the results into two categories - the upper portion presents the outcomes for object classification, while the lower portion provides the results for VideoQA. The empirical findings demonstrate that the maximum enhancement in accuracy is 2.6%, which is attained by the Just Ask method, whereas the minimum improvement is documented at 1.7% by the MViTv2-B method. Furthermore, for a more illustrative understanding, Figure 3 showcases a selection of frames along with the corresponding model outputs prior to and subsequent to the deblurring process. This visualization facilitates a comparison of the quality and detail in the frames. Additionally, Figure 4 presents a compilation of frames extracted from a deblurred video, providing a visual representation of the enhancements achieved through the deblurring process.

Method	Without Deblurring	With Deblurring
ViViT [Arnab et al., 2021]	74.9	76.9 (+2.0)
MViTv2-B [Li et al., 2021]	77.1	78.8 (+1.7)
Just Ask [Yang et al., 2021]	34.9	37.5 (+2.6)
Singularity [Lei et al., 2022]	38.6	40.9 (+2.3)

Table 3: The impact of deblurring Live/Motion Photos. **Top:** Object classification on the ORBIT test set. **Bottom:** VideoQA on the ActivityNet-QA test set.

Method	Live/Motion Photo =3s	Accuracy				All Frames
		Short <15s	Medium 15s> and <30s	Long >30s		
ViViT [Arnab et al., 2021]	74.9	75.8	76.6	77.1	76.7	
MViTv2-B [Li et al., 2021]	77.1	77.9	78.4	79.0	78.5	
Just Ask [Yang et al., 2021]	34.9	36.0	36.9	37.8	37.0	
Singularity [Lei et al., 2022]	38.6	39.6	40.4	41.1	40.6	

Table 4: The results of top-performing methods with different temporal crop lengths. **Top:** Object classification on the ORBIT test set. **Bottom:** VideoQA on the ActivityNet-QA test set. Videos shorter than a targeted crop size are not included in that group.

## 5 Temporal Length Impact

Although Live/Motion Photos are limited to three seconds, it is possible for other applications to implement the same technology but without the capturing limitations. Therefore, in this section, we study the effect of video length on accuracy for object classification and VideoQA tasks. To this aim, we evaluate the video-based methods on three temporal crop size ranges. The ‘Short’ crop range is the random crops of shorter than 15 seconds, the ‘Medium’ range is between 15 to 30, and the ‘Long’ range is longer than 30 seconds. Videos that are shorter than a targeted crop size are not included in that group. Additionally, we evaluate the methods on the whole dataset using all the available frames in the videos. We do not use videos shorter than the required minimum length for the Medium and Long ranges. We use the same setup in Sections 3.1 and 3.2.

For object classification, we employ ViViT [Arnab et al., 2021] and MViTv2-B [Li et al., 2021] and evaluate them on the ORBIT dataset [Massiceti et al., 2021]. The top two methods in Table 4 report the results for object classification using different video lengths. For MViTv2-B, the lowest accuracy is 77.1%, achieved by using Live/Motion Photos, and the highest accuracy is 79.0%, achieved using the longest video crops. For both methods, adding more frames helps improve the accuracy. The accuracy of using all the frames gets slightly worse due to the addition of shorter videos. Since having more frames reveals more data about an object, the longer crops reach higher accuracies.

For VideoQA, we employ Just Ask [Yang et al., 2021] and Singularity [Lei et al., 2022] and train and test them on the ActivityNet-QA dataset [Yu et al., 2019]. The bottom two methods in Table 4 report the results for different video lengths in VideoQA. The lowest accuracy for Singularity is 38.6% by using Live/Motion Photos, and 41.1% is the highest accuracy by using all the frames.

The findings from our ablation study reveal that while there is a positive correlation between the length of video clips and the enhancement in accuracy, the incremental accuracy attained through longer video clips, as compared to Live/Motion Photos, is not significant in contrast to single images. This implies that Live/Motion Photos, constrained to a duration of three seconds, are capable of furnishing a substantial improvement in accuracy that is deemed sufficient for a majority of applications.



Figure 5: Three screenshots showcasing Live Photos functionality on Apple iOS [Apple, 2021].

## 6 Conclusion and Future Directions

Despite significant recent developments, visual assistance applications are still in need of improvement. Current machine learning methods designed to help visually impaired people suffer from the low quality of the images taken by the end users.

In this paper, we made multiple contributions to improving existing methods for visual assisting. We introduced a simple way to evaluate the performance of Live/Motion Photos compared to single images. We employed this approach to show that Live/Motion Photos achieve higher accuracy in common visual assisting tasks. Our experiment revealed that Live/Motion Photos perform better than images in object classification and VideoQA tasks. In addition, we further studied the effect of longer temporal crops and showed how deblurring can improve accuracy.

In future research, it is essential to carry out user studies involving visually impaired individuals. This information will guide us in refining our method, ensuring it is not only technically robust but also practically beneficial for the intended users.

## References

- [Apple, 2021] Apple (2021). Take and edit live photos.
- [Arnab et al., 2021] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.
- [BeMyEyes, 2023] BeMyEyes (2023). Be my eyes.
- [BeSpecular, 2023] BeSpecular (2023). Bespecular.
- [Bhattacharya et al., 2019] Bhattacharya, N., Li, Q., and Gurari, D. (2019). Why does a visual question have different answers? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4271–4280.
- [Caba Heilbron et al., 2015] Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

- [Chiu et al., 2020] Chiu, T.-Y., Zhao, Y., and Gurari, D. (2020). Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3656.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [Hosseini et al., 2022] Hosseini, P., Khoshirsat, S., Jalayer, M., Das, S., and Zhou, H. (2022). Application of text mining techniques to identify actual wrong-way driving (wwd) crashes in police reports. *International Journal of Transportation Science and Technology*.
- [Khoshirsat and Kambhamettu, 2022] Khoshirsat, S. and Kambhamettu, C. (2022). Semantic segmentation using neural ordinary differential equations. *Advances in Visual Computing: 17th International Symposium, ISVC 2022, San Diego, CA, USA, October 3–5, 2022, Proceedings, Part I*, pages 284–295.
- [Khoshirsat and Kambhamettu, 2023] Khoshirsat, S. and Kambhamettu, C. (2023). Embedding attention blocks for the vizwiz answer grounding challenge. *VizWiz Grand Challenge Workshop*.
- [Lei et al., 2022] Lei, J., Berg, T. L., and Bansal, M. (2022). Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*.
- [Li et al., 2022] Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al. (2022). mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.
- [Li et al., 2021] Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., and Feichtenhofer, C. (2021). Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*.
- [Lin et al., 2022] Lin, J., Cai, Y., Hu, X., Wang, H., Yan, Y., Zou, X., Ding, H., Zhang, Y., Timofte, R., and Van Gool, L. (2022). Flow-guided sparse transformer for video deblurring. *arXiv preprint arXiv:2201.01893*.
- [Maserat et al., 2017] Maserat, E., Safdari, R., Aghdaei, H. A., Khoshirsat, A., and Zali, M. R. (2017). 43: Designing evidence based risk assessment system for cancer screening as an applicable approach for the estimating of treatment roadmap. *BMJ Open*, 7(Suppl 1):bmjopen–2016.
- [Massiceti et al., 2021] Massiceti, D., Zintgraf, L., Bronskill, J., Theodorou, L., Harris, M. T., Cutrell, E., Morrison, C., Hofmann, K., and Stumpf, S. (2021). Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10818–10828.
- [Tan and Le, 2021] Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR.
- [Wang et al., 2022] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. (2022). Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.
- [Yang et al., 2021] Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. (2021). Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697.
- [Yu et al., 2019] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019). Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

# Saliency Maps as an Explainable AI Method in Medical Imaging: A Case Study on Brain Tumor Classification

Ayse Keles<sup>1</sup>, Ozan Akcay<sup>2</sup>, Halil Kul<sup>3</sup>, and Malika Bendechache<sup>4</sup>

<sup>1</sup>*General Directorate of Health Information Systems, Ministry of Health, Türkiye*

<sup>2</sup>*Department of Biomechanics, Dokuz Eylül University, Türkiye*

<sup>3</sup>*School of Medicine, Department of Neurosurgery, Ankara Yıldırım Beyazıt University, Türkiye*

<sup>4</sup>*ADAPT Research Centre, School of Computer Science, University of Galway, Ireland*

## Abstract

Explainable Artificial Intelligence (XAI) plays a crucial role in the field of medical imaging, where AI systems are used for clinical decision support and diagnostic processes. XAI aims to develop approaches that make machine learning (ML) models more transparent and interpretable, facilitating human-AI collaboration and improving trust. In medical imaging, early prediction of anomalies is vital, and understanding AI's decision-making process is crucial. Saliency maps are used to highlight important regions in an image and have been found a user-friendly explanation method for deep learning-based imaging tasks. They are widely used in many applications across various domains. There are different methods for generating saliency maps depending on the analysis and temporal occurrence. Ad-hoc methods are model-specific, while ante-hoc and post-hoc methods are independent of the model architecture. Post-hoc methods, such as activation-based, perturbation-based, and gradient-based methods, are commonly used for generating saliency maps. In this case study, we focus on the application of gradient-based saliency maps using Magnetic Resonance Imaging (MRI) images to provide insights into brain tumor classification. To achieve this, we implemented a convolutional neural network (CNN) model on a benchmark brain MRI dataset and generated saliency maps. The results reveal that the tumor and its surrounding pixels play a significant role in the classification of brain MRIs, highlighting the importance of tumor shape in the classification process. Understanding these underlying mechanisms enhances the robustness, reliability, and accountability of AI systems used in brain tumor detection and classification.

**Keywords:** Explainable Artificial Intelligence, XAI, Saliency Map, Deep Learning, Clinical Decision Support, Brain Tumor Classification.

## 1. Introduction

### 1.1. Explainable Artificial Intelligence (XAI) in Medical Imaging

Artificial Intelligence (AI) has inevitably become increasingly involved in clinical practice, in the form of clinical decision support systems (CDSS), assisting clinicians in the detection and diagnostic processes. It uses complex and high-performance machine learning (ML) models trained on large sets of medical data. Since state-of-the-art ML systems are getting more sophisticated, they are also becoming less interpretable, hence black-box solutions. Explainability of the AI model is essential in the medical domain, where understanding the reason for a decision is as important as the decision itself (Amann, Blasimme, Vayena, Frey, & Madai, 2020; Sina, Zarei, & Ragan, 2021). At the point of comprehensibility of AI/ML systems, the paradigm of eXplainable AI (XAI) comes into the scene. To understand, and eventually improve the inner dynamics of a model, XAI aims at developing approaches for making ML models more transparent and interpretable, so users can understand how the model is making decisions. Moreover, explainability provides the ability to overlook the model's possible bottlenecks and identify potential biases, yielding to make necessary curations and calibrations on it, increasing trust, and facilitating human-AI communication and collaboration (Rasheed, et al., 2022). In the field of medical imaging, the purpose of AI is to support clinicians to make an informed decision in the diagnostic procedures as early as possible since it is vital to predict initial stages in case of a malignancy. Proper classification of clinical findings needs to be

evidence-based and trustworthy. Medical experts in studies tend to review explanations provided by AI when its predictions align with their hypotheses or help resolve disagreements, especially in medical imaging tasks (Jin, Li, & Hamarneh, 2021). Saliency maps are one of the preferred user-friendly explanation methods.

## 1.2. Saliency Maps in Computer Vision and Medical Imaging

In computer vision (CV), saliency maps are seen as a visualization technique to highlight the most key features of an input that contribute to the output of an AI model (Lamichhane, Carli, & Battisti, 2023). Therefore, they provide a way to understand which parts of the input data are influencing the model's decision-making process. This makes them widely used in Deep Learning (DL)-based image processing tasks, such as object detection, image segmentation, and visual attention modelling (Nauta, et al., 2022). Saliency maps have proven to be valuable in various applications and domains, not only in computer vision, but also in natural language processing and generation, recommendation systems, fraud, and anomaly detection, and human-computer interaction (Rahimi & Jain, 2022; RichardWebster, Hu, Fieldhouse, & Kitware, 2022; Andrushia, Sagayam, Dang, Pomplun, & Quach, 2021), etc. Several methods have been proposed to generate saliency maps in XAI. It is important to design methods that are specifically suited for the medical field due to its unique characteristics (Reyes, et al., 2020). Additionally, these methods vary depending on factors such as the type of input data (e.g., acquisition modality, additional clinical data) and the architecture of the AI model (Singh, Sengupta, & Lakshminarayanan, 2020). Saliency maps are also classified in terms of temporal occurrence. Ad-hoc, ante-hoc, and post-hoc saliency maps are three distinct approaches for analysing and interpreting saliency in various domains. Each approach serves different purposes in saliency analysis and interpretation. Ante-hoc saliency maps, also known as pre-hoc saliency maps, are created in advance before any specific task or analysis aiming to capture and represent the inherent salient features or regions in the input data, irrespective of any specific task or application. Post-hoc saliency maps are generated after the completion of a specific task or analysis. Post-hoc saliency maps help in understanding the underlying factors that influenced the results and provide insights into the decision-making process or the causes of certain outcomes (Ali, et al., 2023). Three most common approaches to generate *Post-Hoc saliency maps*; *activation-based methods* that work by aggregating neuron activation signals in a deep neural network and visualizing signals in the input space, *perturbation-based methods* that operate by manipulating input features and observing their respective output to measure their difference from the original output, and finally *gradient-based methods*, that function by calculating the partial derivative of a prediction for an input feature. *Gradient-based methods* compute the importance of input features by examining the gradients of the output class for the input data. These methods leverage the gradient information to identify which input features contribute most significantly to the final prediction (Vries, et al., 2023).

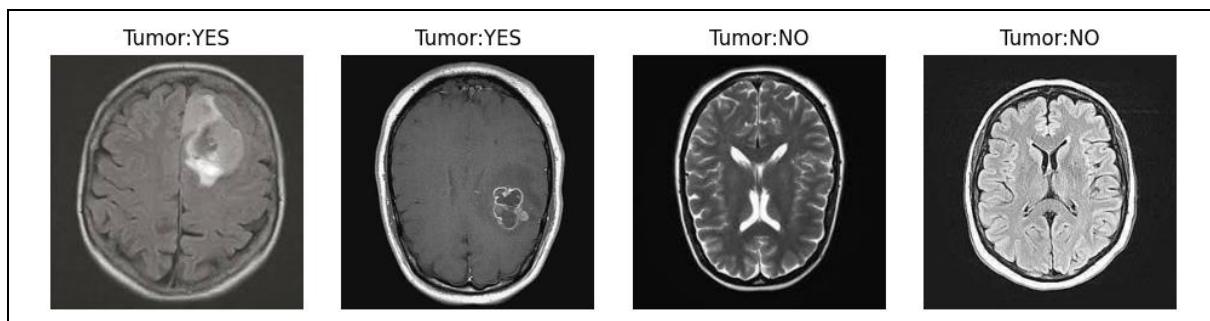
## 2. A Case-Study

Some research studies in the field of medical image classification have undertaken analyses to find out the significance of individual input image pixels in influencing the final prediction using different saliency methods. In their study, Ayhan et al. (Ayhan, et al., 2022) validated perturbation-based saliency maps compared with expert annotations for the detection of diabetic retinopathy and neovascular age-related macular degeneration using retinal fundus images and optical coherence tomography scans. In another study (Wang, 2021), Wang examined gradient-based saliency maps of an AI model and provided valuable insights into the regions of input X-ray images that hold significance in the model's estimation of hand bone age. Amorim et al (Amorim, Abreu, Santos, Cortes, & Vila, 2023) assessed the fidelity of the saliency maps by introducing inherent perturbations in histopathological images of breast cancer, subsequently discovering that saliency maps based on gradients exhibit a correlation with the presence of tumor evidence within the image. Brain tumor detection is one of the hottest topics in medical image research (Muhammad, Khan, Ser, & Albuquerque, 2021; Ranjbarzadeh, Caputo, Tirkolaee, Ghoushchi, & Bendechache, 2022). Understanding how AI models classify brain MRI images is crucial for reliable and accountable brain tumor detection. In their study, Wojciech et al. (Chmiel, Kwiecie, & Motyka,

2023) demonstrate that integrating saliency regions with various CNNs is effective for brain tumor detection in X-ray images. However, MRI imaging is extensively employed in the diagnosis of brain tumors and various medical conditions, surpassing X-ray imaging in terms of superior soft tissue visualization and non-invasiveness. This research study focuses on the interpretability of the CNN model in tumor classification by generating post-hoc, gradient-based saliency maps on brain MRI images as a novelty. The BR35H::Brain Tumor Detection 2020 dataset (Hamada, 2020) of brain MRIs dataset is used in the case study. In brain MRIs, it is seen that the tumor and its surrounding pixels are of high importance in the classification of images with tumors. These results highlight that the tumor and its shape in brain MRIs play a significant role in classification.

## 2.1. Dataset

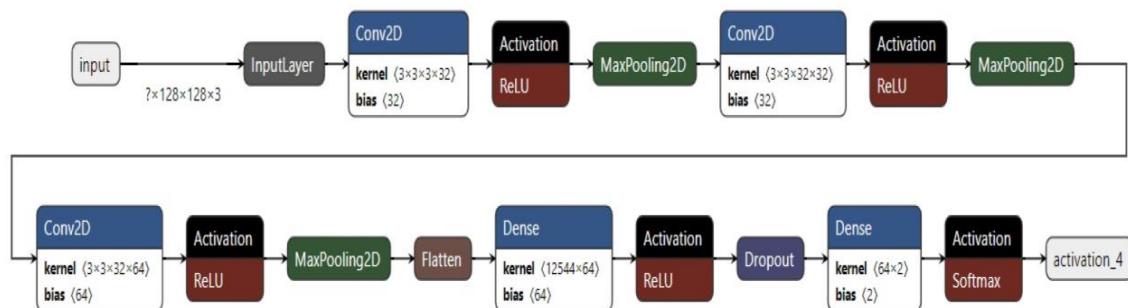
BR35H dataset (Hamada, 2020) contains 1500 negative and 1500 positive MRIs of brain tumors. The dataset is publicly accessible at <https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection>. Each MRI image is two-dimensional (2D) and has different dimension sizes. Twenty percent of the images from this dataset are used for testing the model. Before feeding to the DL model, the original images are pre-processed. They rescaled to a size 128x128, and the data type converted to float32. Figure 1 shows some examples from the used dataset. The title of each image indicates the image's label (yes/no).



**Figure 1 Examples of MR images from BR35H Dataset.**

## 2.2. The Deep Learning Model

A comparatively straightforward CNN architecture was implemented. The Keras package is used to build the proposed models, with TensorFlow (Abadi, et al., 2016) as the backend. The proposed model consists of two convolutional layers followed by two dense layers preceding the softmax layers and 2x2 max-pooling was applied to each convolutional layer's output of models. The 2D model's input data shape is 128x128. The activation function of the convolution output of the model was ReLU. In the training and validation processes, implemented models were trained for fifty epochs. The architecture of the model is demonstrated in Figure 2.

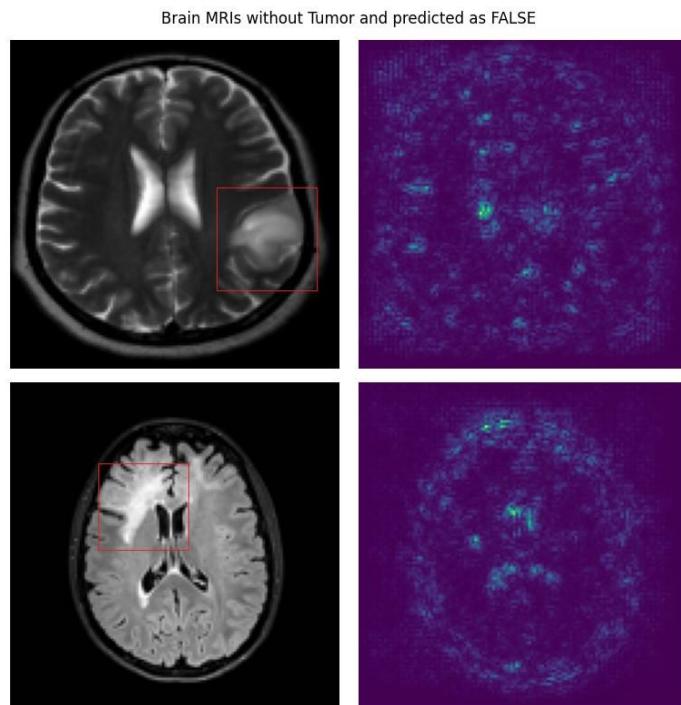


**Figure 2: The Architecture of implemented CNN Model**

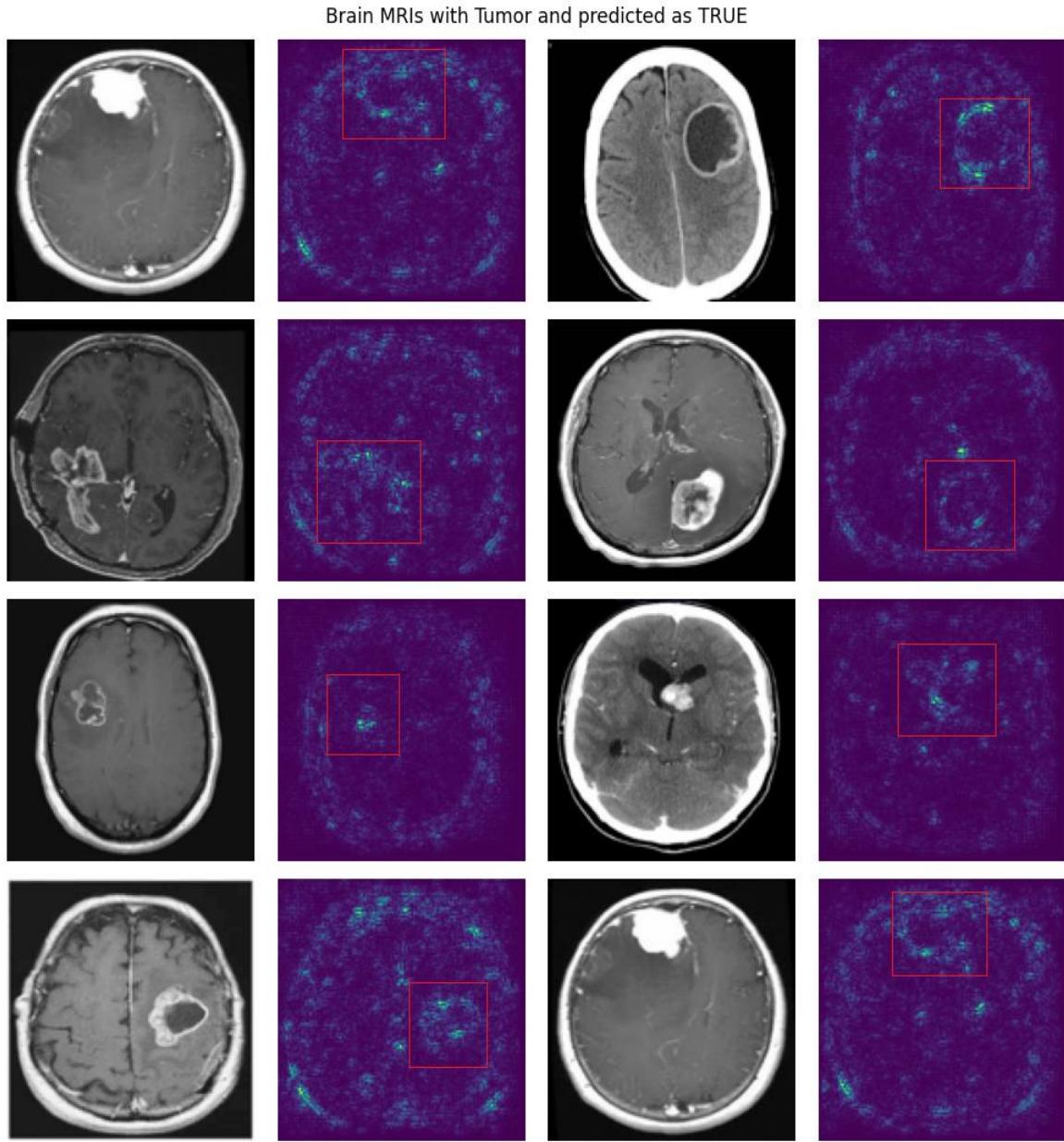
### 2.3. Saliency Maps of the models:

Mathematically, a saliency map is the derivative of the class probability of the input image. The calculated saliency maps were created with the TensorFlow library which provides a built-in function. The study involved the calculation of saliency maps for MRI images containing tumors, specifically examining two groups: correctly predicted MRIs (comprising eight images) and misclassified MRIs (consisting of two images). As the primary focus of this investigation was to evaluate the influence of tumor presence on classification, saliency maps for healthy brain images (those without tumors) were not generated. By directing attention solely to tumor-related images, the study aimed to gain valuable insights into the impact of tumor presence on the classification process. The third figure (Figure 3) of this study presents a comparative analysis comprising two brain MRI images containing tumors that were misclassified by the DL model, along with their respective saliency maps. Within this figure, the tumor regions are visually emphasized through square markings overlaid on the MRI images. Notably, an observation is made that the saliency maps for these images fail to effectively highlight the tumor regions. Consequently, it becomes evident that these tumors occupy a more extensive spatial extent within the brain compared to the accurately predicted tumors (Figure 4), leading to blurred boundaries and a lack of distinctiveness in the corresponding saliency maps.

Figure 4 in the study displays the MRI collection, including brain tumor cases accurately predicted by the model. In contrast to Figure 3, Figure 4 visually highlights regions associated with the presence of tumors on the salience maps using squares, rather than on MRI images. In this way, distinctive pixels are shown to be concentrated around the tumor, presenting accurately classified MRI images and their corresponding salience maps. This demonstration provides valuable information that contributes significantly to the prediction of the model.



**Figure 3 shows FALSE predicted 2 MRIs with brain tumors and related saliency maps of them. Tumors in MRI images were visually depicted using squares to show regions.**



**Figure 4 shows TRUE predicted 8 MRIs with brain tumors and related saliency maps of them. To illustrate the specific regions which coincided with the tumor coordinates in the MRI images, salient regions on the maps were visually depicted using squares.**

### 3. Results

The Confusion Matrix (CM) is a fundamental tool for assessing the performance of a binary or categorical classifier. The F1 score (or Dice similarity coefficient) metric was commonly used to measure the ratio of overlap for both classes by taking harmonic mean precision formulated as  $TP/(TP + FP)$  and recall(sensitivity) formulated as  $TP/(TP + FN)$ , derived from the confusion matrix. The CNN model's CM and calculated metrics from CM are presented in Table 1.

**Table 1 Confusion Matrix for 2-Class Classification**

Confusion Matrix			Calculated Metrics				
		Predicted					
Actual	No Yes		Precision	Recall	F1-score	Sample Number	
	No	29	0.85	0.83	0.84	35	
	Yes	5	0.73	0.76	0.74	21	

## 4. Discussions

The primary objective of this study was not to develop a highly accurate model. Rather, our focus was on interpreting the model's estimation using saliency maps. However, the accuracy metrics of the implemented model are presented in the results section to provide a quantitative measure of its predictive capabilities. Consequently, we will beware of discussing a model with improved accuracy in this context.

The saliency maps generated using gradient-based analysis revealed that the DL model's predictions strongly focus on areas corresponding to tumor presence in the given MRI image. This finding provides valuable insights into how the model makes decisions. However, they often lack a clear threshold to determine the significance of salient regions and may struggle to capture context and global relationships in complex scenes or images. During the subsequent discussion, we interpreted the saliency maps alongside available information and expert input from the field. We took these limitations into account, aiming to provide reliable and meaningful results.

Upon careful examination of the saliency maps depicted in the provided figures, it becomes apparent that the regions corresponding to the bones in the skull-stripping influence the decision-making process of the DL model. Consequently, incorporating a preprocessing step that involves bone extraction before training the model can improve accuracy. This bone extraction process would effectively shift the attention of the model towards the soft tissue regions, thereby enhancing its capacity to classify brain tumors more accurately.

The misclassified tumor images become evident that these tumors possess more complicated shapes. These unsteady shapes are challenging for the model to accurately detect. Consequently, to increase accuracy in tumor classification, it is necessary to enrich the dataset with a broader range of brain MRI images that include tumors exhibiting complex shapes. By training the model with an expanded dataset comprising such diverse tumor shapes, it is expected to strengthen its ability to discern and classify tumors with higher precision and reliability.

There are advantages and drawbacks to using saliency maps in medical image processing. Interpretability is the most significant of its benefits since the aim is to draw a visual explanation for the model's decision-making process, and eventually, to explain the inner workings and improve in case of necessity. They can also guide attention by indicating the most informative regions in an image, drawing a region of interest (ROI), and enabling more efficient processing on this focus. For the detection tasks, attention guidance can enhance performance by reducing the search space and improving accuracy.

## 5. Conclusions

The field of XAI is still evolving, and there is ongoing research and development aimed at improving the interpretability and explainability of AI models. Although they are not the sole solution, saliency maps remain a valuable tool in this domain, and new methods will continue to emerge as the field progresses. In our conducted case study, we have arrived at the finding that incorporating a preprocessing step that involves bone extraction before model training will be beneficial, as it will redirect the model's attention towards soft tissue regions, leading to improved accuracy in classifying brain tumors using MRI images. Furthermore, the misclassification reveals that the model struggles with tumors exhibiting complex shapes. To enhance classification accuracy, it is essential to enrich the dataset with a broader range of brain MRI images that encompass tumors with diverse and intricate shapes. Finally, the incorporation and careful interpretation of saliency maps can provide valuable insights, improve model performance, and facilitate the understanding of AI model behaviour in medical image processing.

However, it is essential to consider both the benefits and drawbacks of using saliency maps to ensure their appropriate and effective utilization in practical applications.

## 6. Acknowledgement:

This research was supported by Science Foundation Ireland under grant number 13/RC/2106\_P2 (ADAPT SFI Centre for Software). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Abadi et al., 2021] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," arXiv preprint arXiv:1603.04467, 2016.
- [Ali et al., 2023] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., M. Alonso-Moral, J., Confalonieri, R., Guidotti, R., D. Ser, J., Díaz-Rodríguez, N., Herrera, F., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," Information Fusion, p. 101805, 2023.
- [Andrushia et al., 2021] D. Andrushia, A., M. Sagayam, K., Dang, H., Pomplun, M., Quach, L., "Visual-Saliency-Based Abnormality Detection for MRI Brain Images—Alzheimer's Disease Analysis," Applied Science, vol. 11, no. 9, p. 9199, 2021.
- [Amann et al., 2020] Amann, J., Blasimme A., Vayena, E., Frey, D., and I. Madai, V., "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC medical informatics and decision making, vol. 20, pp. 1-9, 2020.
- [Amorim et al., 2021] P. Amorim, J., H. Abreu, P., Santos, J., Cortes, M., Vila, V., "Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations," Information Processing & Management, vol. 60, no. 2, p. 103225, 2023.
- [Ayhan et al., 2022] S. Ayhan, M., B. Kümmeler, L., Kühlewein, L., Inhoffen, W., Aliyeva, G., Ziemssen, F., Berens, P., "Clinical validation of saliency maps for understanding deep neural networks in ophthalmology," Medical Image Analysis, vol. 77, p. 102364, 2022.
- [Chmiel et al., 2021] Chmiel, W., Kwiecie, J., Motyka, K., "Saliency Map and Deep Learning in Binary Classification of Brain Tumors," Sensors, vol. 23, no. 9, p. 4543, 5 May 2023.
- [Hamada, 2021] Hamada, A., "Br35h: Brain Tumor Detection 2020," 2020. [Online]. Available: <https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection>. [Accessed 31 05 2023].
- [Jin et al., 2021] Jin, W., LI, X., HAMARNEH, G., "One map does not fit all: Evaluating saliency map explanation on multi-modal medical images," arXiv preprint, vol. arXiv:2107.05047, 2021.
- [Lamichhane et al., 2023] Lamichhane, K., Carli, M., Battisti, F., "A CNN-based no reference image quality metric exploiting content saliency," Signal Processing: Image Communication, vol. 111, p. 116899, 2023.
- [Muhammad et al., 2021] Muhammad, K., Khan, S., D. Ser, J., H. C. d. Albuquerque, V., "Deep Learning for Multigrade Brain Tumor Classification in Smart Healthcare Systems: A Prospective Survey," IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, vol. 32, no. 2, pp. 507-523, February 2021.

- [Nauta et al., 2021] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., V. Keulen, M., Seifert, C., "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," arXiv preprint, vol. arXiv:2201.08164, 2022.
- [Rahimi and Jain, 2021] Rahimi, A., Jain, S., "Testing the effectiveness of saliency-based explainability in NLP using randomized survey-based experiments," arXiv, vol. arXiv:2211.15351, 2022.
- [Ranjbarzadeh et al., 2021] Ranjbarzadeh, R., Caputo, A., B. Tirkolaei, E., J. Ghoushchi, S., Bendechache, M., "Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools," Computers in Biology and Medicine, p. 106405, 2022.
- [Rasheed et al., 2022] Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., Qadir, J., "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," Computers in Biology and Medicine, vol. 149, p. 106043, 2022.
- [Reyes et al., 2020] Reyes, M., Meier, R., Pereira, S., A. Silva, C., M. Dahlweid, F., V. Tengg-Kobligk, H., M. Summers, R., Wiest, R., "On the interpretability of artificial intelligence in radiology: challenges and opportunities," Radiology: artificial intelligence, vol. 2, no. 3, p. e190043, 2020.
- [RichardWebster et al., 2022] RichardWebster, B., Hu, B., Fieldhouse, K., H. Kitware, A., "Doppelganger Saliency: Towards More Ethical Person Re-Identification," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
- [Sina et al., 2021] Sina, M., Zarei, N., D. Ragan, E., "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 11, no. 3-4, pp. 1-45, 2021.
- [Singh et al., 2020] Singh, A., Sengupta, S., Lakshminarayanan, V., "Explainable deep learning models in medical image analysis," Journal of Imaging, vol. 6, no. 6, p. 52, 2020.
- [Vries et al., 2021] D. Vries, B., J. C. Zwezerijnen, G., L. Burchell, G., H. P. v. Velden, F., W. M.-v. d. H. v. Oordt, C., Boellaard, R., "Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review," frontiers in Medicine, vol. 10, 2023.
- [Wang, 2021] J. Wang, Z., "Probing an AI regression model for hand bone age determination using gradient-based saliency mapping," Scientific reports, vol. 11, no. 1, p. 10610, 2021.

# Domain Generalisation with Bidirectional Encoder Representations from Vision Transformers

Hamza Riaz and Alan F. Smeaton

*Dublin City University, Glasnevin, Dublin 9, Ireland  
alan.smeaton@dcu.ie*

## Abstract

Domain generalisation involves pooling knowledge from source domain(s) into a single model that can generalise to unseen target domain(s). Recent research in domain generalisation has faced challenges when using deep learning models as they interact with data distributions which differ from those they are trained on. Here we perform domain generalisation on out-of-distribution (OOD) vision benchmarks using vision transformers. Initially we examine four vision transformer architectures namely ViT, LeViT, DeiT, and BEiT on out-of-distribution data. As the bidirectional encoder representation from image transformers (BEiT) architecture performs best, we use it in further experiments on three benchmarks PACS, Home-Office and DomainNet. Our results show significant improvements in validation and test accuracy and our implementation significantly overcomes gaps between within-distribution and OOD data.

**Keywords:** Domain generalisation, vision transformers, benchmarking

## 1 Introduction

Machine learning algorithms rely on training data distributions though in certain scenarios computer vision can fail to generalise when applied to out-of-distribution (OOD) data. Real-world systems like autonomous vehicles can show a decline in performance when interacting with even partially different conditions and settings compared to their training data distributions.

Domain generalisation is influenced by three factors: dataset types, network architectures, and model selection criteria. To overcome OOD challenges, much work has been done including solutions like additional data collection for different domains, adversarial learning, and data augmentation for learning generalised invariances from the training domain [Akuzawa et al., 2020]. A range of pointers from the literature encourages us to explore vision transformers for domain generalisation in computer vision. We implemented a pipeline to determine the OOD capability of four available pre-trained vision transformers. Originally each were pre-trained and fine-tuned on ImageNet-21k and ImageNet1k respectively. Using these, we run inference on unseen benchmarks including ImageNet-Sketch, ImageNet-R (endition), Imagenet Adversarial, and Imagenet Corrupted. From these results, we choose BEiT for further analysis and we fine-tune three separate models on three popular benchmarks for domain generalisation namely PACS, Home-Office and DomainNet.

## 2 Vision Transformers

Vision transformers are inherently more appropriate for domain generalisation compared to other CNNs because of factors like global understanding, handling variable-length inputs, fewer parameters, an attention mechanism, and pre-training. A vision transformer uses the transformer architecture to analyse images for various downstream tasks. A simple transformer architecture was initially proposed for natural language processing tasks in [Wolf et al., 2020] which was extended into the vision transformer in order to handle image

data [Dosovitskiy et al., 2020]. The main innovation in the vision transformer is its ability to process an entire image as a sequence of patches rather than as a grid of pixels. Vision transformers use self-attention during the learning process and an attention score is computed by the product of query-key terms in the last layer.

### 3 OOD Inference Experiments with Pre-trained Vision Transformers

To conduct initial OOD experiments we use pre-trained weights from 4 baseline vision transformers namely ViT, LeViT, DeiT and BEIT. These are fine-tuned on ImageNet 2012 1K classes with 224x224 input resolution and 16x16 patch size except LeViT which has 256x256 input resolution. For performance of these on OOD datasets, we use variations of ImageNet as OOD examples namely ImageNet Sketch, ImageNet-R(edition), Imagenet Adversarial, and Imagenet Corrupted. The ImageNet-Sketch dataset has 50,000 images with 1K classes, 50 images for each of the 1,000 ImageNet classes. ImageNet-R contains 30,000 image renditions for 200 ImageNet classes which is a subset of ImageNet-1K. ImageNet-adversarial consists of adversarially filtered real-world images to fool ImageNet classifiers and it also contains 200 classes as a subset of ImageNet-1K. Finally Imagenet Corrupted consist of images with 75 common visual distractions and the goal was to improve and evaluate the robustness of models, it has 1,000 classes.

Table 1: Results using 4 vision transformers (rows) on 4 OOD related benchmarks (columns)

Models	ImageNet-Sketch		ImageNet-R(edition)		Imagenet Adversarial		Imagenet Corrupted	
	Top1 Acc	Top5 Acc	Top1 Acc	Top5 Acc	Top1 Acc	Top5 Acc	Top1 Acc	Top5 Acc
ViT	35.43	57.29	32.82	47.54	12.97	30.04	78.06	94.43
LeViT	0.95	0.72	0.81	0.44	9.13	27.14	73.67	90.98
DeiT	32.58	50.21	31.04	44.42	9.97	24.31	77.95	92.56
BEIT	<b>47.55</b>	<b>71.01</b>	<b>44.72</b>	<b>62.13</b>	<b>22.60</b>	<b>47.74</b>	<b>81.88</b>	<b>96.41</b>

Table 1 presents the top-1 and top-5 accuracy figures for 4 selected transformers on 4 OOD-related benchmarks. Results indicate that BEIT outperforms given transformers with a notable improvement in evaluation metrics for each benchmark. The main reason BEIT surpasses others are its properties including Mask Image Modeling (MIM) with self-supervised learning of large models, a self-attention mechanism, and denoising of corrupted inputs. Thus we selected BEIT for analysis of domain generalisation benchmarks. The next section presents the methodology of our approach.

### 4 Domain Generalisation Experiments

The BEIT vision transformer was applied to the PACS, Office-Home, and DomainNet benchmarks to test the domain generalisation capability of BEIT for small, medium and large datasets. PACS has 9,991 images with 4 domains and 7 classes which is a comparatively smaller dataset. Office-Home has 15,588 images also with 4 domains but it has 65 classes. DomainNet is one of the largest benchmarks for domain generalisation with more than 0.5 million images, 6 domains and 365 classes. Although fine-tuning of any vision transformer is a relatively less time-consuming process than pre-training from scratch, this also depends on the size of the dataset. For instance, PACS and Office-Home take almost 4-6 hours for fine-tuning but in the case of DomainNet our model takes almost 3 days. During the training and validation steps, pre-processing includes image resizing, random horizontal flip, and normalisation. Similarly, testing includes re-sizing, centre cropping, and normalisation.

Inspired by the work in [Bao et al., 2022], we used the based version of BEIT transformer which has 12 transformer layers with 768 hidden and 3,072 feed-forward networks. Each attention layer has 12 attention heads of size 64 and these are responsible for learning self-attention masks. Each image was divided into 14\*14 patches of 16\*16 pixels. BEIT is trained with 8,192 visual tokens. The version of pre-trained weights which we used were pre-trained and fine-tuned on ImageNet 21k.

## 5 Experimental Results on OOD Benchmarks

Following fine-tuning of hyperparameters for OOD datasets, we applied well-trained and shallow networks of BEIT on the unseen testing sets from each benchmark using <https://github.com/huggingface/transformers>.

Table 2: Results of BEIT fine-tuning experiments on three benchmarks – PACS, Office-Home, DomainNet

Benchmarks	Validation Top1 Acc	Target Top1 Acc	Validation Top5 Acc	target Top5 Acc	Gap	Precision
PACS	0.96	0.94	1.0	0.9980	0.02	0.9464
Office-Home	0.8597	0.8691	0.9948	0.9679	-0.0094	0.8754
DomainNet	0.7019	0.6978	0.9347	0.8793	0.0041	0.7111

Table 2 presents results of the vision transformer-based domain generalised model including the top-1 and top-5 scores for validation and target/test data distributions. Ideally, in domain generalisation, the gap is the difference in performance metrics like accuracy, loss or precision for Independent and Identically Distributed (IID) and for Out Of Domain (OOD) test data. Here we consider the validation distribution as IID and the target distribution as OOD and the gap is the difference between these. For all three benchmarks, our vision transformer-based model shows state-of-the-art performance. In the case of PACS, it has 0.94 accuracy and the gap is only 0.02, a sign of good domain generalisation. For Office-Home, the OOD or target accuracy is higher than the validation accuracy of IID which made the gap negative. Although our model has relatively low performance for DomainNet and Office-Home compared to PACS the gap remains small which means the generalisation performs effectively.

Table 3: BEIT accuracy and loss for PACS, Office-Home, and DomainNet for each domain independently

PACS	Photos	Artwork	Cartoon	Sketch		
Accuracy	0.9766	0.9183	0.9578	0.9371		
Loss	0.0493	0.2507	0.1206	0.2227		
Office-Home	Art	Clipart	Product	Real World		
Accuracy	0.7979	0.8488	0.9324	0.8645		
Loss	0.7443	0.5947	0.2502	0.5339		
DomainNet	Clipart	Infograph	Painting	Quickdraw	Real World	Sketch
Accuracy	0.7822	0.3812	0.6893	0.6727	0.8073	0.6764
Loss	0.9129	3.0639	1.4036	1.2023	0.7915	1.4661

Table 3 presents loss and accuracy metrics for each domain. The first rows present the performance of BEIT-PACS which has 4 domains namely photos, artwork, cartoon, and sketch. It is clear that the model has good performance for the samples of photos but lower accuracy and loss for samples related to the artwork domain. The rows in Table 3 show performance of BEIT-Office-Home which also has 4 domains namely art, clipart, product, and real world. Like BEIT-PACS, BEIT-Office-Home also shows lower scores for the artwork domain, possibly because pre-trained weights do not have the high-level features because not enough artwork images were used in training. In the case of BEIT-DomainNet, the model has 6 domains, 365 classes and more than 0.5 million samples from across different domains. The model has poorer performance for samples of infographs as it is a different domain to the others.

To compare our method with the state-of-the-art, Table 4 presents various CNN-based domain generalised algorithms using performance figures taken from [Riaz and Smeaton, 2023],[Gulrajani and Lopez-Paz, 2021]. The columns present IID accuracy, OOD accuracy, and the ensuing gap, for PACS and for Office-Home. For DomainNet we consider target accuracy only. It is clear that our method substantially outperforms state-of-the-art approaches in all benchmarks. If we examine the figures for PACS, our method obtains overall 0.96 and 0.94 IID and OOD accuracy respectively and the gap shrinks to 0.02. Our method also shows similar performance

Table 4: Comparison between our trained model and other state-of-the-art methods for OOD generalisation

Models	PACS			Office-Home			DomainNet
	IID Accuracy	OOD Accuracy	Gap	IID Accuracy	OOD Accuracy	Gap	Target Accuracy
GroupDRO	0.95	0.73	0.22	0.82	0.52	0.30	0.337
ANDMask	0.95	0.72	0.23	0.81	0.44	0.37	*
Mixup	0.97	0.72	0.25	0.83	0.53	0.30	0.396
MMD	0.94	0.69	0.25	0.82	0.52	0.30	0.394
DANN	0.94	0.73	0.21	0.83	0.51	0.32	0.384
CORAL	0.95	0.77	0.18	0.84	0.55	0.29	0.418
VREx	0.97	0.80	0.17	0.76	0.49	0.27	0.336
RSC	0.97	0.77	0.20	0.83	0.50	0.33	0.389
ERM	0.97	0.78	0.19	0.84	0.57	0.27	0.412
<b>Our approach</b>	<b>0.96</b>	<b>0.94</b>	<b>0.02</b>	<b>0.86</b>	<b>0.87</b>	<b>-0.0094</b>	<b>0.70</b>

for Office-Home with a negative because our model perform slightly better on OOD than IID which is a good sign for domain generalisation. Table 4 also shows that overall target accuracy is not high for all approaches but our method still outperforms all existing approaches.

## 6 Conclusions

We present an investigation into vision transformers for domain generalisation in computer vision. We fine-tuned a base version of the BEiT model for domain generalisation benchmarks and investigated the generalisation of other pre-trained vision transformers on OOD versions of ImageNet. A per-domain analysis and detailed comparison with other domain generalisation algorithms showed our model yields state-of-the-art results or better for all benchmarks significantly reducing the gaps between IID and OOD scores in all benchmarks.

**Acknowledgements:** HR is funded under the ML-Labs SFI Centre for Researcher Training in Machine Learning (18/CRT/6183) and AS is part-funded by SFI [12/RC/2289\_P2] at Insight the SFI Research Centre for Data Analytics at DCU.

## References

- [Akuzawa et al., 2020] Akuzawa, K., Iwasawa, Y., and Matsuo, Y. (2020). Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 315–331. Springer.
- [Bao et al., 2022] Bao, H., Dong, L., Piao, S., and Wei, F. (2022). BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.
- [Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [Gulrajani and Lopez-Paz, 2021] Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. In *International Conference on Learning Representations*.
- [Riaz and Smeaton, 2023] Riaz, H. and Smeaton, A. F. (2023). Vision based machine learning algorithms for out-of-distribution generalisation. In *Computing Conference, London, UK*.
- [Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing*, pages 38–45.

# Dynamic Cost Volumes with Scalable Transformer Architecture for Optical Flow

Vemburaj Yadav, Alain Pagani, Didier Stricker

*Augmented Vision, German Research Center for Artificial Intelligence (DFKI)  
Kaiserslautern, Germany  
[{vemburaj.yadav, alain.pagani, didier.stricker}@dfki.de](mailto:{vemburaj.yadav, alain.pagani, didier.stricker}@dfki.de)*

## Abstract

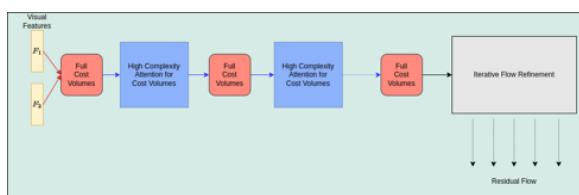
We introduce DCV-Net, a scalable transformer-based architecture for optical flow with dynamic cost volumes. Recently, *FlowFormer* [Huang et al., 2022], which applies transformers on the full 4D cost volumes instead of the visual feature maps, has shown significant improvements in the flow estimation accuracy. The major drawback of *FlowFormer* is its scalability for high-resolution input images, since the complexity of the attention mechanism on the 4D cost volumes scales to  $O(N^4)$ , with  $N$  being the number of visual feature tokens. We propose a novel architecture where we obtain the *FlowFormer* type enrichment of matching cost representations, but using light-weight attention on the visual feature maps with quadratic ( $O(N^2)$ ) complexity. Firstly, we generate sequential updates to the visual feature representations and, consequently, the cost volumes using lightweight attention layers. Secondly, we interleave this sequence of cost volumes with iterations of flow refinement, thereby modeling the update operator to handle dynamic cost volumes. Our architecture, with two orders of computational complexity lower than that of *FlowFormer*, demonstrates strong cross-domain generalization on the Sintel and KITTI datasets. We outperform *FlowFormer* on the KITTI dataset and achieve highly competitive flow estimation accuracies on the Sintel dataset.

**Keywords:** Transformer, Dynamic Cost Volumes, Scalability, etc.

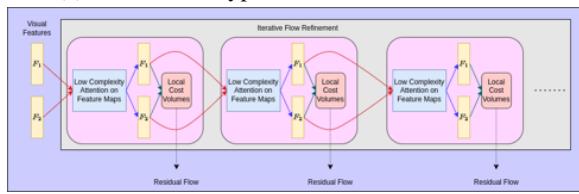
## 1 Introduction

Optical flow is a crucial concept in computer vision, aimed at estimating the per-pixel motion between frames and providing dense correspondences as valuable information for various downstream video tasks. Although optical flow has been studied since the seminal work of Horn and Schunck [Horn and Schunck, 1981], it remains a challenging problem due to difficulties such as fast-moving objects, occlusions, motion blur, and textureless surfaces.

Recently, optical flow estimation has witnessed a paradigm shift from the traditional optimization-based approach to deep learning-based methods. *RAFT* [Teed and Deng, 2020] makes it plausible to model near-global search space by constructing a  $H \times W \times H \times W$  cost volume that measures similarities between all pairs of pixels of the  $H \times W$  image pair. This is then complemented by an iterative framework



(a) FlowFormer type attention for Cost Volumes



(b) Our Dynamic Cost Volumes with scalable attention

Figure 1: Comparison of our DCV-Net with FlowFormer a) Static Cost Volume for refinement with Non-scalable attention blocks b) Dynamic Cost Volume for refinement with Scalable attention blocks.

that retrieves local costs within local windows to perform flow residual regression. The current state-of-the-art methods are all based on such RAFT-style architectures.

The feature representations being obtained from a CNN backbone, it lacks the global context, and thus the matching costs remains unreliable for occluded pixels and pixels undergoing large motions. Several follow-up works like *GMFlow* [Xu et al., 2022] and *GMFlowNet* [Zhao et al., 2022] incorporated self and cross attention layers into the overall architecture to learn global context-aware feature representations. *FlowFormer* [Huang et al., 2022], on the other hand, directly processes the 4D cost volumes with transformer blocks. They adopt an encoder-decoder architecture to first encode the matching costs into a latent cost space (cost memory), and then iteratively query this cost memory (similar to cost volume lookup) to regress the flow.

Although, *FlowFormer* has led to state-of-the-art flow estimation accuracy on several benchmarks, it has a major drawback in terms of its scalability to the resolution of the input image. Cost volume computation and storage in itself is one of the major bottlenecks for high resolution optical flow and real-time inference, since it scales quadratically ( $O(N^2)$ ) with the number of visual tokens  $N$  in the visual feature map. Since context aggregation by attention in itself scales quadratically to the number of input tokens, applying transformer blocks directly on the 4D cost volumes induces a complexity of  $O(N^4)$ . On the other hand, applying attention layers to the feature maps retains the complexity of the attention layers to  $O(N^2)$ . So, a clear argument could be made to still use this style of feature enhancement instead of the attention on 4D cost volumes, especially for high resolution inputs. We demonstrate in this work that we achieve state-of-the-art performances and very good cross-domain generalization, yet still using the light weight attention on the feature maps instead of the cost volumes.

Another major drawback, not limited to *FlowFormer*, but also to any transformer based architecture for optical flow, is the application of sequential attention blocks for feature enhancement in addition to the sequential nature of the iterative flow refinement. In the RAFT-style architectures, even though a full 4D cost volume is first computed and stored, only local costs within a local window are being looked up for each pixel at any given iteration. The local cost volumes could theoretically be computed on the fly at each iteration, thereby enabling significant reduction in memory requirements. However, these architectures do not consider updating and enriching the matching costs over the refinement iterations, thus limiting the update operator to work only with a one-time computed static cost volume. We derive motivations from these observations and propose an architecture, where the feature maps are contextualized with attention layers sequentially concurrent to the sequential regression of residual flow, thus updating the matching costs over the iterations, and thereby modeling the cost volume to be dynamic inside the architecture.

In this paper, we propose DCV-Net (Dynamic Cost Volume Network), a scalable transformer-based architecture for optical flow estimation with dynamic cost volumes. We emulate the *FlowFormer* type transformer-based updates to the matching costs, but at the same time enjoying the scalable attribute of the whole architecture for higher resolutions by leveraging the lower computational effort that comes with applying transformers to update the feature representations and computing on-the-fly dynamic local cost volumes, as shown in Fig. 1. We also propose a softmax based matching layer to transform the local matching costs into their corresponding matching distributions. This allows us to seamlessly incorporate a deep transformer architecture inside the refinement stage, while still respecting the RAFT-style optimization for flow regression with weight-tied update operators. We summarize our key contributions as follows:

1. A scalable transformer-based architecture for optical flow with dynamic cost volumes, which at the same time achieves competitive performance on all the benchmarks. Our approach could easily be scaled not only while running with high-resolution images, but also if the whole architecture is required to run with feature maps at a scale higher. In comparison, *FlowFormer* suffers from a poor scalability under both these scenarios.
2. A differentiable softmax-based matching layer to convert the dynamic local matching costs into matching distributions, thereby enabling a seamless integration of deep transformer based feature updates with the iterations of the flow refinement

3. A state-of-the-art cross validation performance on the KITTI dataset, and near-state-of-the-art cross validation performance on the Sintel dataset, while being only trained on synthetic datasets of Flying Chairs and Flying Things.
4. To our knowledge, this is the first attempt to demonstrate an architecture with transformers for optical flow, which is both scalable to higher resolution inputs, but at the same time delivers state-of-the art cross-domain generalization on real world datasets. Most of the previous works lacks in terms of one or more out of these 3 crucial aspects: scalability, generalization and performance.

## 2 Related Work

**Deep Learning for Optical flow:** Compared with traditional optimization-based flow estimation methods, data-driven methods directly learn to estimate optical flow from labeled data. Since *FlowNet* [Dosovitskiy et al., 2015], learning optical flow with neural networks has demonstrated superior flow estimation accuracies. Earlier works along these lines directly regress the optical flow by applying convolutional layers on the computed local cost volume. However, these architecture styles poses constraints on the channel dimension of the cost volumes, which restricts the search space to a local range, making it hard to tackle large displacements. Prominent works in these directions are *FlowNet* and *PWC-Net* [Sun et al., 2018], with the former predicting the flow from a local cost volume computed from the visual similarities of downsampled feature maps at a single resolution, while the latter builds hierarchical local cost volumes with warped features and progressively estimates flows from such local costs.

**Transformers for Optical Flow:** Exploiting the long-range modeling capacity of attention layers [Vaswani et al., 2017], Recent works ([Xu et al., 2022, Zhao et al., 2022, Huang et al., 2022]) employs transformers to further weaken the network-determined bias and learn feature relationships from data. *GMFlowNet* [Zhao et al., 2022] proposed to first contextualize the individual feature maps of both the frames by a sequential application of several self-attention layers, followed by the RAFT-style cost volume computation and iterative flow refinement. *GMFlow* [Xu et al., 2022], on the other hand enhances the feature maps with alternating applications of self and cross attention layers, thereby providing also the inter-frame context to the feature maps. However, the dense correspondence field is directly regressed from the cost volume using the softmax based differentiable matching layer. The lack of iterative refinement, along with the presence of heavy transformer blocks in *GMFlow* massively degrades the generalization ability of the architecture, which being demonstrated from their inferior cross-validation performance on the KIITI dataset. *FlowFormer* achieves state-of-the-art accuracy by replacing the CNN based backbone for feature extraction with a transformer, and directly applying attentions on the 4D cost volumes.

## 3 Method

Given a pair of source and target RGB images  $I_1$  and  $I_2$  respectively, optical flow aims at recovering a dense correspondence field  $(f^1, f^2)$  for every source pixel  $(u, v)$  in  $I_1$ , such that it maps to the locations  $(u', v')$  in  $I_2$ , with  $(u', v') = (u + f^1(u, v), v + f^2(u, v))$ . Our network design is based on the amalgamation of attention layers into the successful *RAFT* [Teed and Deng, 2020] architecture. Fig. 2 represents an overall visual representation of our network architecture. For completeness, refer to our supplementary material for a brief description of 1) the key components of *RAFT* architecture and 2) the attention mechanism in transformers.

### 3.1 Dynamic Cost Volume

The core contribution of our work lies at the heart of modeling the 4D cost volumes to be dynamic inside the architecture, but at the same time maintaining the scalability of the model to the input image resolution. The temporal nature of the cost volumes is inherently suited to our proposed integration of them within the iterative refinement stage of the architecture, which also operates in a temporal fashion. We therefore first present the

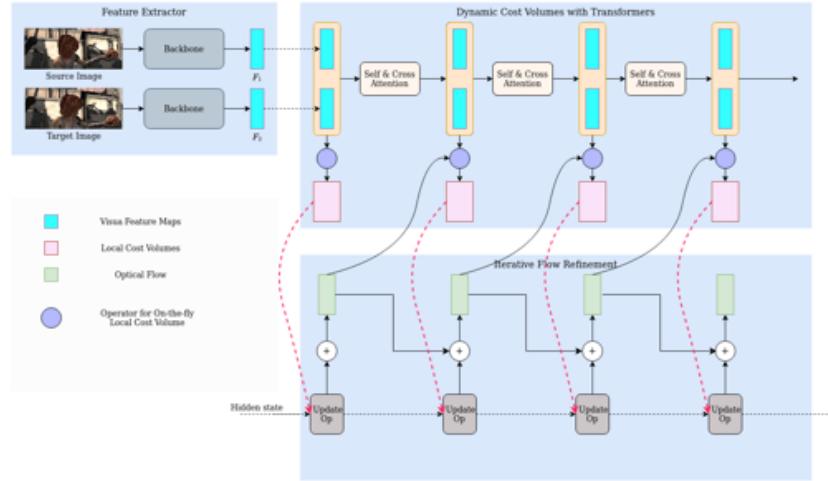


Figure 2: Architecture of DCV-Net. DCV-Net consists of 3 main components: 1) A sequence of visual feature maps generated by layers of a transformer architecture. 2) A sequence of Local Cost Volumes (on-the-fly computed variants of Dynamic Cost Volumes). 3) Iterative Flow refinement stage which recurrently regresses the residual flow using the Dynamic Cost Volumes.

mathematical formulation of our dynamic cost volume with scalable attention. To provide a comprehensive understanding of our proposed approach in comparison to previous works, we introduce two additional formulations. First, we consider an approach similar to *FlowFormer*, where transformer layers are directly applied to the 4D cost volume, resulting in a sequence of cost volumes. Second, we consider approaches similar to *GMFlow* and *GMFlowNet* where transformer layers are applied to the visual feature maps, generating a sequence of feature maps with richer representations. In both scenarios, the cost volume remains static during the refinement stage.

**Attention on Cost Volumes:** Given the feature maps  $F_1$  and  $F_2$  extracted from a CNN backbone, a 4D cost volume  $\mathbf{C}$  is constructed to store similarity scores for potential matches between the frames. Enhancing the matching distribution for a specific source pixel involves considering the distributions of nearby and contextually similar source pixels. This fosters consensus among neighboring matches, resulting in more accurate flow estimation, particularly for occluded pixels. To generate a sequence of cost volumes  $\mathbf{C}^{[i]}$  with improved matching costs, the cost volume is passed through a transformer architecture that employs self-attention using all the  $H \times W \times H \times W$  tokens. Let  $M$  transformer blocks, each parameterized by  $\theta^{[i]}$ , be used to generate  $L$  stages of cost volumes  $\mathbf{C}^{[j]}$ , where  $\mathbf{C}^{[0]}$  is the initial cost volume  $\mathbf{C}$ . The final cost volume  $\mathbf{C}^{[L]}$  would then be looked-up in the Iterative Residual Refinement (IRR) stage to regress the optical flows  $\mathbf{f}^{[k]}$  for  $K$  iterations. A mapping function  $p$  determines the index  $i$  of the transformer block used in the  $j^{th}$  stage of the architecture, allowing for sharing of transformer blocks across multiple stages, which is advantageous for our formulation with dynamic cost volumes. In a straightforward case of  $L$  stages and  $L$  transformer blocks, each transformer block is used in a single stage, resulting in the mapping function  $p(j) = j$ .

$$\begin{aligned} \mathbf{C}^{[j]} &= \text{Transformer}(\mathbf{C}^{[j-1]}; \theta^{p(j)}), \quad j = 1, 2, \dots, L \\ \mathbf{f}^{[k]} &= \text{IRR}(\mathbf{C}^{[L]}, \mathbf{f}^{[k-1]}), \quad k = 1, 2, \dots, K \end{aligned} \tag{1}$$

With  $N$  as the number of visual feature tokens, attention on the cost volumes with  $N^2$  tokens induces a complexity of  $O(N^4)$  both in terms of computational effort and memory requirements. *FlowFormer* tries to mitigate this by patchifying the 4D cost tokens and using a complex latent cost encoder and decoder architecture. On the other hand, our architecture enjoys the simplicity in terms of its integration into a RAFT-style architecture.

**Attention on Feature Maps and Static Cost Volume:** Here, instead of processing the 4D cost volume with a transformer architecture, the feature maps are sequentially enhanced using transformer blocks with self and cross attention layers. Let  $\mathbf{F}_1^{[i]}, \mathbf{F}_2^{[i]}$  be the sequence of feature maps generated over  $L$  stages using  $M$  distinct transformer blocks as before. The feature maps  $\mathbf{F}_1^{[L]}, \mathbf{F}_2^{[L]}$  from the final stage would then be used to construct the cost volume  $\mathbf{C}$ , which would then be used inside IRR. Unlike the previous formulation with attention on cost volumes, this formulation presents much more light-weighted architecture and superior scalability due to the quadratic complexity of the attention blocks being applied on visual feature maps.

$$\begin{aligned} (\mathbf{F}_1^{[j]}, \mathbf{F}_2^{[j]}) &= \text{Transformer}(\mathbf{F}_1^{[j-1]}, \mathbf{F}_2^{[j-1]}; \theta^{p(j)}), \quad j = 1, 2, \dots, L \\ \mathbf{C} &= f(\mathbf{F}_1^{[L]}, \mathbf{F}_2^{[L]}) \\ \mathbf{f}^{[k]} &= \text{IRR}(\mathbf{C}, \mathbf{f}^{[k-1]}), \quad k = 1, 2, \dots, K \end{aligned} \quad (2)$$

**Attention on Feature Maps and Dynamic Cost Volume:** Previous works like *GMFlow* and *GMflowNet* employing the formulation with attention on feature maps lags behind significantly in terms of flow estimation accuracy, in comparison to *FlowFormer* type approaches where directly the cost volumes are contextualized using attention. Also, in both both formulations above, cost volume remains static over the iterations of the update operator. We, therefore propose a formulation where we combine the advantages of the both the mentioned formulations, i.e, 1) we maintain the scalability of the architecture by sequentially enhancing the feature maps using attention 2) we simultaneously generate a sequence of cost volumes from the sequence of feature maps, and make this whole sequence of cost volumes dynamic by utilizing it over the iterations of the flow refinement. Unlike the previous 2 formulations where the refinement stage utilizes a fixed pre-computed cost volume, we concurrently update the cost volumes over the iterations of the flow refinement. A sequence of  $L$  feature maps  $\mathbf{F}_1^{[i]}, \mathbf{F}_2^{[i]}$  and their corresponding cost volumes  $\mathbf{C}^{[i]}$  are obtained using a module with  $M$  transformer blocks. Since, we have  $L$  stages of the cost volume and  $K$  iterations of refinement, the first  $L$  iterations of refinement benefits from the usage of dynamic cost volumes, and for the remaining  $K - L$  iterations, the cost volume from the final stage would be utilized. Since only a portion of the full cost volume is being utilized at each refinement iteration, we compute these local cost volumes on-the-fly using the implementation proposed by authors from RAFT. We take advantage of the significant reduction in the memory requirements that this implementation brings for our dynamic cost volumes.

$$\left. \begin{array}{l} (\mathbf{F}_1^{[j]}, \mathbf{F}_2^{[j]}) = \text{Transformer}(\mathbf{F}_1^{[j-1]}, \mathbf{F}_2^{[j-1]}; \theta^{p(j)}) \\ \mathbf{C}^{[j]} = f(\mathbf{F}_1^{[j]}, \mathbf{F}_2^{[j]}) \\ \mathbf{f}^{[j]} = \text{IRR}(\mathbf{C}^{[j]}, \mathbf{f}^{[j-1]}) \end{array} \right\} j=1, 2, \dots, L$$

$$\mathbf{f}^{[k]} = \text{IRR}(\mathbf{C}^{[L]}, \mathbf{f}^{[k-1]}), \quad k = L + 1, L + 2, \dots, K$$

### 3.2 Cost Volume LookUp based on Matching Probability

The GRU based update operator in the IRR is responsible for regressing the residual flows based on the previous flow estimate, context features, hidden state and local matching costs sampled from the cost volume. The parameters of this update operator are weight-tied to mimic its functionality as a optimization solver. For static cost volumes, the distribution of the similarity score values encountered by the update operator does not change over the refinement iterations. On the other hand, the distribution of our feature map activations differs from one iteration to the other since they are the output of transformer blocks with non-shared parameters. This leads to unstable training of the parameters of update operator. To mitigate this issue, we propose to transform the sampled local matching costs into their corresponding matching probabilities. We accomplish this by applying *softmax* to the matching costs over the sampled co-ordinate locations. The update operator therefore receives such probability measures as input, irrespective of the distributions of matching costs generated by distinct transformer blocks.

Training Data	Method	Sintel		KITTI	
		Clean	Final	F1-epe	F1-all
C + T	HD3	3.84	8.77	13.17	24.0
	LiteFlowNet	2.48	4.04	10.39	28.5
	PWC-Net	2.55	3.93	10.35	33.7
	LiteFlowNet2	2.24	3.78	8.97	25.9
	S-Flow	1.30	2.59	4.60	15.9
	RAFT	1.43	2.71	5.04	17.4
	FM-RAFT	1.29	2.95	6.80	19.3
	GMA	1.30	2.74	4.69	17.1
	GMFlow	1.08	2.48	-	-
	GMFlowNet	1.14	2.71	4.24	15.4
	CRAFT	1.27	2.79	4.88	17.5
	SKFlow	1.22	2.46	4.47	15.5
	FlowFormer	<b>0.94</b>	<b>2.33</b>	4.09	14.72
	DCV-Net (Ours)	0.99	2.43	<b>3.91</b>	<b>14.27</b>

Table 1: Results on Sintel and KITTI datasets. We test the generalization performance after training on FlyingChairs(C) and FlyingThing(T)

## 4 Experiments

Following previous works, we train our models first on the FlyingChairs [Dosovitskiy et al., 2015] and then on the FlyingThings3D [Mayer et al., 2016]. Please refer to the supplementary material for the implementation details and the training schedules. We then evaluate our model on the training splits of Sintel [Butler et al., 2012] and KITTI [Geiger et al., 2013] to demonstrate cross-domain generalization.

### 4.1 Quantitative Experiments

Table 1 shows the cross-domain generalization performance of our model in comparison with the recent works from the literature. Although our DCV-Net has two orders of complexity than that of *FlowFormer*, we outperform it on the KITTI dataset, while staying highly competitive on the Sintel dataset. Also, we outperform both *GMFlow* and *GMFlowNet* by a significant margin on all the benchmarks, even though we employ transformer blocks of similar complexity on the visual feature maps.

### 4.2 Qualitative Experiments

We visualize flows estimated by our DCV-Net and *FlowFormer* for four examples in Fig. 3. We show that the flow estimated by our DCV-Net is similar in quality to that of *FlowFormer*, since we are able to recover high quality motions in overlapping object regions, occluded areas and regions with large displacements, with a scalable model of much lower complexity.

### 4.3 Ablation Study

We perform a set of ablations to demonstrate the effectiveness of our proposed contributions. We start with *RAFT* as the baseline with a CNN backbone and a static cost volume being decoded by a recurrent decoder to iteratively regress the optical flow. We then gradually alter different components of the baseline with our proposed components and show their significance on the performance of flow estimation.

**Static vs Dynamic Cost Volume:** We demonstrate the importance of having the cost volumes to be dynamic over the iterations of flow refinement and compare it with the case with the RAFT-style static cost volume. It is

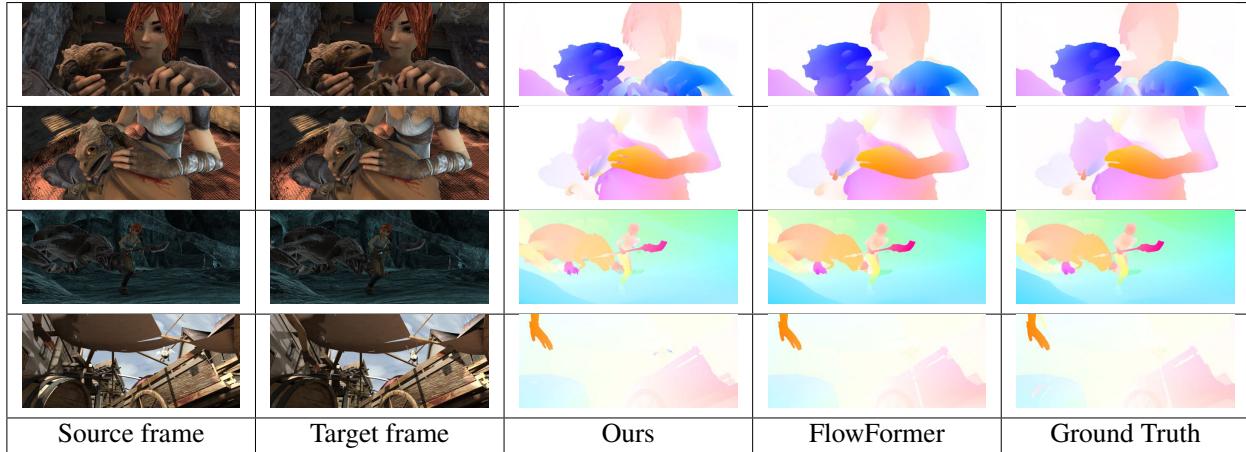


Figure 3: Qualitative comparison on the Sintel training set

Cost Volume	Sintel		KITTI	
	Clean	Final	F1-epc	F1-all
Static	1.51	2.65	4.61	16.11
Dynamic	<b>1.23</b>	<b>2.63</b>	<b>4.21</b>	<b>15.99</b>

(a) Static vs Dynamic Cost Volume

Transformer Blocks	Sintel		KITTI	
	Clean	Final	F1-epc	F1-all
No weight sharing	1.05	2.47	4.05	14.77
Shared weights	<b>1.02</b>	<b>2.41</b>	<b>3.99</b>	<b>14.58</b>

(b) Transformer blocks with and without weight sharing

Feature Enhancement	Sintel		KITTI	
	Clean	Final	F1-epc	F1-all
Before IRR	<b>1.00</b>	2.44	4.29	15.46
Inside IRR	1.02	<b>2.41</b>	<b>3.99</b>	<b>14.58</b>

(c) Feature enhancement: before refinement vs concurrent to refinement

Cost Representation	Sintel		KITTI	
	Clean	Final	F1-epc	F1-all
Matching Cost	1.22	2.76	4.56	16.24
Matching Probability	<b>1.02</b>	<b>2.41</b>	<b>3.99</b>	<b>14.58</b>

(d) Matching cost vs matching probability based cost volume lookup

Table 2: Ablation studies

evident from Table 2a that having the cost volumes inside the IRR improves the flow estimation accuracy, with upto a reduction of 18.5 % of the EPE on the Sintel Clean benchmark.

**Transformer Weight-sharing:** We compare the nature of our cost volume updates using the transformer architecture with both weight-tied and distinct transformer blocks. From Table 2b, it is evident that weight-sharing consistently improves the cross-domain generalization on all the benchmarks while maintaining the number of parameters of the transformer architecture to nearly half of the one without weight sharing.

**Feature enhancement relative to IRR:** Here, we test the effectiveness of contextualizing feature maps concurrent to flow refinement iterations as compared to before IRR. Model parameters, computational overhead and memory requirements remains exactly the same in both the scenarios. However, the cost volumes are dynamic inside IRR in the former setting, whereas the update operator decodes the same cost volume over the refinement iterations in the latter setting. We observe an EPE reduction of 7 % for the non-occluded pixels and a reduction of 5.7 % for all the pixels of the KITTI dataset (Table 2c).

**Matching probability based Lookup** Transforming the matching costs into their corresponding matching distributions stabilizes the training of the weight-tied update operator of IRR. From the performance point of view, we see a reduction in the EPE of 16.7 % and 12.4 % on the Clean and Final passes of the Sintel dataset, whereas the error reduction for all the pixels in the KITTI dataset is 10.2 % (Table 2d).

## 5 Conclusion

We have demonstrated a new transformer based architecture for Optical flow with dynamic cost volumes, which delivers optical flows with near state-of-the-art cross-domain generalization, but at the same time maintaining very good scalability of the overall model. Having a deep transformer architecture inside the refinement stage, however presents opportunities for future works to use the sparse residual flows to further design efficient self and cross attention layers for visual feature enhancement.

**Acknowledgments:** This research has been partially funded by the German BMBF project SocialWear (01IW20002) and EU project CORTEX2 (Grant Agreement: Nr 101070192).

## References

- [Butler et al., 2012] Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer.
- [Dosovitskiy et al., 2015] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766.
- [Geiger et al., 2013] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.
- [Horn and Schunck, 1981] Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.
- [Huang et al., 2022] Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K. C., Qin, H., Dai, J., and Li, H. (2022). Flowformer: A transformer architecture for optical flow. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 668–685. Springer.
- [Mayer et al., 2016] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048.
- [Sun et al., 2018] Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. (2018). Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943.
- [Teed and Deng, 2020] Teed, Z. and Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- [Xu et al., 2022] Xu, H., Zhang, J., Cai, J., Rezatofighi, H., and Tao, D. (2022). Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130.
- [Zhao et al., 2022] Zhao, S., Zhao, L., Zhang, Z., Zhou, E., and Metaxas, D. (2022). Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601.

# DeepSky dataset: A new benchmark for ground-based cloud classification using all-sky images

Dimitrios Tsourounis<sup>1</sup>, Dimitris Kastaniotis<sup>1</sup>, Panagiotis Tzoumanikas<sup>2</sup>, George Andrianakos<sup>3</sup>, Orestis Panagopoulos<sup>2</sup>, Andreas Kazantzidis<sup>2</sup>, Christos Theocharatos<sup>3</sup>, and George Economou<sup>1</sup>

<sup>1</sup> Electronics Laboratory, Physics Department, University of Patras, Rio Patras, 26504, Greece,

<sup>2</sup> Laboratory of Atmospheric Physics, Physics Department, University of Patras, Rio Patras, 26504 Greece,

<sup>3</sup> Irida Labs, Patras InnoHub – Kastritsiou 4, Magoula Patras, 26504, Greece

## Abstract

Cloud observation methods have attracted growing interest due to their significant role in climate prediction and weather forecasting applications since clouds serve as important indicators of weather conditions. Ground-based cloud image classification, which involves categorizing all-sky images into predefined classes, is a key approach to address this challenge. Deep learning methods have demonstrated their effectiveness in various computer vision tasks; however, their performance is inherently tied to the size and quality of the dataset employed for training and evaluation. Many ground-based all-sky image datasets are available, each fulfilling various criteria, including the number of samples, the duration of data collection, the interval between collected samples, the geographical location of data collection, the resolution of images, the types of sky images, and the annotation process used to assign labels to the images. In this paper, we introduce the DeepSky dataset, a novel ground-based all-sky image classification dataset captured over a span of two years. The dataset comprises over seven thousand images and includes seven cloud categories, making it one of the largest datasets in the field. Additional to the dataset, we present an effective cloud classification model based on SWIN Transformers, which achieves state-of-the-art accuracy when evaluated to the other large all-sky image dataset. We are confident that the new dataset, the analysis on cross-dataset evaluation, and the transformer-based proposed method will provide researchers with new opportunities and challenges to explore and expand upon in their investigations. The dataset and supplementary code materials can be accessed from: <https://github.com/dimkastan/DeepSky-classification-dataset>.

**Keywords:** ground-based cloud observation, all-sky images, cloud classification, deep learning

## 1 Introduction

Clouds observation plays a crucial role in numerous applications, including weather forecasting, climate research as well as environmental monitoring. Cloud observations can be classified into space-based remote sensing, where satellites capture images of the Earth's atmosphere from above. Additionally, ground-based observations can be obtained using various sky imaging techniques.

These techniques produce different types of sky images, including sky patch images (SPIs), total sky images (TSIs), and all-sky images (ASIs). The SPIs are captured using wide-angle lenses, resulting in images that cover a portion of the sky instead of the entire sky hole. On the other hand, TSIs are obtained by utilizing a hemispherical chrome-plated mirror to reflect the sky onto a camera positioned above the mirror. However, these images often have low resolution and are hindered by a black sun-blocking shadow band and camera supporting arm, limiting the view of the entire sky. ASIs are typically acquired using a camera equipped with a fish-eye lens, enabling a detailed view

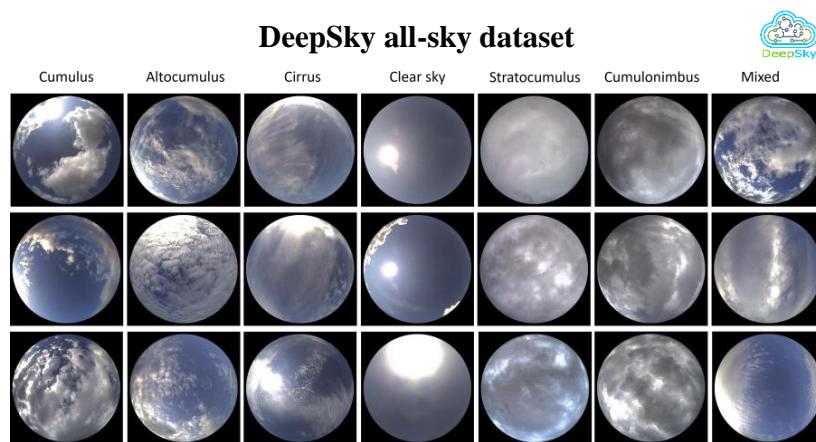


Figure 1: Examples of images from the DeepSky dataset, showcasing the seven different cloud categories.

of the whole sky dome. Ultimately, each type of sky image offers unique features and insights into cloud formations, catering to specific user requirements.

Labeling sky images with relevant information is a crucial aspect that enhances their usability across various applications. For example, incorporating irradiance and photovoltaic power generation data enables the study of clouds impact on solar energy production while segmentation maps, which delineate cloud boundaries, are useful for analyzing cloud morphology and tracking cloud movements. Cloud category labels provide insights into cloud dynamics and their impact on weather patterns. Some of these labeled data are derived from additional measurement instruments, while others require human annotation. Nonetheless, their utilization facilitates researchers to explore a wide range of research questions and develop more comprehensive models for cloud observation.

Over the past decade, significant progress has been made in developing data-driven algorithms for the automated analysis of cloud formations in the sky (Taravat et al., 2015). Although the continuous progress of learning models through new proposed topologies and methods, ranging from hand-crafted techniques (Cheng & Yu, 2015; Dev et al., 2015; Gan et al., 2017; Q. Li et al., 2016; Oikonomou et al., 2019; Xiao et al., 2016) to Convolutional Neural Networks (CNNs) (M. Li et al., 2019; S. Liu, Li, Zhang, Xiao, et al., 2020; Shi et al., 2017; Tsourounis et al., 2022; Ye et al., 2017; Zhang et al., 2018; Zhu et al., 2022), Graph CNNs (S. Liu et al., 2022; S. Liu, Li, Zhang, Cao, et al., 2020a), and Transformers (X. Li et al., 2022), their success heavily relies on the availability of high-quality and diverse datasets for training and evaluation. Cloud formation is influenced by multiple natural factors, resulting in regional and seasonal variations, making localization criteria important. The geographic location and the collection period also impact the images captured by cameras due to local environmental conditions. Additionally, the camera specifications and settings introduce additional challenges. Therefore, the creation of new all-sky datasets containing thousands of images helps capturing a wider range of characteristics, further enhancing the development of cloud analysis methods. Furthermore, the presence of multiple cloud image datasets facilitates the evaluation of feature transferability across various cloud classification tasks. It enables researchers to analyze how well features learned from one dataset and generalize to another, providing insights into the models' adaptability and transfer learning capabilities. Moreover, the existence of diverse datasets enables sophisticated error analysis, allowing for in-depth exploration and understanding of the limitations and challenges faced by current algorithms. This comprehensive analysis helps in driving advancements in the field by identifying areas for improvement and inspiring the development of innovative solutions.

To the best of our knowledge, the largest publicly available ground-based all-sky image dataset with cloud category labels is the Ground-based Remote Sensing Cloud Database (TJNU-GRSCD) (S. Liu, Li, Zhang, Cao, et al., 2020b). This dataset collected in Tianjin by the Meteorological Observation Centre of the China Meteorological Administration for the task of cloud image classification. It contains all-sky colored images captured using a fisheye lens, with a resolution of  $1024 \times 1024$  pixels in color JPEG format. The dataset spans from 2017 to 2018 and consists of 8000 images, with 4000 for training and 4000 for testing. The images are divided into seven sky types according to the presented clouds: 1) cumulus, 2) altocumulus and cirrocumulus, 3) cirrus and cirrostratus, 4) clear sky, 5) stratocumulus, stratus, and altostratus, 6) cumulonimbus and nimbostratus, 7) mixed cloudiness, following the international cloud classification system criteria published in the World Meteorological Organization (WMO) as well as the visual similarity of cloud in practice. The distinguishing features of the GRSCD include its significantly larger number of all-sky images and the provision of separate training and test sets compared to other available datasets (Nie et al., 2022).

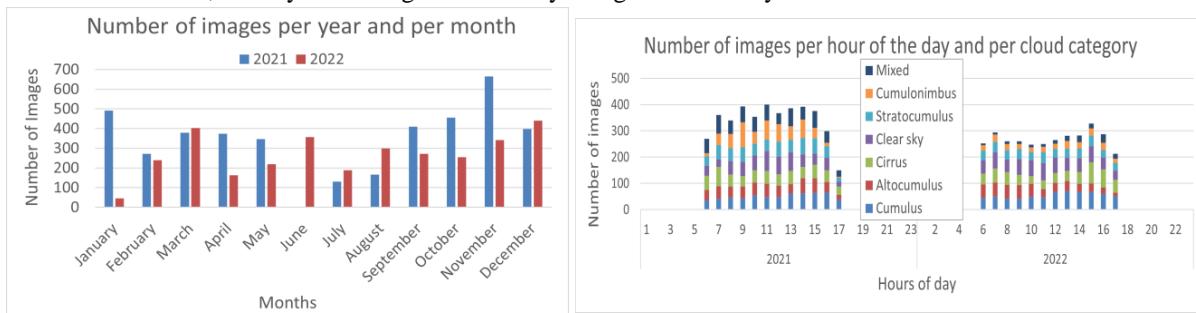
This work introduces a new dataset specifically curated for cloud classification using all-sky images. The dataset covers a wide range of weather conditions, cloud types, and atmospheric phenomena, spanning a two-year period. Our main contribution lies in creating a benchmark dataset to facilitate robust and accurate cloud classification models that generalize well in real-world scenarios. State-of-the-art (sota) deep learning models, including the ResNet-50 CNN and Swin Transformer, are evaluated to establish baseline performance on the new dataset. The presented dataset is categorized among the two largest all-sky image datasets with cloud category labels, sharing an equivalent number of images. Also, the two datasets share the same cloud categories and thus, a comparative analysis could be conducted between them. Finally, the proposed transformer-based approach achieves state-of-the-art performance, surpassing previous best results on the GRSCD.

## 2 Proposed Method

The contribution of the proposed method is two-fold. Firstly, a new dataset is introduced, facilitating the evaluation of different approaches using various evaluation criteria, including cross-dataset generalization using existing datasets. Secondly, a voting scheme is presented, which improves the models' performance by 1%, simply applying augmentations on the test images and utilizing majority voting to determine the final classification prediction.

### 2.1 DeepSky dataset

The DeepSky dataset collected using a Mobotix Q26B-6D016 hemispheric camera with a fisheye lens B016 (focal length 1.6 mm, f/2.0,  $180^\circ \times 180^\circ$ ), featuring a 1/1.8"CMOS, 6MP ( $3072 \times 2048$ ), Progressive Scan on  $1024 \times 768$  resolution, and stored in 24-bit color JPEG format along. A Circle of Interest (COI) with a fixed radius set at 324 pixels is defined to ensure that the analysis focuses solely on relevant cloud information. Then, the images are cropped to the final size of  $678 \times 678$  pixels, removing the black space of COI surrounding the image. Some image examples are shown on Figure 1. The distribution of samples across the recording time is presented in the following Figure 2. The system was captured samples only during daytime to avoid recording samples to ensure visibility of cloudiness in the images. The images were annotated by professional meteorologists to assign the dominant cloud category in each image. Although the system acquires one image per 11 seconds, we selected approximately one image per 30 minutes to form the dataset. Considering that the recordings span a period of two years, we found prominent to split the training and test data in a yearly basis. Towards this goal, the recordings from 2021 were utilized for training and validation, while the recordings from 2022 were reserved for testing. For the training and validation split, researchers have the flexibility to employ various cross-validation approaches; however, we recommend employing the Leave N samples out method, with N set to 10% of the training data across all categories. The test set represents a projection into future events, encompassing potential natural deviations that may not be evident within short time intervals. By doing so, we aim to provide a more accurate reflection of the true distribution of cloud formations, thereby enhancing the reliability and generalizability of classification models.



**Figure 2: Distributions of images in the DeepSky dataset.**

The variation in the number of images per class can be attributed to the geographical environmental conditions of the camera's operation, as different cloud categories are closely linked to the climate of camera's location. To ensure a balanced training set, suitable for end-to-end data-driven learning methods, the training set is carefully pre-processed. Cloud categories that are overrepresented, such as the clear sky class is reduced by randomly selecting a subset of images. On the other hand, underrepresented categories, like cumulonimbus and mixed clouds, are enlarged following two approaches respectively for the two classes. The visual characteristics of cumulonimbus images enable us to increase the sampling rate and select images at 15-minute intervals on days when cumulonimbus clouds are present, while still maintaining the diversity of this class. The mixed class comprises images with multiple clouds, making it a combination of some of other cloud categories. To enhance its representation, image blending techniques are applied using training images from different cloud categories within the dataset. A human expert then selects the most photorealistic output images for inclusion in the mixed class. Consequently, the mixed training class is an amalgam of original and synthetic images. Detailed metadata accompanies the original and synthetic images within the dataset to provide comprehensive information about the blended images. A detailed description the dataset is presented on Table 1.

## 2.2 Swin Transformer and TTA

In this study, we employ Vision Transformers, and specifically the Swin Transformers (Z. Liu et al., 2021), for the image classification task. The Swin Transformers are chosen due to their distinctive architecture, combining the advantages of processing both local and global information. Through the utilization of hierarchical shifting windows, these transformers effectively capture local dependencies while concurrently enabling the incorporation of larger receptive fields to capture global spatial context. This adaptability empowers the models to effectively comprehend and model input patterns, ultimately leading to superior generalization performance across tasks and datasets. Given the inherent characteristics of all-sky image classification, which necessitates the capture of long-range spatial context, the transformers emerge as the most suitable approach for this specific task. Furthermore, demonstrate that by leveraging transformers in conjunction with a Test-Time Augmentation (TTA) technique (Shanmugam et al., 2020), we can achieve an additional improvement in classification performance.

When working with all-sky images, it is important to consider their specific limitations when applying augmentations. It is not recommended to use colour augmentations without careful consideration, as they can potentially modify the cloud characteristics and impact the assigned cloud category. Furthermore, cropping images could isolate parts of clouds or even exclude clouds entirely, which may not align with the assigned class label of the image. Hence, it is crucial to employ augmentation techniques that maintain the original cloud category in the images. Taking into consideration these observations, the TTA pipeline follows 12 equally spaced image rotations. Additionally, for each rotated version of the image, a five-crop strategy is implemented, where only 10% of the image boundaries are cropped. The total number of 60 images is passed from the network and the classification decisions are incorporated through majority voting. The query image is then classified into the category that receives the highest number of votes.

Classes - Cloud type	Description	Year	
		2021	2022
1. <i>Cumulus</i>	Puffy clouds	661	610
2. <i>Altocumulus</i>	Altocumulus & Cirrocumulus	464	521
3. <i>Cirrus</i>	Cirrus & Cirrostratus	559	572
4. <i>Clear sky</i>	Cloudless or a very few cloudiness	655	601
5. <i>Stratocumulus</i>	Stratocumulus, Stratus, & Altostratus	431	570
6. <i>Cumulonimbus</i>	Cumulonimbus & Nimbostratus	568	266
7. <i>Mixed</i>	More than two genera of clouds	645	185
Total		3983	3325

Table 1: DeepSky dataset, Number of images per class and per year.

## 3 Experimental Results

Our experimental analysis focuses on three main aspects. First, we evaluate the performance of the SWIN transformers and TTA method on the GRSCD dataset. Second, we present the performance of this method on the new DeepSky dataset. Third, we compare the transferability between the two datasets.

All models were trained using PyTorch on a Linux machine with four NVIDIA GPUs GeForce RTX 2080 with 8GB. The SWIN base transformer was trained with input image size of 224×224 pixels, patch size of 4 pixels and window size of 7 pixels, which achieves the best performance surpassing by almost 2% the ResNet-50 on the GRSCD classification problem. The SWIN transformer was trained with a cosine learning rate scheduler and a starting learning rate of 0.001. The image augmentations were designed to take into consideration the limitations imposed by the annotation and task definition, as described in the previous section, and thus, images were only rotated and color-augmented with color jitter parameters that do not heavily alter the images.

### 3.1 Intra-dataset analysis

During our experiments, we found that the SWIN base Transformer combined with TTA surpasses all previous methods in the GRSCD, as demonstrated in Table 2. This can be attributed to the expressive capabilities of Transformer networks and their inherent ability to learn high-level relationships between different regions of the image. For the task of all-sky image classification, the attention mechanism in Transformers plays a crucial role in preserving local information and effectively associating it with other regions. This is particularly important as different cloud formations can coexist in the image, along with parts of clear sky. Additionally, Table 2 presents baseline results for the newly introduced DeepSky dataset using ResNet-50 and SWIN Transformer.

Figure 3 presents the confusion matrices for the best performing method on both datasets. The mixed cloudiness class poses the greatest challenge, as it is often confused with multiple other categories, as expected. On the DeepSky dataset, the accuracy for the mixed class is relatively low; however, individual cloud categories within the mixed cloudiness are still detected in the images, suggesting that a post-processing mechanism could be beneficial for this particular class, but it beyond the scope of this work. Moreover, the DeepSky dataset shows a higher level of confusion between stratocumulus and cumulonimbus images, given their similar texture structure. In the GRSCD, the cirrus images are primarily confused with the mixed category, consistent with findings in the relevant literature (M. Li et al., 2019).

References	Method	Accuracy (%)
<b>GRSCD</b>		
(S. Liu, Li, Zhang, Xiao, et al., 2020)	SIFT + BoW	66.13
(S. Liu, Li, Zhang, Xiao, et al., 2020; Zhang et al., 2018)	CloudNet	79.92
(S. Liu, Li, Zhang, Xiao, et al., 2020; Shi et al., 2017)	DCAFs	82.67
(M. Li et al., 2019)	ResNet-50 + DGL	85.28
(S. Liu, Li, Zhang, Xiao, et al., 2020)	ResNet-50 + Attentive Net	86.25
(Tsourounis et al., 2022)	Fusion: ResNet-18 & SIFT-ResNet-18	87.22
(S. Liu, Li, Zhang, Cao, et al., 2020b)	TGCNN (with ResNet-50)	89.48
(Zhu et al., 2022)	Fusion: ICN, ResNet-50 & VGG-16	90.08
Ours	ResNet-50 + TTA	89.12
Ours	SWIN transformer	90.25
Ours	SWIN transformer + TTA	<b>91.25</b>
<b>DeepSky dataset</b>		
Ours	ResNet-50 + TTA	76.32
Ours	SWIN transformer + TTA	<b>76.83</b>

Table 2: Classification accuracy using sota methods on GRSCD and DeepSky datasets.

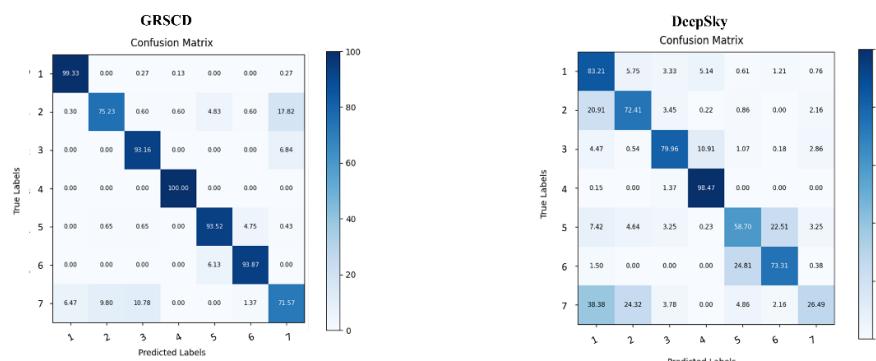


Figure 3: Confusion matrices using SWIN transformer + TTA, evaluated on GRSCD and DeepSky datasets.

### 3.2 Cross-dataset analysis

It is crucial to assess the generalization ability of a model trained on one dataset when applied on the other. To investigate this, we employed the trained models on one dataset and evaluated their performance on a new dataset, as shown in Table 3 and the accompanying confusion matrices in Figure 4. The results reveal substantial differences between the two datasets. Particularly, when training on the GRSCD and evaluating on the DeepSky dataset, a performance drop of about 10% was observed compared to training on the DeepSky dataset. Conversely, when training on the DeepSky dataset and evaluating on the GRSCD dataset, a drop of over 20% was observed. These findings emphasize the disparities between the two datasets and highlight the challenges in generalizing models from one dataset to another, even when they share similar classes. Furthermore, we are investigating the impact of learning the joint distribution of the dataset by merging the training sets of both the GRSCD and DeepSky datasets. Subsequently, we evaluate the model's performance on each individual test set. This approach enables us to assess the extent to which the model possesses the capability to learn from multiple distributions. From the results presented in Table 3, it is evident that the GRSCD can improve the performance on the DeepSky as the joint distribution allows the model to learn some common characteristics across datasets. Also, as shown in the confusion matrices in Figure 4, some characteristics from one dataset are transferred to the other, as for example the performance on the mixed category on GRSCD comparing results from Figure 3 and Figure 4.

Method	Trained on	Evaluated on		Accuracy (%)
		Cross-dataset	Both-datasets	
Swin Transformer + TTA	DeepSky	GRSCD	68.75	
	GRSCD	DeepSky	66.35	
	Both-datasets			
	GRSCD + DeepSky	GRSCD	86.27	
	GRSCD + DeepSky	DeepSky	77.14	

Table 3: Investigation using cross dataset analysis and using jointly both training sets.

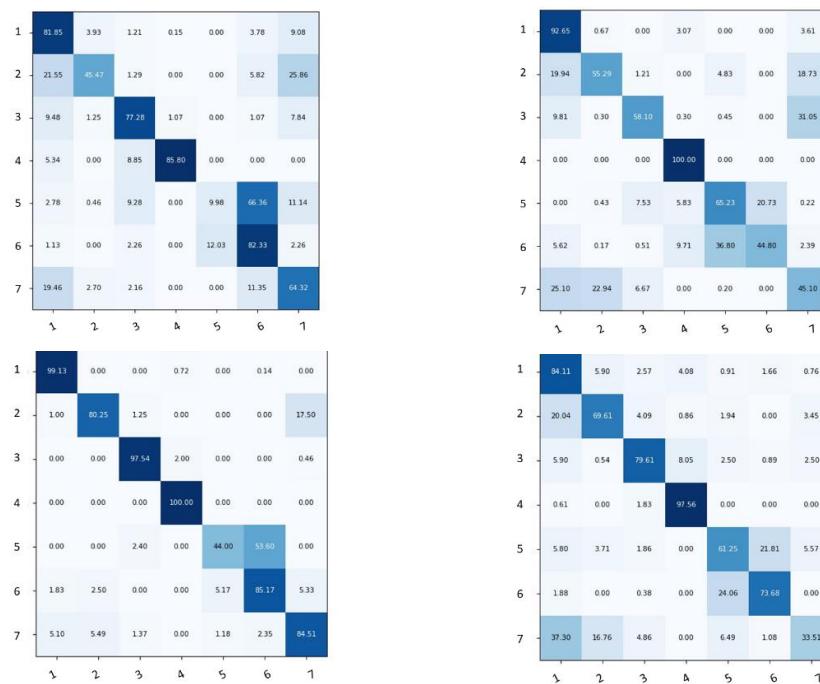
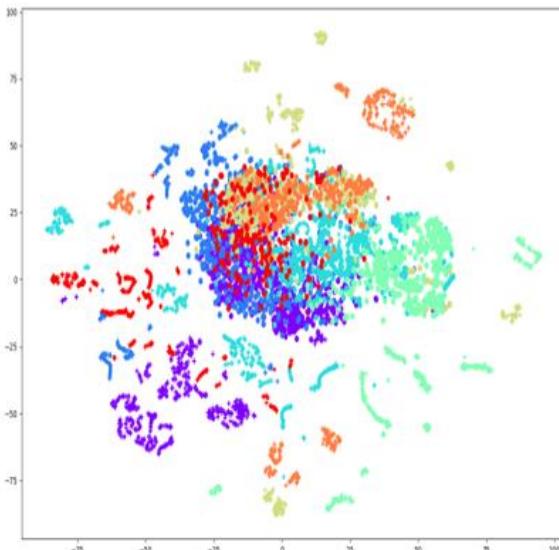


Figure 4: Confusion matrices (True Response vs. Predicted Response) using SWIN transformer + TTA.  
 Upper left: Trained on DeepSky, evaluated on GRSCD. Upper right: Trained on GRSCD, evaluated on DeepSky.  
 Down left: Trained on both, evaluated on GRSCD. Down right: Trained on both, evaluated on DeepSky.

By employing a feature extraction technique using a ResNet-50 model pretrained on ImageNet, we generated biplots that allow for visual analysis of the data distribution. The penultimate fully connected layer output served as the feature vector, and t-SNE was utilized to project these vectors into a 2-dimensional space. Through this approach, we can examine the distribution of the newly introduced DeepSky dataset and its relationship with the GRSCD. The visualization of Figure 5 clearly reveals that the DeepSky dataset exhibits a more uniform distribution, in contrast to the GRSCD, which shows distinct trajectories. In this manner, the DeepSky dataset features a more spherical distribution as compared to the GRSCD, where specific clusters are appeared both on its training and test sets.



**Figure 5: Feature vector embedding from GRSCD (crosses) and DeepSky dataset (circles) in 2D space.**

## 4 Conclusions

In this work we focused on the problem of all-sky image classification and proposed a straightforward yet powerful approach utilizing Transformers and Test Time Augmentations. Furthermore, we introduced a novel dataset consisting of recordings spanning two years. This dataset allows for comprehensive evaluations of model generalization and learning capabilities across datasets, particularly considering that the DeepSky dataset shares cloud types with the GRSCD. Notably, the DeepSky and GRSCD datasets are currently the largest existing all-sky datasets with cloud category labels. Future plans include the extension of the DeepSky dataset using both the images of next year and additional environmental multimodal information. Additionally, we aim to explore the utilization of image sequences to train Transformer-based architectures, further enhancing the performance of our models.

## Acknowledgements

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK – 00681, MIS 5067617).

## References

- Cheng, H.-Y., & Yu, C.-C. (2015). Block-based cloud classification with statistical features and distribution of local texture features. *Atmospheric Measurement Techniques*, 8(3), 1173–1182. <https://doi.org/10.5194/amt-8-1173-2015>
- Dev, S., Lee, Y. H., & Winkler, S. (2015). Categorization of cloud image patches using an improved texton-based approach. *2015 IEEE International Conference on Image Processing (ICIP)*, 422–426. <https://doi.org/10.1109/ICIP.2015.7350833>
- Gan, J., Lu, W., Li, Q., Zhang, Z., Yang, J., Ma, Y., & Yao, W. (2017). Cloud Type Classification of Total-Sky Images Using Duplex Norm-Bounded Sparse Coding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(7), 3360–3372. <https://doi.org/10.1109/JSTARS.2017.2669206>
- Li, M., Liu, S., & Zhang, Z. (2019). Dual Guided Loss for Ground-Based Cloud Classification in Weather Station

- Networks. *IEEE Access*, 7, 63081–63088. <https://doi.org/10.1109/ACCESS.2019.2916905>
- Li, Q., Zhang, Z., Lu, W., Yang, J., Ma, Y., & Yao, W. (2016). From pixels to patches: A cloud classification method based on a bag of micro-structures. *Atmospheric Measurement Techniques*, 9(2), 753–764. <https://doi.org/10.5194/amt-9-753-2016>
- Li, X., Qiu, B., Cao, G., Wu, C., & Zhang, L. (2022). A Novel Method for Ground-Based Cloud Image Classification Using Transformer. *Remote Sensing*, 14(16), Article 16. <https://doi.org/10.3390/rs14163978>
- Liu, S., Duan, L., Zhang, Z., Cao, X., & Durrani, T. S. (2022). Ground-Based Remote Sensing Cloud Classification via Context Graph Attention Network. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11. <https://doi.org/10.1109/TGRS.2021.3063255>
- Liu, S., Li, M., Zhang, Z., Cao, X., & Durrani, T. S. (2020b). Ground-Based Cloud Classification Using Task-Based Graph Convolutional Network. *Geophysical Research Letters*, 47(5). <https://doi.org/10.1029/2020GL087338>
- Liu, S., Li, M., Zhang, Z., Xiao, B., & Durrani, T. S. (2020). Multi-Evidence and Multi-Modal Fusion Network for Ground-Based Cloud Recognition. *Remote Sensing*, 12(3), Article 3. <https://doi.org/10.3390/rs12030464>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. <https://doi.org/10.1109/ICCV48922.2021.00986>
- Nie, Y., Li, X., Paletta, Q., Aragon, M., Scott, A., & Brandt, A. (2022). *Open-Source Ground-based Sky Image Datasets for Very Short-term Solar Forecasting, Cloud Analysis and Modeling: A Comprehensive Survey* (arXiv:2211.14709). arXiv. <https://doi.org/10.48550/arXiv.2211.14709>
- Oikonomou, S., Kazantzidis, A., Economou, G., & Fotopoulos, S. (2019). A local binary pattern classification approach for cloud types derived from all-sky imagers. *International Journal of Remote Sensing*, 40(7), 2667–2682. <https://doi.org/10.1080/01431161.2018.1530807>
- Shanmugam, D., Blalock, D. W., Balakrishnan, G., & Guttag, J. V. (2020). Better Aggregation in Test-Time Augmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1194–1203.
- Shi, C., Wang, C., Wang, Y., & Xiao, B. (2017). Deep Convolutional Activations-Based Features for Ground-Based Cloud Classification. *IEEE Geoscience and Remote Sensing Letters*, 14(6), 816–820. <https://doi.org/10.1109/LGRS.2017.2681658>
- Taravat, A., Del Frate, F., Cornaro, C., & Vergari, S. (2015). Neural Networks and Support Vector Machine Algorithms for Automatic Cloud Classification of Whole-Sky Ground-Based Images. *IEEE Geoscience and Remote Sensing Letters*, 12(3), 666–670. <https://doi.org/10.1109/LGRS.2014.2356616>
- Tsourounis, D., Kastaniotis, D., Theoharatos, C., Kazantzidis, A., & Economou, G. (2022). SIFT-CNN: When Convolutional Neural Networks Meet Dense SIFT Descriptors for Image and Sequence Classification. *Journal of Imaging*, 8(10), Article 10. <https://doi.org/10.3390/jimaging8100256>
- Xiao, Y., Cao, Z., Zhuo, W., Ye, L., & Zhu, L. (2016). mCLOUD: A Multiview Visual Feature Extraction Mechanism for Ground-Based Cloud Image Categorization. *Journal of Atmospheric and Oceanic Technology*, 33(4), 789–801. <https://doi.org/10.1175/JTECH-D-15-0015.1>
- Ye, L., Cao, Z., & Xiao, Y. (2017). DeepCloud: Ground-Based Cloud Image Categorization Using Deep Convolutional Features. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10), 5729–5740. <https://doi.org/10.1109/TGRS.2017.2712809>
- Zhang, J., Liu, P., Zhang, F., & Song, Q. (2018). CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network. *Geophysical Research Letters*, 45(16), 8665–8672. <https://doi.org/10.1029/2018GL077787>
- Zhu, W., Chen, T., Hou, B., Bian, C., Yu, A., Chen, L., Tang, M., & Zhu, Y. (2022). Classification of Ground-Based Cloud Images by Improved Combined Convolutional Network. *Applied Sciences*, 12(3), Article 3. <https://doi.org/10.3390/app12031570>

# Haptic Gloves (SM-EXO) for Multi-Users in Pick-and-Place Collaborative Robot Simulated Environment

Rupal Srivastava, Eber Lawrence Souza Gouveia, Niall Murray, Declan Devine

*SFI Confirm Smart Manufacturing, TUS Athlone, Ireland*

## Abstract

The use of collaborative pick-and-place robots has been gaining popularity in recent years, given their ability to work alongside humans in various industrial and manufacturing environments. The use of smart wearables as the Industrial Internet of Wearable Things (IIoWT), to facilitate efficiency of such environments is receiving vast attention. In this article, we present the development of a novel smart glove (SM-EXO) integrated with the RobotStudio simulation software for use in a multi-user pick-and-place collaborative robot-simulated environment. The proposed smart glove uses an intuitive force-feedback concept to provide users with a sense of constraint while moving their fingers, thus giving them a more organic feeling. A series of experiments were conducted in the simulated environment pairing the SM-EXO with several robotic arms to evaluate the feasibility and repeatability of the concept. Our results show that the SM-EXO effectively improves human control over the robotic arm in the simulated environment. These results can be directly translated into the real environment as well as AR/VR environments for training purposes. To facilitate intelligent learning and prediction of human intentions, we propose a finger-based gesture understanding approach using Shape Memory Alloy wires and displacement sensors embedded in the smart glove. This study provides a foundation for further research in Industry 4.0, with future directions and applications domains discussed with a focus on two specific use cases. The first use case is worker safety and convenience in hazardous environments in Industry 4.0. The second is based on rehabilitation using SM-EXO controlled simulated set of tasks.

**Keywords:** Shape Memory Alloy, Multi-User Collaboration, Smart Gloves, Human-Computer Interaction, AR-VR Technology

## 1 Introduction

The term "Industry 4.0," also known as the fourth industrial revolution, originated from the high-tech strategy of the German federal government in 2011. Industry 4.0 aims to utilize the internet, digital technologies, and quantum sciences to advance autonomous and intelligent cyber-physical systems. As Industry 4.0 continues to develop, it can be defined as the integration of cyber-physical systems, cloud technology, the Internet of Things, and the Internet of Services, interacting with humans in real-time to maximize value creation. By combining the physical and virtual worlds, interoperability, advanced artificial intelligence, and autonomy will become integral parts of this new industrial era [Hwang, 2016]. Integration of smart wearables as a part of the Industrial Internet of Wearable Things (IIoWT) has further revolutionised the way we perceive modern manufacturing. The major focus of these wearables is on worker safety and data collation.

Smart wearables have become an increasingly popular technology in many industries due to their ability to enhance efficiency, safety, and productivity. In manufacturing, smart wearables such as smart glasses and headsets can provide workers with hands-free access to real-time data, allowing them to complete tasks more efficiently and with greater accuracy. In healthcare, wearables can track vital signs and provide early detection of potential health issues for patients, as well as provide health professionals with remote access to patient data. Smart wearables are also being used in logistics and warehousing, where they can provide workers with

real-time inventory information, route optimization, and enhanced safety features. With the ability to provide a wealth of information in real-time, smart wearables are poised to revolutionize many industries, providing workers with valuable insights and enabling them to make better-informed decisions [Srivastava et al., 2022a].

Industrial wearables can be categorized into four major groups, namely Monitoring, Supporting, Training, and Tracking. The Monitoring category is not limited to analyzing the vital signs of workers but also includes monitoring the workplace's environmental conditions. Supporting wearables not only augments the physical abilities of workers through exoskeletons but also enables communication among them. Wearables that monitor the accuracy of a worker's actions and provide detailed reports, such as those utilizing the biomechanical analysis to determine proper posture, fall under the Training category. Lastly, Tracking wearables are responsible for keeping tabs on the location of workers and machinery, providing a comprehensive overview of the production line for safer worker-machine interactions [Svertoka et al., 2021].

Simulated environments play a crucial role when working with industrial robots. These virtual environments offer a valuable testing ground for the system under development, allowing workers to assess its performance and functionality before deploying it in a physical cell. By simulating real-world conditions and scenarios, such as collision detection, path planning, and task execution, engineers can identify potential issues, optimize robot behaviour, and validate the system's overall design. Simulated environments not only save time and resources but also contribute to a safer working environment for the initial testing phase, as they eliminate the risk of accidents or damage to physical equipment [Oyekan et al., 2019]. Additionally, simulated environments enable the reuse of code, making it easier to transfer and apply algorithms and programs from the virtual environment to the physical robot, thereby streamlining the development process and enhancing efficiency [Al-Ahmari et al., 2016]. RobotStudio, the official platform for working with ABB industrial robots, offers a wide range of tools to develop and test projects in both the physical and digital layers of the system. Among these tools, RobotStudio includes its native programming language (RAPID), which is capable of reading and writing data from/to other programming languages using TCP/IP protocols. This feature allows for the development of more complex applications beyond the RAPID language, where only the final results, such as a robot target, need to be sent.

As human-robot collaboration becomes more commonplace, there is an increasing demand for safe, efficient, and cost-effective solutions. One such solution is the development of smart gloves, which are equipped with sensors that can accurately capture and analyze data on the wearer's movements. This technology holds great potential for improving workplace safety, productivity, and overall efficiency. However, effective collaboration between humans and robots through smart gloves still requires significant improvements. This study presents an innovative approach to real-time human-robot interaction in collaborative tasks through the development of an Arduino-embedded and ROS-compatible smart glove system. The approach is implemented using RAPID programming in RobotStudio, with three baseline gestures from thumb, index, and middle finger, used alongside appended labels and corresponding sensor data. Real-world validation of the proposed system and approach shows potential applications in Industry 4.0, healthcare, and daily assistance for rehabilitation. The future applications and research scope in this field are vast, especially using the same concept for user training purposes in AR and VR environments, which will be the further directions in this research.

## 2 Shape Memory Alloy-based Exoskeleton (SM-EXO)

The selection of material was a critical aspect of designing the glove. To ensure that the glove could be used for extended periods, particularly by children, it was essential to select a soft material, preferably fabric, suitable for everyday use. Therefore, the prototype of the glove is designed with an external attachment of SMA wire and sensors (SM-EXO) that can be attached to any glove. This eliminates the need to design multiple gloves based on the size and shape of the hand. The external mechanism fits the fingers using ring-like 3D printed attachments, serving as guides for the SMA wires and attachment mechanisms for any glove. The ring's size can be adjusted based on the finger's diameter. The SMA wires pass through these rings and are fixed at one end, while linear position sensors are attached to the other end. The position sensors are fixed to the glove using industrial Velcro tapes, allowing the sensor's location to be changed based on the subject's hand size and

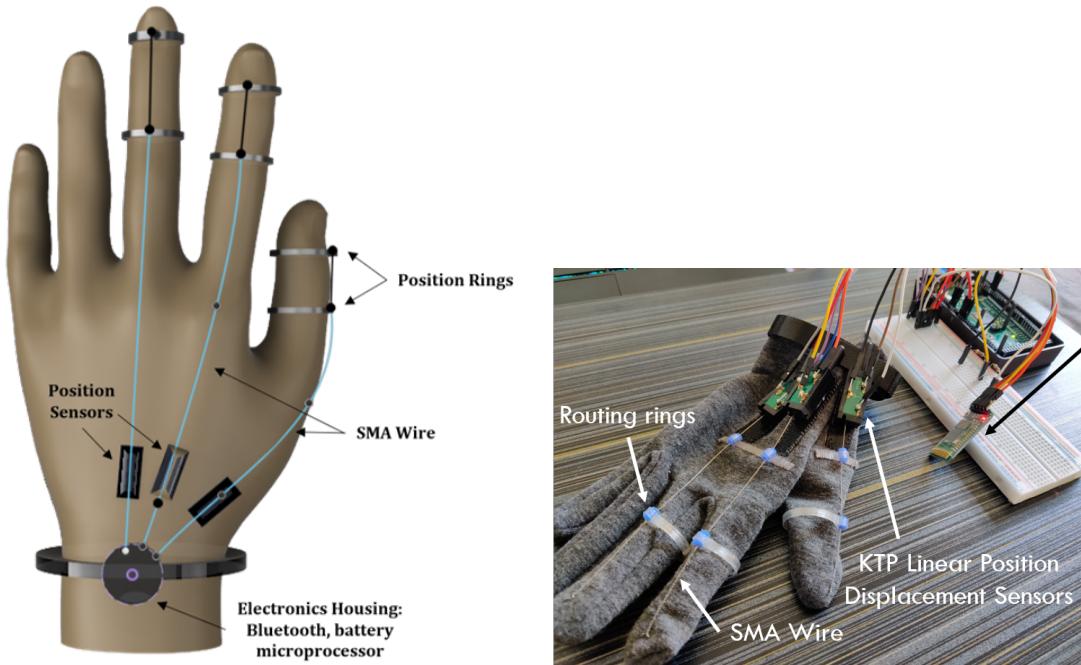


Figure 1: Design and mechanism of the Smart Glove and Proof-of-concept. (a) Computer-Aided Design of the SM-EXO, Position Rings, and Tubing for the SMA routing [Srivastava et al., 2022b]. (b) Detailed view of the first SM-EXO prototype [Srivastava et al., 2022c].

dimension. Finally, the glove's base is attached with a Velcro strip, which controls the glove's diameter and enables it to be worn by subjects of any age, size, or gender. Figure 2 depicts the Computer-Aided Design of the glove made in Autodesk Fusion 360, and the complete design of the glove is shown in Figure 2, with wire routing and Velcro strips. The user-centric design of the glove is unique as it offers variable geometry. As the SM-EXO attachment on the glove is easily detachable, users can further customize the glove design based on their preferences and weather conditions. Therefore, the proposed modular SM-EXO design is not only innovative as a smart glove concept, but its customizable potential sets it apart from any state-of-the-art glove. The advantages of the proposed design over available gloves are discussed in detail in the Conclusions and Table 1.

## 2.1 Design and Mechanism

Before finalizing the SMA wire routing method for the glove, various designs were proposed and discussed. One idea involved embedding the wires within a custom-designed glove, while another was to pass the wire through tubing attached externally to the glove from the tip of the finger to the base of the finger. Ultimately, the current design was chosen, which involves passing the wire through 3D printed self-adjustable rings. Although the embedded wire-sensor design is being considered for advanced phases of the concept, the proposed prototype design is deemed a seamless, lightweight, and hassle-free alternative to the tubing design. It significantly reduces the complexity, material requirement, and weight of the glove. Additionally, the rings' variable diameter ensures the firm grip of the SM-EXO attachment to the glove, acting as an SMA wire guide between the fixed and potentiometer-connected ends. The rings are designed with holes corresponding to the 0.8 mm diameter of the NiTiNOL wire used, securing and allowing free movement of the wire through them. Figure 1 illustrates the fixed end of the wire and the ring with tubing attachment for SMA wire passing and movement.

The KTP Linear Position Sensor with spring feedback is used to attach the SMA wires. These sensors measure the displacement of the wires when the fingers move and generate a voltage response. The open hand

position is considered as the reference point for this measurement, and the bending of the fingers results in a series of voltage responses that vary depending on the angle of bending. The KTP linear position sensors employed in this study exhibit a high degree of repeatability, with a tolerance range of +/-0.0127, and are engineered for use in joystick or robotics applications with high shock and vibration resistance. Moreover, the sensor features an integral slider, spring return design, infinite high resolution, and an operating life of 2 million cycles. The position sensors capture the displacement data of the SMA wire as the fingers move and transmit it to a computer for storage and processing. The processed data is then utilized as actuation signals for the simulated environment. The authors have discussed the working of the glove and the proof-of-concept in more detail in their previous works [Srivastava et al., 2022b], [Srivastava et al., 2022c].

### 3 RobotStudio Simulated Environment

RobotStudio is a powerful software tool developed by ABB Robotics for designing, simulating, and programming industrial robots. It provides a virtual environment for testing and optimizing robot programs before they are implemented in real-world applications. RobotStudio includes features such as robot calibration, collision detection, and path planning, making it an essential tool for robot system integration and programming. It supports a wide range of robot models and applications, and its user-friendly interface makes it accessible to both novice and expert users. It is a powerful tool used by robotics engineers, automation experts, and manufacturing professionals to create, simulate, and optimize robot systems. With RobotStudio, users can create a virtual model of a robot and its surroundings, allowing them to test the robot's movements and functions before deploying it on the shop floor. The software also allows for offline programming (RAPID), where users can program the robot without needing to connect it to the actual hardware, saving time and reducing downtime. Additionally, RobotStudio supports a wide range of robot models and configurations, making it a versatile tool for various industrial applications [Connolly, 2009].

RAPID is a high-level programming language that is used to control industrial robots, particularly those manufactured by ABB. It was developed by ABB Robotics in the 1990s to provide a powerful and flexible programming interface for their robots. RAPID stands for "Robot Application Programming Interface for Developers," and it is designed to be easy to use, even for those with little or no programming experience. RAPID is a structured programming language that uses a combination of keywords, variables, and functions to define the robot's movements and actions. It allows programmers to control the robot's movements in three-dimensional space, as well as to interact with sensors and other external devices. RAPID programs can be written offline and then uploaded to the robot's controller, or they can be created and edited directly on the robot itself. One of the key advantages of RAPID is its flexibility. It allows programmers to create complex programs that can be easily modified and adapted to changing production needs. It also provides a wide range of built-in functions for common robotic tasks, such as path planning, collision detection, and error handling. This makes it easier and faster to develop new robotic applications, even for those without extensive programming experience [Holubek et al., 2014].

#### 3.1 Integrating SM-EXO

The smart glove is able to send data from the microcontroller to the computer via Bluetooth, WiFi, or Serial Connection. For the sake of this article, we connected the glove to the computer using a serial connection. The voltage signal generated from the bending of the fingers is received and identified using Python and is sent to the RobotStudio simulation environment using the RAPID programming language. We define the voltage limits from the position sensor between 0 and 1. Hence, 0 corresponds to no voltage, whereas 1 represents the highest voltage which is 5V. This allows us to define the multiple combinations of the movement of the x-y-z axis of the end effector on the robotic arm. Table 1 shows the truth table for the axis movements and the corresponding fingers that are causing the movement. We have, furthermore, defined the coordinates of the movement of the end effector between (0,0,0) and (1,1,1) locations on the x, y, and z axis respectively.

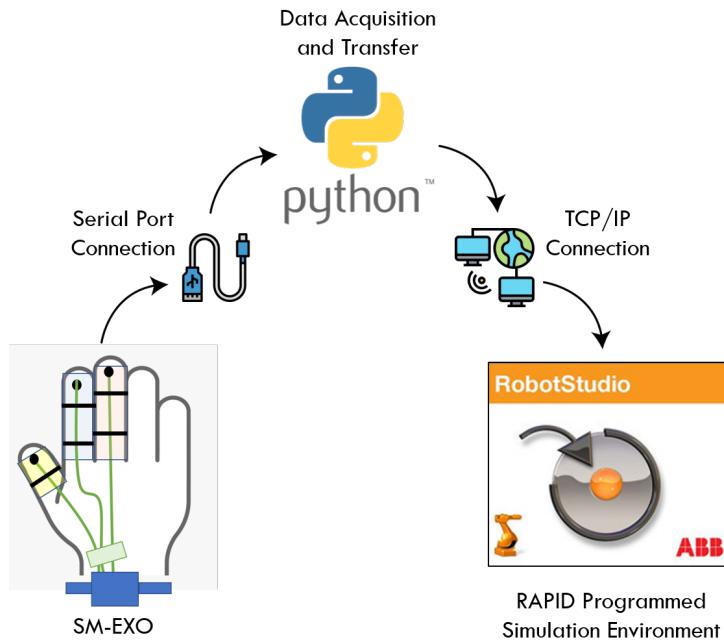


Figure 2: The flow architecture of the Cyber-Physical system in the glove-robot simulated environment.

Table 1: Truth Table for the movement configurations of the end-effector based on the finger movement.

X	Y	Z	Output (Movement)	Finger
0	0	0	0	None
0	0	1	1	Middle
0	1	0	1	Index
0	1	1	1	Index and Middle
1	0	0	1	Thumb
1	0	1	1	Thumb and Middle
1	1	1	1	All three

## 4 Multi-Users Collaborative Concept

Multiuser collaborative robots, also known as cobots, are designed to work alongside humans in a shared workspace. Unlike traditional industrial robots that are usually confined to a fenced-off area, cobots are equipped with advanced sensors and algorithms that allow them to operate safely around humans. One of the key advantages of multiuser collaborative robots is their ability to work in a team with human workers. Cobots can assist humans in tasks that require precision, strength, and speed, thereby reducing the workload and improving productivity. Additionally, cobots can also take on repetitive or hazardous tasks, freeing up human workers to focus on more complex or creative tasks. Another advantage of cobots is their flexibility and ease of use. Many cobots are designed to be programmed by non-experts, using intuitive interfaces or even by physically moving the robot arm to teach it a new task. This makes it easier for workers to adapt to new tasks and reduces the need for specialized programming skills. Overall, multiuser collaborative robots have the potential to transform manufacturing and other industries by enabling a more efficient, safe, and flexible way of working [Rodriguez-Guerra et al., 2021].

## Use Case

Several robots in a manufacturing assembly unit work together in a shared workspace to accomplish different tasks, such as stud welding or spot welding a car body in the automotive industry's body-in-white process. Figure 1 demonstrates how four cooperative robots operate together in an assembly station. Efficient robot operations demand that the robots complete their tasks in the shortest possible time while preventing collisions with each other. Consequently, identifying a collision-free path for multiple robots to finish their allotted tasks is critical. This is a vital factor from the operator's standpoint [Xin et al., 2020].

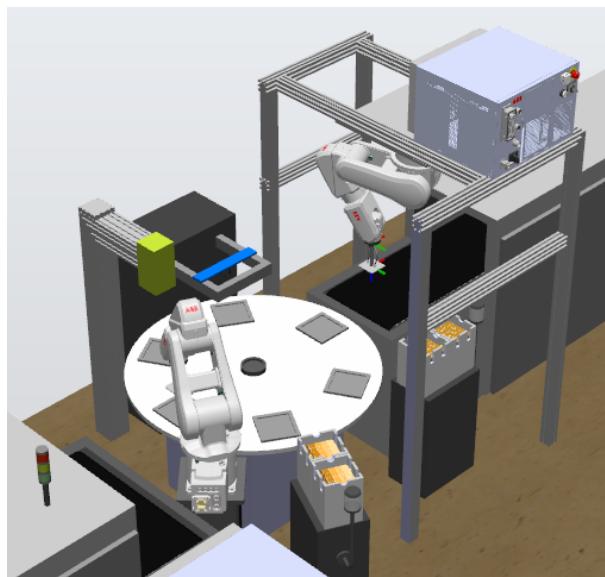


Figure 3: Demo from RobotStudio showing a Multi-Robot collaborative setup

In this Use Case, we discuss integrating two smart gloves, individually for each hand, with the multi-robot collaborative system. While one hand controls the robotic arm handling the fragile/hazardous on one conveyor belt and puts it on another, the other glove controls another arm for another pick-and-place action. The application is limited, however, to robotic grippers with 3-finger grips.

The process begins with the user putting on the control glove and establishing a connection with the robots. In this case, there are two steps of communication, i.e., the first is between the glove and Python script and the second is between Python script and RAPID script. For the first one, it's being used a serial communication for reading data from the glove and send it to the Python script. For the second, it's being used a TCP/IP protocol to establish the connection between the Python and RAPID scripts. After the connection is established, the user has a series of predefined finger movements that allows them to control the robot. As mentioned before, each robot is controlled by one of the gloves, so that all the control gestures are the same for both. The user needs to close and open the fingers twice to enter in the control mode. After entering in the control mode, the user can move their fingers as explained in the section 3.1 to move the robot to the target position. The user can also exit the control mode by closing and opening the fingers twice again. In this way, it is possible move their hands freely when out of the control mode. In this application we focus on the manipulation of fragile/hazardous elements, however, it might be applied in any product manipulation task.

## ACKNOWLEDGEMENT

This research was funded by Marie Skłodowska-Curie grant agreement No. 847577, co-funded by the European Regional Development Fund and Science Foundation Ireland (SFI) under Grant Number SFI/16/RC/3918 (CONFIRM).

## References

- [Al-Ahmari et al., 2016] Al-Ahmari, A. M., Abidi, M. H., Ahmad, A., and Darmoul, S. (2016). Development of a virtual manufacturing assembly simulation system. *Advances in Mechanical Engineering*, 8(3):1687814016639824.
- [Connolly, 2009] Connolly, C. (2009). Technology and applications of abb robotstudio. *Industrial Robot: An International Journal*, 36(6):540–545.
- [Holubek et al., 2014] Holubek, R., Delgado Sobrino, D. R., Košt’ál, P., and Ružarovský, R. (2014). Offline programming of an abb robot using imported cad models in the robotstudio software environment. In *Applied Mechanics and Materials*, volume 693, pages 62–67. Trans Tech Publ.
- [Hwang, 2016] Hwang, J. S. (2016). The fourth industrial revolution (industry 4.0): intelligent manufacturing. *SMT Magazine*, 3:616–630.
- [Oyekan et al., 2019] Oyekan, J. O., Hutabarat, W., Tiwari, A., Grech, R., Aung, M. H., Mariani, M. P., López-Dávalos, L., Ricaud, T., Singh, S., and Dupuis, C. (2019). The effectiveness of virtual environments in developing collaborative strategies between industrial robots and humans. *Robotics and Computer-Integrated Manufacturing*, 55:41–54.
- [Rodriguez-Guerra et al., 2021] Rodriguez-Guerra, D., Sorrosal, G., Cabanes, I., and Calleja, C. (2021). Human-robot interaction review: Challenges and solutions for modern industrial environments. *IEEE Access*, 9:108557–108578.
- [Srivastava et al., 2022a] Srivastava, R., Alsamhi, S. H., Murray, N., and Devine, D. (2022a). Shape memory alloy-based wearables: A review, and conceptual frameworks on hci and hri in industry 4.0. *Sensors*, 22(18).
- [Srivastava et al., 2022b] Srivastava, R., Kuts, V., Souza Gouveia, E. L., Murray, N., Devine, D., and O’Connell, E. (2022b). SMA-Based Haptic Gloves Usage in the Smart Factory Concept: XR Use Case. volume Volume 2B: Advanced Manufacturing of ASME International Mechanical Engineering Congress and Exposition. V02BT02A017.
- [Srivastava et al., 2022c] Srivastava, R., Singh, M., Gomes, G. D., Murray, N., and Devine, D. (2022c). Smexo: Shape memory alloy-based hand exoskeleton for cobotic application. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1277–1284.
- [Svertoka et al., 2021] Svertoka, E., Saafi, S., Rusu-Casandra, A., Burget, R., Marghescu, I., Hosek, J., and Ometov, A. (2021). Wearables for industrial work safety: A survey. *Sensors*, 21(11).
- [Xin et al., 2020] Xin, J., Meng, C., Schulte, F., Peng, J., Liu, Y., and Negenborn, R. R. (2020). A time-space network model for collision-free routing of planar motions in a multirobot station. *IEEE Transactions on Industrial Informatics*, 16(10):6413–6422.

# Do the Frankenstein, or how to achieve better out-of-distribution performance with manifold mixing model soups

Hannes Fassold

*JOANNEUM RESEARCH - DIGITAL, Austria*

## Abstract

The standard recipe applied in transfer learning is to finetune a pretrained model on the task-specific dataset with different hyperparameter settings and pick the model with the highest accuracy on the validation dataset. Unfortunately, this leads to models which do not perform well under distribution shifts, e.g. when the model is given graphical sketches of the object as input instead of photos. In order to address this, we propose the *manifold mixing model soup*, an algorithm which mixes together the latent space manifolds of multiple finetuned models in an optimal way in order to generate a fused model. We show that the fused model gives significantly better out-of-distribution performance (+3.5 % compared to best individual model) when finetuning a CLIP model for image classification. In addition, it provides also better accuracy on the original dataset where the finetuning has been done.

**Keywords:** Latent space manifold, transfer learning, finetuning, distribution shift, image classification

## 1 Introduction

Large pretrained visual foundation models like CLIP [Radford et al., 2021] or CoCa [Yu et al., 2022] got very popular recently due to their great performance for a variety of computer vision tasks, either as zero-shot learner (without finetuning) or serving as a base for task-specific finetuning on a smaller dataset.

Typically, multiple models are finetuned with different hyperparameters (like learning rate, weight decay or data augmentation strategy), using the same pretrained model as initialization. From those, the model with the best accuracy on the validation dataset is selected. Unfortunately, this procedure leaves out important information which has been learned in the latent space manifolds (individual layers or a collection of layers) of the remaining finetuned models. As shown in [Wortsman et al., 2022a], even fusing multiple finetuned models in a very straightforward way by averaging them makes the fused model already significantly more robust to distribution shifts in the data.

Motivated by this, we propose the *manifold mixing model soup* (*ManifoldMixMS*) algorithm. Instead of simple averaging, it uses a more sophisticated strategy to generate the fused model. Specifically, it partitions a neural network model into several latent space manifolds (which can be individual layers or a collection of layers). Afterwards, from the pool of finetuned models available after hyperparameter tuning, the most promising ones are selected and their latent space manifolds are mixed together individually. The optimal mixing coefficient for each latent space manifold is calculated automatically via invoking an optimization algorithm. The fused model we retrieve with this procedure can be thought as sort of a "Frankenstein" model, as it integrates (parts of) individual model components from multiple finetuned models into one model.

The remainder of the work is organized as follows. In section 2 we revise related work. Section 3 presents the proposed manifold mixing model soup algorithm. Section 4 presents the experiments and evaluation, which show the advantage of the proposed algorithm with respect to the state of the art, especially with respect to distribution shifts in the data. Finally, section 5 concludes the paper.

## 2 State of the Art

A variety of methods has been proposed recently for merging several models into one fused model, with the aim of increasing the generalization capability and robustness to distribution shifts of the fused model.

A classical method is *stochastic weight averaging* (SWA) [Izmailov et al., 2018], which produces the fused model by averaging the weights of a set of models sampled from the final stages of a single training run. They show that SWA leads to solutions of the optimization problem that are wider than the optima found by standard SGD, which in turn leads to a better generalization of the fused model.

The authors of [von Oswald et al., 2021] propose to replicate and learn in parallel a subset of weights (e.g. the batch-norm and classification layers) in a late phase of neural network learning. These late-phase weights define an ensemble of models which share every other weight. These parameters are then optimized independently and subsequently averaged.

In [Jolicoeur-Martineau et al., 2023] an algorithm called *population parameter averaging* (PAPA) is presented, which trains a population of models in parallel (with different learning rate, data augmentation strategies etc.). It improves overall performance by infrequently replacing the weights with the population average and frequently pushing all model weights slightly towards the population average. A disadvantage of this method is the high memory consumption, as the gradients for *several* parallel training runs have to be kept up.

The work of [Wortsman et al., 2022a] shows that averaging the weights of multiple models finetuned with different hyperparameter configurations often improves accuracy and out-of-distribution performance of the averaged model. They propose two different averaging algorithms (which they call "souping"), *uniform soup* and *greedy soup*. The uniform soup is a very simple procedure, as it does an averaging of all finetuned models. In contrast, the greedy soup is constructed by sequentially adding each model as a potential ingredient to the soup, and only keeping the model if it improves the performance of the averaged model. Our proposed *manifold maxing model soup* algorithm (see section 3) is inspired by their greedy soup algorithm. But in contrast to them, we are partitioning the model into several components (latent space manifolds) and do an optimization in order to calculate the optimal mixing factor for each component.

In [Matena and Raffel, 2022] it is shown that uniform averaging of several finetuned models corresponds to making an isotropic Gaussian approximation to their posteriors. The authors propose an alternative merging procedure based on the Laplace approximation, where each model's posterior is approximated as a Gaussian distribution whose precision matrix (inverse of the covariance matrix) corresponds to its Fisher information.

The authors of [Wortsman et al., 2022b] found that while fine-tuning a pretrained vision model improves performance on the downstream task, it also tends to decrease accuracy on the original pretraining task. They therefore propose a robust finetuning procedure called *WiSE-FT* that computes a weighted average of the original pretrained parameters and the finetuned parameters. Different weighting values produce different trade-offs between pretraining and finetuning task performance.

## 3 Manifold mixing model soup

In the following, we outline the proposed algorithm for generating a fused model – the *manifold mixing model soup* – from its ingredients (the finetuned models after hyperparameter tuning). The algorithm pseudocode can be seen in Algorithm 1.

We first sort all  $n$  finetuned models  $\theta_i$  (with  $i = 0, \dots, n - 1$ ) in descending order, based on their validation accuracy  $ValAcc(\theta_i)$  on the original dataset which was used for finetuning. So  $\theta_0$  is the model (to be precise, its finetuned parameters) with the highest validation accuracy, whereas  $\theta_{n-1}$  is the one with the lowest validation accuracy.

Each model  $\theta_i$  is partitioned into  $m$  components  $\theta_i^j$ , where  $\theta_i^j$  corresponds to a single latent space manifold, and  $j = 1, \dots, m$ . Each latent space manifold comprises either a single layer or a collection of layers, corresponding to one building block of the model. Typically, we partition a model into 10 – 30 components. A finer partitioning makes the subsequent optimization more difficult, whereas a too coarse partitioning provides not enough freedom to optimize the mixing of the latent space manifolds individually. The motivation for aggregating a collection of

layers into one component is to reduce the number of variables during optimization, which makes it easier for the optimizer to find a good optimum. For each model, of course the same partitioning is employed.

The fused model  $\Psi$  is now calculated in an sequential way, by iteratively mixing promising ingredient models with it. At first, the fused model is set to the best finetuned model via  $\Psi = \theta_0$ , and the variable  $k$ , which counts the number of models which have been mixed so far into the fused model, is set to 1.

In each iteration (for  $i = 1, \dots, n - 1$ ), we try now to mix the candidate model  $\theta_i$  with the current fused model  $\Psi$  in an optimal way, with the aim of increasing the validation accuracy of the updated fused model  $\Psi'$  (which includes  $\theta_i$ ) on the original dataset.

In order to save computation time, we skip the optimization step for a candidate model  $\theta_i$  for which it is unlikely that we get an increase in the validation accuracy by mixing  $\theta_i$  into the current fused model  $\Psi$ . For that, we generate the "approximate average" model  $\tilde{\Psi}$  via

$$\tilde{\Psi} = \frac{k}{k+1} \cdot \Psi + \frac{1}{k+1} \cdot \theta_i \quad (1)$$

and test whether the condition  $ValAcc(\tilde{\Psi}) > \tau \cdot ValAcc(\Psi)$  is fulfilled. If so, we continue with this iteration. If it is not fulfilled, we skip the following steps of this iteration, so candidate model  $\theta_i$  will not be taken into account. The motivation for the specific combination provided in Equation (1) is that  $\tilde{\Psi}$  calculated in this way corresponds approximately to the *average* of all candidate models (like in [Wortsman et al., 2022a]) which have been mixed so far into the fused model (including  $\theta_i$ ), if we assume that the optimization did not change the mixing coefficients drastically from their provided initial values. We set the constant  $\tau$  to 0.998.

Having identified  $\theta_i$  as a promising candidate model, in the next step we determine the optimal factors for mixing its latent space manifolds into the current fused model  $\Psi$ . For this, we define the updated fused model  $\Psi'(\lambda)$  as a *component-wise* convex combination of  $\Psi$  and  $\theta_i$  via

$$\Psi'(\lambda)^j = \lambda^j \cdot \Psi^j + (1 - \lambda^j) \cdot \theta_i^j \quad (2)$$

for all components  $j = 1, \dots, m$ . Note that  $\Psi'(\lambda)$  is a function of the mixing vector  $\lambda$ . The mixing factor  $\lambda^j \in [0, 1]$  determines how much of the  $j$ -th component (latent space manifold) of the candidate model  $\theta_i$  is mixed into the current fused model  $\Psi$ . The component-wise convex combination of the two models allows an optimizer to explore the latent space manifolds of the models  $\Psi$  and  $\theta_i$  in a very flexible way, in order to find the optimal mixing vector  $\lambda^* \in \mathbb{R}^m$  which gives the highest validation accuracy for the updated fused model  $\Psi'$ .

For the subsequent optimization step, we set up the optimization problem to solve as

$$\lambda^* = \arg \max_{\lambda \in [0, 1]^m} (ValAcc(\Psi'(\lambda))) \quad (3)$$

where  $[0, 1]^m$  is the  $m$ -dimensional unit interval. Via the constraint  $\lambda \in [0, 1]^m$  we ensure that a convex combination is done for each component, so we are in fact *interpolating linearly* between the latent space manifolds  $\Psi^j$  and  $\theta^j$ . The model  $\Psi'(\lambda)$  can be calculated via Equation (2).

For solving this optimization problem, we employ the *Nevergrad*<sup>1</sup> optimization package. It provides a large variety of black-box *derivative-free* optimization algorithms together with a sophisticated heuristic [Liu et al., 2020] to select the best optimizer based on the characteristic (number of variables, allowed budget for function evaluations etc.) of the optimization problem. As the initial value for the mixing factors, we set  $\lambda^j = k/(k+1)$  for  $j = 1, \dots, m$  with a similar motivation as explained earlier for Equation (1).

We invoke now the optimizer in order to calculate the optimal mixing vector  $\lambda^*$  which give the highest validation accuracy on the dataset used for finetuning. The optimal updated fused model can be calculated now via  $\Psi'^* = \Psi'(\lambda^*)$ . We check now whether the condition  $ValAcc(\Psi'^*) > ValAcc(\Psi)$  is fulfilled. If so, we have found a better fused model  $\Psi'^*$  by mixing  $\theta_i$  into it. Consequently, we replace  $\Psi$  by  $\Psi'^*$  and increase  $k$  by 1. If this is not the case, we keep the current fused model  $\Psi$  and  $k$  as they are.

After iterating over all candidate models  $\theta_i$  for  $i = 1, \dots, n - 1$  we retrieve a final fused model  $\Psi$  (the *manifold mixing model soup*), which mixes together the  $k$  selected candidate models / ingredients in an optimal way.

<sup>1</sup><https://facebookresearch.github.io/nevergrad/>

**Algorithm 1** Manifold mixing model soup algorithm

---

**Require:** Finetuned models  $\{\theta_0, \dots, \theta_{n-1}\}$  as result of hyperparameter tuning  
**Require:** Partitioning of a model  $\zeta$  into  $m$  components (latent space manifolds)  $\zeta^j$  for  $j = 1, \dots, m$   
**Require:** Function  $ValAcc(\zeta)$  which calculates validation accuracy for  $\zeta$  on dataset used for finetuning

```

 $\{\theta_0, \dots, \theta_{n-1}\} \leftarrow sort(\{\theta_0, \dots, \theta_{n-1}\})$                                 ▷ Sort  $\{\theta_0, \dots, \theta_{n-1}\}$  in descending order based on  $ValAcc(\theta_i)$ 
 $k \leftarrow 1$                                          ▷ Number of candidate models mixed into fused model
 $\Psi \leftarrow \theta_0$                                          ▷ Set initial fused model to best finetuned model  $\theta_0$ 
 $\tau \leftarrow 0.998$                                          ▷ Tolerance factor for promising candidate models
for  $i = 1, \dots, n - 1$  do
     $\tilde{\Psi} = \frac{k}{k+1} \cdot \Psi + \frac{1}{k+1} \cdot \theta_i$ 
    if  $ValAcc(\tilde{\Psi}) > \tau \cdot ValAcc(\Psi)$  then
         $\Psi'(\lambda)^j = \lambda^j \cdot \Psi^j + (1 - \lambda^j) \cdot \theta_i^j$ 
         $\lambda^* = \operatorname{argmax}_{\lambda \in [0,1]^m} (ValAcc(\Psi'(\lambda)))$ 
         $\Psi'^* = \Psi'(\lambda^*)$ 
        if  $ValAcc(\Psi'^*) > ValAcc(\Psi)$  then
             $k \leftarrow k + 1$ 
             $\Psi \leftarrow \Psi'^*$ 
        end if
    end if
end for
return  $\Psi$                                          ▷ Return final fused model

```

---

## 4 Experiments and Evaluation

The setup for our experiments is very similar to the one for the vision models given in the *model soup* paper [Wortsman et al., 2022a]. We summarize it in the following for clarity and completeness.

The model employed for finetuning is the *CLIP* model [Radford et al., 2021]. CLIP is a powerful multi-modal zero-shot neural network, which has been pretrained with contrastive learning on a huge dataset of image-text pairs. Specifically, we use the *CLIP ViT-B/32* variant specified in Table 20 of [Radford et al., 2021] and provided in the *OpenCLIP* package<sup>2</sup>. Finetuning of the pretrained model is performed end-to-end (all parameters are modified), as it typically leads to better performance than training only the final linear layer. Before finetuning, the final layer is initialized with a linear probe as described in [Kumar et al., 2022]. The loss function employed for finetuning is the cross-entry loss.

The original dataset employed for finetuning is *ImageNet* [Deng et al., 2009]. Since the official ImageNet validation dataset is typically used as the test dataset, we use roughly 2% of the ImageNet training dataset as held-out validation dataset for calculating the validation accuracy in our proposed algorithm (see section 3 and the pseudocode provided in Algorithm 1).

For measuring the out-of-distribution performance (robustness to distribution shifts) of our proposed algorithm, we employ five datasets derived from ImageNet with natural (not synthetically generated) distribution shifts. They corresponds to datasets with naturally occurring variations of the data samples due to different lighting, viewpoint, geographic location, image style (e.g. sketch instead of photo), crowdsourcing and more. The five datasets with distribution shifts we use are:

- ImageNet-V2 (IN-V2) [Recht et al., 2019] is a reproduction of the ImageNet test set with distribution shift. The dataset was collected by closely following the original labelling protocol.
- ImageNet-R (IN-R) [Hendrycks et al., 2021a] contains renditions (e.g., sculptures, paintings) for 200 ImageNet classes.

<sup>2</sup>[https://github.com/mlfoundations/open\\_clip](https://github.com/mlfoundations/open_clip)

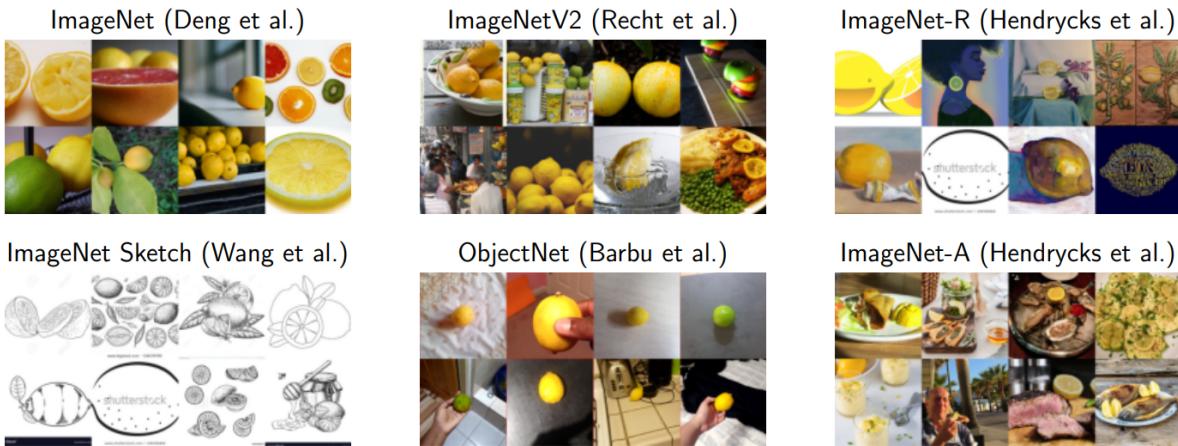


Figure 1: Samples for class *lemon*, from the original ImageNet dataset and the five datasets with natural distribution shifts. Image courtesy of [Wortsman et al., 2022b]

- ImageNet-Sketch (IN-Sketch) [Wang et al., 2019] contains sketches instead of natural images. It contains only sketches in "black-and-white" color scheme.
- ObjectNet [Barbu et al., 2019] provides objects in various scenes with 113 classes overlapping with ImageNet.
- ImageNet-A (IN-A) [Hendrycks et al., 2021b] is a test set of natural images misclassified by a ResNet-50 model for 200 ImageNet classes.

See Figure 1 for an illustration of samples for each of the datasets with natural distribution shifts. For all datasets (the original used for finetuning and the ones with distribution shifts), we take the top-1 accuracy on the respective test set for measuring the performance of a model. We calculate the overall out-of-distribution performance of a model as the average of its test accuracy over all five datasets with distribution shifts.

We partition the CLIP ViT-B/32 model into 8, 15 and 26 components. A too fine partitioning (e.g. one component for each layer of the model) makes the optimization much more difficult, whereas a too coarse partitioning provides not enough flexibility for mixing the latent space manifolds individually in an optimal way. The structure of the partitioning is done roughly according to the hierarchy of the building blocks of the CLIP model. We denote the respective variant of our proposed algorithm with 8, 15 and 26 components as ManifoldMixMS-C8, ManifoldMixMS-C15 and ManifoldMixMS-C26.

We parametrize the Nevergrad optimizer with a maximum budget for the number of function evaluations (of the objective function to optimize) of roughly 250 function evaluations for all ManifoldMixMS variants. The employed optimizer is automatically selected by the Nevergrad optimization package (see [Liu et al., 2020]). For our cases, it always selects the *Cobyla* [Powell, 1994] optimization algorithm. The Cobyla algorithm is one of the best derivative-free algorithms for optimization of continuous variables with bound constraints, especially when the allowed number of function evaluations is quite small.

For the evaluation of our proposed manifold mixing model soup algorithm, we compare mainly with the *greedy soup* and *uniform soup* algorithms which have been proposed in [Wortsman et al., 2022a]. Additionally, we compare our proposed algorithm also against ensemble models. We compare against the same ensemble models as done in [Wortsman et al., 2022a] and take also the accuracy numbers reported there for them. Of course, one should take into account that the computational cost for inference of an ensemble model is much higher –  $K$  times higher for an ensemble model consisting of  $K$  individual models – than for our proposed ManifoldMixMS algorithm which produces only a single fused model.

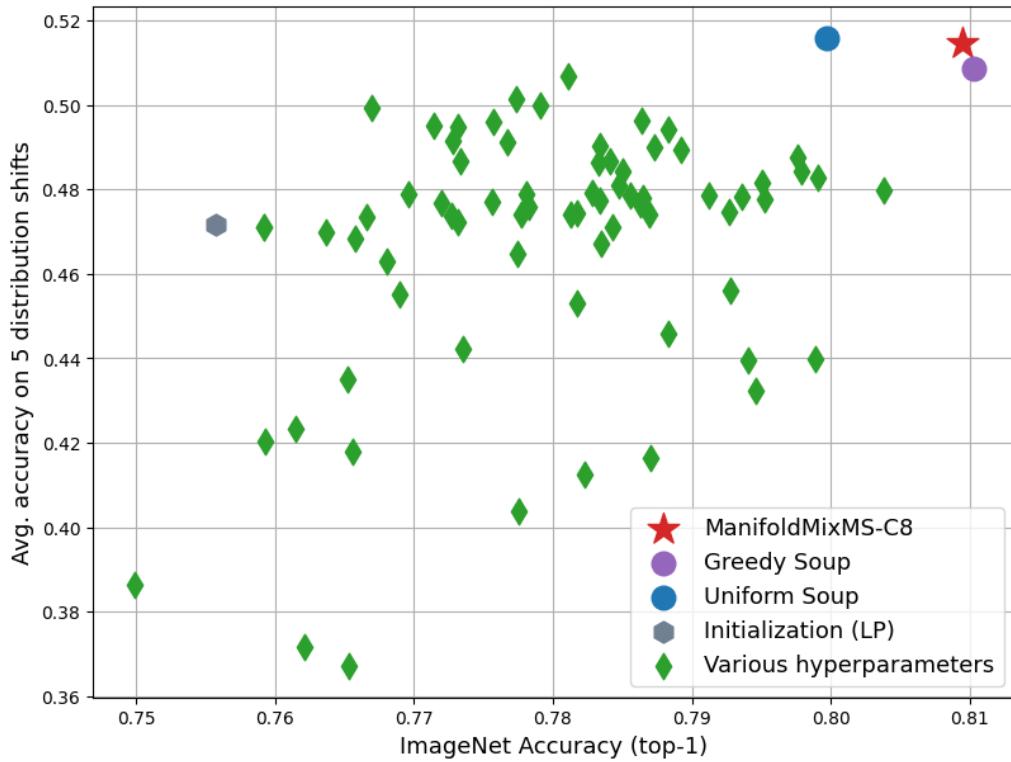


Figure 2: Comparison of our proposed manifold mixing soup algorithm (with 8 components) against greedy soup and uniform soup algorithm from [Wortsman et al., 2022a] and the individual finetuned models.

The scatterplot in Figure 2 shows how our proposed ManifoldMixMS-C8 algorithm (the overall best variant) performs compared to the greedy soup and uniform soup algorithm from [Wortsman et al., 2022a] and to the individual finetuned models.

Furthermore, Table 1 gives a detailed evaluation of our proposed variants of the manifold mixing soup algorithm with 8, 15 and 26 on the five datasets with distribution shifts (ImageNet-V2, ImageNet-R, ImageNet-Sketch, ObjectNet, ImageNet-A) as well as on the original dataset used for finetuning (ImageNet).

One can see clearly from the scatterplot that our proposed manifold mixing model soup (especially the preferred variant with 8 components) algorithm combines the best properties of the uniform model soup and greedy soup algorithm. Specifically, it has practically the same good out-of-distribution accuracy as the uniform soup algorithm and still keeps the good accuracy of the greedy soup algorithm on the original ImageNet dataset. In contrast, the uniform soup algorithm performs on the original ImageNet dataset even worse than the best individual finetuned model.

It is significantly better with respect to the best finetuned model both on the datasets with distribution shifts (+3.5%), but also on the original ImageNet dataset (+0.6%). The difference grows even bigger when comparing with the second-best finetuned model.

Surprisingly, it has also a significantly better out-of-distribution accuracy than both Ensemble methods, although its accuracy on the original ImageNet dataset is worse especially when compared with the greedy ensemble method. As already mentioned, one should take into account that Ensemble methods have a much higher computational cost.

Table 1: Detailed comparison of our proposed manifold mixing soup algorithm variants for the CLIP ViT-B/32 neural network with the best and second-best finetuned model, the model soup algorithms from [Wortsman et al., 2022a] and for completeness also with Ensemble methods. The top-1 accuracy (in %) on the respective test dataset is employed. The column "Avg OOD" corresponds to the average over all 5 datasets with distribution shifts. The best and second-best result for each dataset (without taking into account the Ensemble methods as they have a much higher computational cost) is marked in red and blue.

Method	ImageNet	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	Avg OOD
Best finetuned model	80.38	68.44	44.51	60.63	42.62	23.64	47.97
Second-best finetuned model	79.89	67.91	41.49	54.58	37.98	18.01	44.01
Uniform soup	79.97	68.51	47.71	66.54	45.95	29.17	51.57
Greedy soup	81.03	69.55	47.77	64.20	44.90	27.89	50.86
ManifoldMixMS-C8	80.95	69.67	48.15	64.81	45.66	29.06	51.47
ManifoldMixMS-C15	80.80	69.61	47.89	64.76	44.45	28.39	51.02
ManifoldMixMS-C26	80.85	69.58	48.04	64.79	45.75	28.88	51.41
Ensemble	81.19	—	—	—	—	—	50.77
Greedy ensemble	81.90	—	—	—	—	—	49.44

## 5 Conclusion

We propose the *manifold mixing model soup* algorithm, which mixes together the latent space manifolds of multiple finetuned models in an optimal way in order to generate a fused model. Experiments show that the fused model gives significantly better out-of-distribution performance (+3.5 % compared to best finetuned model) when finetuning a CLIP model for image classification.

In the future, we plan to evaluate the proposed algorithm on other neural network architectures, for both computer vision as well as natural language processing tasks. Furthermore, we plan to do a theoretical analysis of the properties of the proposed algorithm in order to get a better insight why it provides a better out-of-distribution performance.

## Acknowledgment

This work was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media.

## References

- [Barbu et al., 2019] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. (2019). Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.
- [Hendrycks et al., 2021a] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. (2021a). The many faces of robustness: A critical analysis of out-of-distribution generalization. *International Conference on Computer Vision (ICCV)*.

- [Hendrycks et al., 2021b] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. (2021b). Natural adversarial examples. *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Izmailov et al., 2018] Izmailov, P., Podoprikin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 876–885. AUAI Press.
- [Jolicoeur-Martineau et al., 2023] Jolicoeur-Martineau, A., Gervais, E., Fatras, K., Zhang, Y., and Lacoste-Julien, S. (2023). Population parameter averaging (PAPA). *CoRR*, abs/2304.03094.
- [Kumar et al., 2022] Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- [Liu et al., 2020] Liu, J., Moreau, A., Preuss, M., Rapin, J., Rozière, B., Teytaud, F., and Teytaud, O. (2020). Versatile black-box optimization. In *GECCO '20: Genetic and Evolutionary Computation Conference, Cancún Mexico, July 8-12, 2020*, pages 620–628. ACM.
- [Matena and Raffel, 2022] Matena, M. and Raffel, C. (2022). Merging models with fisher-weighted averaging. In *NeurIPS*.
- [Powell, 1994] Powell, M. J. D. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*.
- [Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139, pages 8748–8763. PMLR.
- [Recht et al., 2019] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*. <https://arxiv.org/abs/1902.10811>.
- [von Oswald et al., 2021] von Oswald, J., Kobayashi, S., Sacramento, J., Meulemans, A., Henning, C., and Grewe, B. F. (2021). Neural networks with late-phase weights. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- [Wang et al., 2019] Wang, H., Ge, S., Lipton, Z., and Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/1905.13549>.
- [Wortsman et al., 2022a] Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Lopes, R. G., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. (2022a). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*.
- [Wortsman et al., 2022b] Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. (2022b). Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*.
- [Yu et al., 2022] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022.

# Plant Disease Detection on Multispectral Images using Vision Transformers \*

Dane Brown and Malithi De Silva

*Rhodes University, South Africa*

## Abstract

Precise identification of plant diseases facilitates effective agricultural management. However, classifying plant diseases poses formidable challenges due to various factors. Traditional detection methods have inherent drawbacks, including time-consuming processes, labour-intensive efforts, and the need for specialised expertise. Overcoming these limitations and developing improved approaches is crucial.

Deep learning, particularly hybrid vision transformers (ViTs), has emerged as a promising non-invasive plant disease identification solution. Despite advancements, accurate classification remains complex. The intricate nature of plant diseases and variations in symptoms across species and environmental conditions are significant obstacles.

This study focuses on harnessing cutting-edge classification algorithms, specifically a hybrid ViT framework for plant disease identification. Combining convolutional neural networks (CNNs) and ViTs enhances accuracy and efficiency, improving agricultural practices.

To evaluate the proposed method, a multispectral dataset was collected outdoors using lens filters, covering visible and Near-infrared (NIR) ranges. This comprehensive dataset was photographed on outdoor plants with a complex background such that good *in-the-wild* performance and thorough analysis of plant diseases is possible in a real-world setting. Experimental findings demonstrate the superiority of cutting-edge hybrid ViT models with an accuracy of 88.86%. These outcomes underscore the substantial potential of hybrid ViT models in accurately identifying plant diseases, even in previously unseen images captured under varying environmental conditions.

**Keywords:** Computer Vision, Deep Learning, Transformer, CNN, Multispectral

## 1 Introduction

Agriculture plays a vital role in the global economy as a significant source of employment, income, and exports for many countries, with developing nations relying heavily on agriculture, which accounts for over half of their workforce [Brown and Mazibuko, 2023]. Moreover, agriculture is the primary global food source, ensuring food security and stabilizing prices.

However, the agricultural sector faces numerous threats that can significantly reduce crop yields, leading to substantial production losses and financial hardships for farmers. Among these threats, plant diseases emerge as a prominent challenge, capable of severely impacting harvest production. Plant diseases encompass various conditions and abnormal growth patterns that adversely affect the health and vitality of plants. Factors such as fungi, bacteria, viruses, parasites, and environmental stressors contribute to the prevalence of these diseases, affecting all plant components, including roots, stems, leaves, and fruit [Walton, 1997].

Timely identification of diseases is crucial to minimize crop yield reductions, mitigate food shortages, and prevent escalating food prices that contribute to hunger and malnutrition. Conventional disease management

---

\*This work was undertaken in the Distributed Multimedia CoE at Rhodes University.

often relies on increased pesticide usage, incurring high costs and leading to negative environmental consequences like water pollution and harm to wildlife. Consequently, plant diseases significantly adversely impact the global economy, hindering crop yields, reducing farmer income, and limiting food accessibility.

To address these challenges, effective disease management strategies must focus on early disease identification to protect plants and their surroundings. Deep learning models have gained recognition for their effectiveness in detecting plant diseases at their early stages, leading to improved disease management outcomes. This paper combines the state-of-the-art vision transformer models with CNNs to identify plant diseases during their early development.

The approach involves capturing photographs using multiple filters obtained from Kolari Vision under diverse weather and lighting conditions. These specialised filters are meticulously designed to capture specific segments of the near-infrared (NIR) spectrum, which proves invaluable for imaging and analysis purposes. The NIR spectrum spans wavelengths ranging from approximately 700-2500 nm, occupying the region between the visible and mid-infrared ranges of the electromagnetic spectrum. By leveraging NIR radiation, which lies beyond the visible spectrum, this imaging methodology reveals critical details imperceptible to the human eye, offering exceptional computer vision possibilities for various agricultural applications.

Following contributions were made in this paper.

- Acquisition of a [new multispectral dataset](#) using four Kolari vision lenses (BlueIR, K590, K850, and Hot Mirror) in diverse natural environmental settings with varying weather conditions.
- Review of multiple hybrid models combining state-of-the-art ViT and CNN models with the multispectral dataset to identify the most effective combination for early plant disease detection.
- Utilisation of two different augmentation mechanisms to assess the impact of small and medium augmentations on early plant disease detection with the new dataset.
- Comparative analysis of multiple lenses to determine their suitability for early plant disease identification research in terms of imaging techniques.
- Results comparison on multiple CNN and ViT models with the hybrid models using the same dataset to select the most suitable deep learning model for early disease identification research.

The remaining sections of this paper are organised: Section 2 provides a review of the relevant literature, Section 3 describes the materials and methodology employed, Section 4 presents the experimental results, and finally, Section 5 concludes the paper with a discussion of the results and future directions.

## 2 Related Works

Many deep learning approaches have gained popularity in plant disease detection research as non-invasive methods. The following review highlights state-of-the-art studies in the context of plant disease detection.

[Li and Li, 2022] proposed a lightweight hybrid model for apple disease identification by combining convolutional neural networks (CNNs) and transformers. They used CNNs to extract local features and transformers to capture long-range dependencies and global context information. The proposed method achieved 98.2% accuracy on apple disease images, surpassing several existing state-of-the-art CNN and Vision Transformer (ViT) methods.

In 2022, [Lu et al., 2022] developed a hybrid model to diagnose grape leaf diseases and pests effectively. Their approach involved a pre-processing step to enhance image quality, a ghost-convolutional neural network (GCNN) for feature extraction, and a transformer-based architecture to capture long-range dependencies and global context information. The model achieved an accuracy of 98.14% on a dataset of 12615 grape leaf images, demonstrating improved performance compared to conventional CNN models.

[Zhou et al., 2023] introduced a method for identifying rice leaf diseases using a residual-distilled transformer (RDT) architecture. The RDT architecture combined residual learning with a transformer-based approach. The proposed method achieved an accuracy of 92

In 2023, [Öğrekçi et al., 2023] investigated the classification of sugarcane leaf diseases using deep learning methods. They compared the performance of CNN, ViT, and ViT+CNN models on a dataset of 2,521 images categorised into five classes. The ViT model outperformed the other two models, achieving accuracies of 93.34

These studies highlight the advancements in plant disease identification using hybrid ViT models. Furthermore, while previous works focused on RGB images, there is a growing interest in adding near-infrared (NIR) to the spectra to effectively capture early plant disease symptoms under varying conditions – as NIR imaging can provide additional information beyond the visible range of the electromagnetic spectrum.

### 3 Methodology

This section presents the experimental setup for this work, which includes data collection, augmentation, image processing, and the architectures of the deep learning models.

#### 3.1 Data Collection

Images of plant leaves were captured, including the natural background using a Canon EOS 800D camera equipped with Kolari Vision filters, namely BlueIR, IR 590 nm, Hot Mirror, and IR K850, as shown in Figure 1. Each filter allows specific electromagnetic spectrum ranges to be captured, enabling the acquisition of relevant features [Brown and Poole, 2023]. The K590 filter captures wavelengths in 590-1000 nm, while the K850 filter allows 850-1000 nm wavelengths. The BlueIR filter captures both blue (400-500nm) and infrared (700-1000nm) ranges, while the Hot Mirror filter only allows the visible portion of the electromagnetic spectrum by blocking other wavelengths.

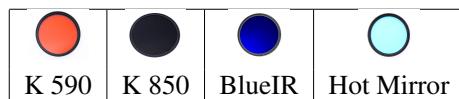


Figure 1: Camera lenses for filtering.

Table 1 indicates that 3016 images were captured in the morning hours (8:00-11:59 a.m.) under various weather conditions, including sunny, windy, cloudy, and rainy, with temperatures ranging from 22-37 °C. The image collection focused on five fruit plants: avocados, tomatoes, limes, passion fruit, and gooseberries. These plants are susceptible to various diseases, such as bacterial spot, mold, rust, and common viruses. Examples of the collected image samples are shown in Figure 2. Due to environmental conditions, capturing the same number of samples per species is not feasible. Thus, data augmentation is used as a mitigating factor to improve training. Each image contained at least one leaf and included the natural background.

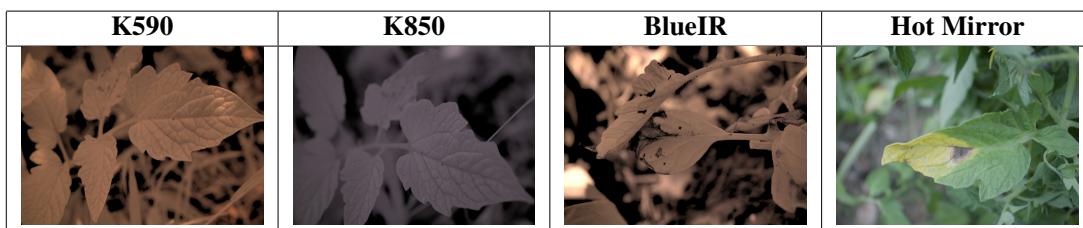


Figure 2: Image samples of the NIR dataset.

Data augmentation is crucial in this work as the dataset is relatively small for CNNs and more so for transformer-based algorithms. Deep learning models require a large amount of data, and data augmentation

Table 1: Plant species that were captured and their sample distribution.

Plant Status	Lens Type			
	K850	Hot Mirror	BlueIR	K590
Avocado Diseased	99	77	50	37
Avocado Healthy	206	144	91	42
Lime Diseased	142	124	64	50
Lime Healthy	177	104	89	38
Tomato Diseased	111	67	35	25
Tomato Healthy	111	144	114	22
Gooseberry Diseased	23	37	14	22
Gooseberry Healthy	140	127	116	133
PassionFruit Diseased	-	21	-	-
PassionFruit Healthy	100	45	75	-

helps enhance the diversity and size of the training dataset, leading to improved model generalisation and robustness. Data augmentation creates new samples by applying various transformations or modifications to the existing data, reducing overfitting and enhancing the model's ability to recognize patterns in unseen data.

However, it is vital to note that excessive use of augmentation may result in diminishing returns. To determine the most suitable augmentation strategy for the dataset, two different mechanisms were employed: small and medium augmentation. Small augmentations included random vertical and horizontal flipping, as well as random resizing. On the other hand, medium augmentation involved random cropping, random flipping, and colour jittering. These techniques were applied with the *ViT\_r26\_s32* model.

### 3.2 Feature Extraction and Classification

The feature extraction process is crucial for classification, as it involves identifying and extracting the most relevant and discriminative features from raw data, enabling differentiation between different classes.

This research adopts hybrid Vision Transformers (ViTs), which combine the strengths of convolutional neural networks (CNNs) and transformers. CNNs excel at extracting local features from images, while transformers effectively capture long-range dependencies. The proposed approach uses CNNs for feature extraction from input images, while the ViT model focuses on the classification or recognition task [Agarwal and Dash, ].

The hybrid model takes an input image and passes it through a CNN or ResNet, generating a feature map. This feature map is subsequently divided into patches, flattened, and fed into the ViT model, as shown in Figure 3. The ViT model treats these patches as a sequence of tokens and utilizes transformer blocks to encode the spatial and channel-wise relationships among the patches. Finally, the encoded sequence is processed by a linear classifier to produce the final predictions.

This paper used *ViT\_r26\_s32* and *ViT\_r50\_l32* models for classification. Then the impact of the size of the CNN model and the ViT model can be evaluated by examining the outcome of the models.

#### 3.2.1 ViT\_r26\_s32

The *ViT\_r26\_s32* model adopts a hybrid architecture for computer vision tasks. It utilizes a ResNet-26 architecture with 26 convolutional layers for local feature extraction. The input image is divided into non-overlapping patches of size 32x32 pixels, which are then processed by the ResNet-26 backbone. To capture long-range dependencies, a Transformer-based head with eight Transformer layers is employed for global self-attention [Darvish et al., 2022].

Two variations of the *ViT\_r26\_s32* model were utilised: *ViT\_r26\_s32\_smallaug*, which incorporates a small augmentation mechanism, and *ViT\_r26\_s32\_medaug*, which employs a medium augmentation mechanism.

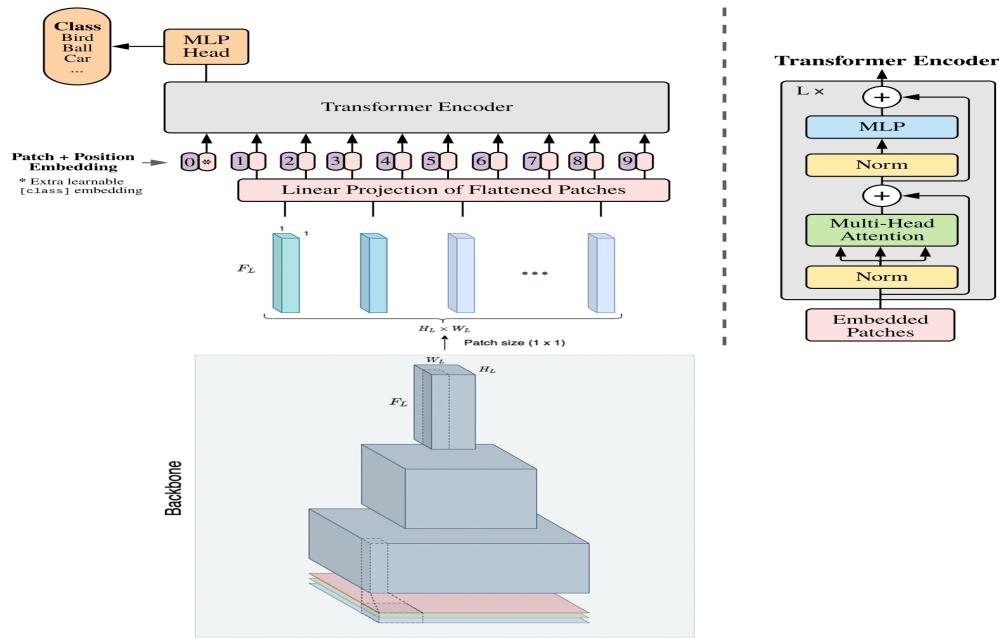


Figure 3: Hybrid Vision Transformer

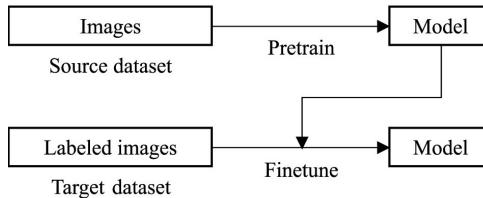


Figure 4: Transfer Learning.

### 3.2.2 ViT\_r50\_l32

This model uses a ResNet-50 backbone for local feature extraction and a Transformer-based head for global self-attention. The ResNet-50 architecture comprises 50 convolutional layers, and the patch size used in the ViT model is 32x32 pixels. The "l" prefix indicates the large ViT model, which consists of 24 transformer layers [Garcia-Martin and Sanchez-Reillo, 2023].

The hybrid ViT models in this study were trained using transfer learning. This involves initializing the model's weights with those of a pre-trained model and fine-tuning it for a different task or domain. Figure 4 illustrates the transfer learning process. Due to the limited number of images in the dataset, the hybrid ViT models were pre-trained on the ImageNet-21k dataset, and cross-validation was not feasible from a time and dataset size perspective.

To prevent overfitting and enhance generalisation, dropout layers were implemented in the hybrid ViT models. Dropout layers randomly drop out units during training, helping to prevent the model from overfitting to the training data. The Adam optimizer was used with a batch size of 32 and an initial 1e-04 learning rate.

Early stopping was implemented to prevent overfitting. It monitored the validation accuracy and stopped training if it did not improve for several epochs. In this case, the learning rate was reduced if the validation accuracy did not improve for 15 epochs.

The dataset was initially divided into two subsets: the training set and the testing set, with a ratio of 80:20. The training set was used to train the model. In contrast, the testing set evaluated the trained model's predictions

to test the generalisation capability of the model.

The effectiveness of the models was primarily evaluated based on accuracy, which is a reliable metric for measuring the model's performance. Accuracy is the ratio of correct predictions to the total number of predictions made.

## 4 Results

The experiments in this study involved using different lenses, namely K850, K590, Blue IR, and Hot mirror lenses. Each lens was tested individually, and the following experimental setups were employed.

### 4.1 Experiment 1

The *ViT\_r26\_s32* model was trained on the datasets obtained with the above lenses. Small and medium augmentation mechanisms were applied to compare the disease identification capability of different lenses and assess the impact of augmentation on the model's performance.

### 4.2 Experiment 2

Both the *ViT\_r26\_s32* and *ViT\_r50\_l32* models were trained on the datasets obtained with the above lenses. This experiment aimed to compare the disease identification capability of different lenses and evaluate the effect of model size on performance.

### 4.3 Results

This study used two hybrid ViT models with different augmentation mechanisms for classification. The results obtained for each model are presented in Table 2.

Table 2: Experiment 1 : Accuracy (%) of *ViT\_r26\_s32\_smallaug* and *ViT\_r26\_s32\_medaug* models.

Lens	<i>ViT_r26_s32_smallaug</i>		<i>ViT_r26_s32_medaug</i>	
	Train	Test	Train	Test
Blue IR	91.28	85.32	91.87	83.32
K590	89.41	76.86	90.16	76.92
Hot Mirror	91.41	82.75	90.86	81.88
K850	92.23	88.17	92.83	88.86

The models trained with small augmentation demonstrated superior performance to those trained with medium augmentation. Among the lenses tested, the K850 lens exhibited the highest accuracy in both the training and testing phases for both types of augmentation.

Specifically, the model trained with small augmentation achieved a higher testing accuracy of 85.32% for the Blue IR lens, surpassing the accuracy of the model trained with medium augmentation (83.32%). However, for the K590 and K850 lenses, the model with medium augmentation showed a slightly higher testing accuracy than the model with small augmentation, although the difference was not statistically significant. It is worth noting that the K590 lens presented the most challenging classification task, consistently yielding lower accuracies across all configurations compared to the other lenses.

These findings indicate that small augmentation is an effective technique for enhancing model performance across various lenses. Small augmentation facilitates the learning of robust features that introduce less noise and variations in the data. Conversely, applying medium or larger augmentations did not yield significant improvements in model performance for the tested lenses.

Given the limited impact observed with medium augmentation, further investigation was conducted to assess the influence of model size by comparing the *ViT\_r26\_s32* and *ViT\_r50\_l32* models using small augmented data during training.

Table 3: Experiment 2 : Accuracy (%) of *ViT\_r26\_s32* and *ViT\_r50\_l32* models.

Lens	<i>ViT_r26_s32</i>		<i>ViT_r50_l32</i>	
	Train	Test	Train	Test
Blue IR	91.28	85.32	92.42	84.87
K590	90.74	76.86	90.45	76.09
Hot Mirror	91.41	82.75	91.35	80.83
K850	92.23	88.17	91.73	87.28

Table 3 provides a comparison of the results, revealing that the models trained with the *ViT\_r26\_s32* architecture consistently exhibited higher accuracy scores for both training and testing across all lenses when compared to the models trained with the *ViT\_r50\_l32* architecture. Notably, the K850 lens consistently achieved the highest accuracy scores across all configurations in both models, while the K590 lens proved to be the most challenging to classify, displaying the lowest accuracies.

The inferior performance of the *ViT\_r50\_l32* model may be attributed to its larger size, characterised by more parameters and layers. This increased complexity presents challenges in training and optimisation, resulting in lower accuracy scores. Additionally, the larger model is more prone to overfitting the training data, leading to diminished generalisation performance on the test data. It is plausible that the smaller *ViT\_r26\_s32* model, with its optimal capacity, effectively learns the necessary features for accurate classification, rendering the larger model unnecessary and providing no additional benefits.

It is worth noting that previous studies [De Silva and Brown, 2023a, De Silva and Brown, 2023b] have explored the analysis with similar lens filters and experimental setups. DenseNet121 yielded the lowest test accuracy (76.69%), while the ViT-B16 model achieved the highest test accuracy by 10%. The study attributed this outcome to adverse weather conditions, such as rain and minimal sunlight, which affected CNNs more negatively than vision transformers. In particular, images containing water drops on leaves could have posed challenges for the CNN models in accurately classifying healthy and diseased plants. The collected dataset for this paper did not collect images when leaves were wet, but the results were similar – hybrid models outperformed CNNs and, to a lesser extent, vision transformers. The consistency of inference run on unseen data and the additional improvements from hybrid ViT models thus provide promising prospects for deploying it in agricultural systems.

## 5 Conclusion

This study aimed to address the challenges of precise identification and early detection of plant diseases by applying deep learning techniques, specifically hybrid vision transformers (ViTs) combined with convolutional neural networks (CNNs).

Section 1 showed notable contributions to a new multispectral dataset encompassing diverse natural environmental settings and weather conditions, including the plant and the naturally complex background. This dataset provides a valuable resource for future research in early plant disease detection and imaging techniques. Moreover, evaluating multiple hybrid models and augmentation mechanisms provides a results contribution to help establish practical approaches for early plant disease identification.

The experimental findings demonstrated the superiority of hybrid ViT models, particularly the *ViT\_r26\_s32* architecture, with state-of-the-art performance in accurately identifying plant diseases. Notably, the K850 lens dataset exhibited remarkable training accuracy of 92.83% and testing accuracy of 88.86%. These outcomes

underscore the substantial potential of hybrid ViT models in early disease detection, even in previously unseen images captured under varying environmental conditions.

The potential impact of advanced plant disease identification techniques on precision agriculture and crop management is substantial. Timely and accurate disease diagnosis, enabled by applying deep learning models, can significantly improve agricultural practices, disease prevention strategies and ultimately optimize crop yield. By minimizing production losses, reducing reliance on pesticides, and promoting sustainable farming practices, these advancements contribute to food security, economic stability, and environmental preservation.

In conclusion, this study highlights the effectiveness of hybrid ViT models and their potential in early plant disease detection. Future research can build upon these findings by identifying diseases at their earliest stages, expanding the dataset size, and refining the deep learning models. These advancements will enhance the practical application of plant disease identification techniques, benefiting farmers, researchers, and the wider agricultural community.

## References

- [Agarwal and Dash, ] Agarwal, S. and Dash, R. Conv-vit-hgr: A transformer based hybrid model for hand gesture recognition in complex background. *Available at SSRN 4084524*.
- [Brown and Mazibuko, 2023] Brown, D. and Mazibuko, S. (2023). Efficient plant disease detection and classification for android. *Inventive Systems and Control. Lecture Notes in Networks and Systems*, 672.
- [Brown and Poole, 2023] Brown, D. and Poole, L. (2023). Enhanced plant species and early water stress detection using visible and near-infrared spectra. *Computational Vision and Bio-Inspired Computing. Advances in Intelligent Systems and Computing*, 1439.
- [Darvish et al., 2022] Darvish, M., Pouramini, M., and Bahador, H. (2022). Towards fine-grained image classification with generative adversarial networks and facial landmark detection. In *2022 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–6. IEEE.
- [De Silva and Brown, 2023a] De Silva, M. and Brown, D. (2023a). Plant disease detection using multispectral imaging. *Lecture Notes in Networks and Systems, In Press*, (1).
- [De Silva and Brown, 2023b] De Silva, M. and Brown, D. (2023b). Plant disease detection using vision transformers on multispectral natural environment images. In *2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), In Press*. IEEE.
- [Garcia-Martin and Sanchez-Reillo, 2023] Garcia-Martin, R. and Sanchez-Reillo, R. (2023). Vision transformers for vein biometric recognition. *IEEE Access*, 11:22060–22080.
- [Li and Li, 2022] Li, X. and Li, S. (2022). Transformer help CNN see better: A lightweight hybrid apple disease identification model based on transformers. *Agriculture*, 12(6):884.
- [Lu et al., 2022] Lu, X., Yang, R., Zhou, J., Jiao, J., Liu, F., Liu, Y., Su, B., and Gu, P. (2022). A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. *Journal of King Saud University-Computer and Information Sciences*, 34(5):1755–1767.
- [Öğrekçi et al., 2023] Öğrekçi, S., Ünal, Y., and Dudak, M. N. (2023). A comparative study of vision transformers and convolutional neural networks: sugarcane leaf diseases identification. *European Food Research and Technology*, pages 1–11.
- [Walton, 1997] Walton, J. D. (1997). 13 biochemical plant pathology. *Plant biochemistry*, page 487.
- [Zhou et al., 2023] Zhou, C., Zhong, Y., Zhou, S., Song, J., and Xiang, W. (2023). Rice leaf disease identification by residual-distilled transformer. *Engineering Applications of Artificial Intelligence*, 121:106020.

# Quality of Multimedia Experience Prediction using Peripheral Physiological Signals

Sowmya Vijayakumar<sup>1</sup>, Ronan Flynn<sup>1</sup>, Peter Corcoran<sup>2</sup>, and Niall Murray<sup>1</sup>

<sup>1</sup>*Department of Computer and Software Engineering, Technological University of the Shannon, Athlone, Ireland*

<sup>2</sup>*Electrical and Electronic Engineering, University of Galway, Galway, Ireland*

## Abstract

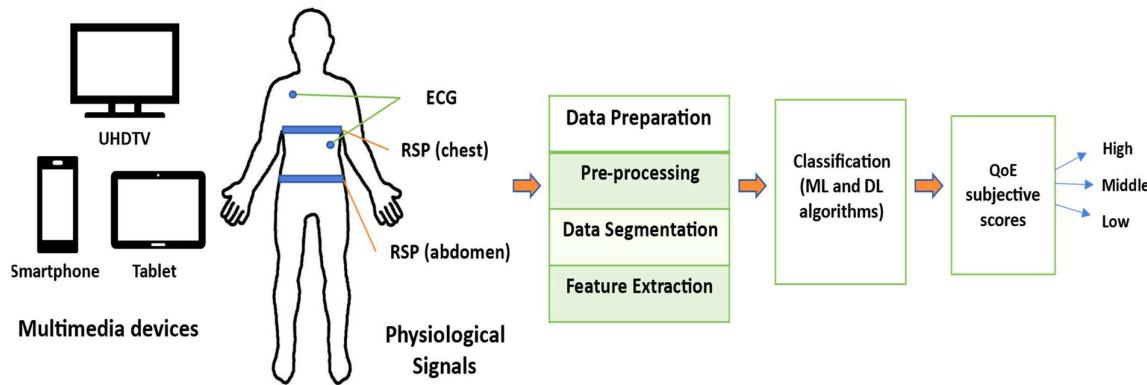
This paper proposes the utilization of physiological signals for quality of experience (QoE) assessment by employing machine learning and deep learning classifiers. Accurately predicting user QoE by analysing physiological signals holds significant potential in diverse fields, including human-computer interaction, healthcare, and education. To predict various QoE factors from physiological signals, the experiments were conducted on two datasets: SoPMD Dataset 1 and SoPMD Dataset 2. The bidirectional long-short-term memory (BLSTM), support vector machine, k-nearest neighbour and random forest algorithms were evaluated using fused electrocardiogram and respiration signals to predict subjective QoE scores, including perceived quality levels, user preference, and the sense of presence. The results demonstrate the effectiveness of the models, with BLSTM emerging as the top-performing algorithm across most experiments, achieving high classification F1-scores. These findings suggest that the physiological signals can be effectively used in the classification of subjective QoE scores.

**Keywords:** Deep Learning, Machine Learning, Physiological Signals, Quality of Experience.

## 1 Introduction

In today's digital era, the quality of user experience (QoE) has become a crucial aspect of the success of applications and services. QoE, a multidimensional concept, is defined as "*the degree of delight or annoyance of the user of an application or service, resulting from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users' personality and current state*" [Möller et al., 2014]. Traditional QoE assessment approaches that rely on subjective ratings and self-reports may be biased and time-consuming. On the other hand, objective approaches utilize various metrics to predict perceived quality, but they may not accurately reflect the perceived QoE. Therefore, developing reliable and accurate QoE models is essential to effectively capture users' perception of quality.

Recent studies have explored the use of physiological signals for QoE assessment, presenting a promising approach to understanding users' experiences. Wearables enable continuous monitoring of physiological signals in facilitating convenient and real-time QoE assessment. Analysing these signals provides insights into users' cognitive and affective states, enhancing our understanding of their experiences. While previous QoE assessment studies [Engelke et al., 2017] relied on statistical correlations, recent research has explored the potential of machine learning (ML) and deep learning (DL) techniques for predicting QoE using physiological signals [Vijayakumar et al., 2022]; [Perrin et al., 2015]. Despite the widespread use of artificial intelligence (AI) techniques in various domains, their application in QoE assessment using physiological signals is still limited. This study aims to investigate the effectiveness of physiological signals in predicting QoE, employing ML and DL models to enhance the accuracy of physiology-based QoE predictions.



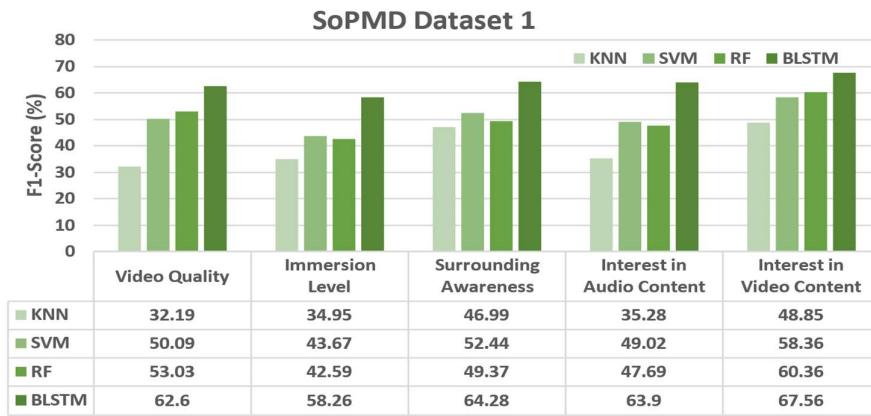
**Figure 1:** The proposed architecture of the QoE prediction from physiological signals for typical multimedia consumption scenarios using AI techniques.

## 2 Methods

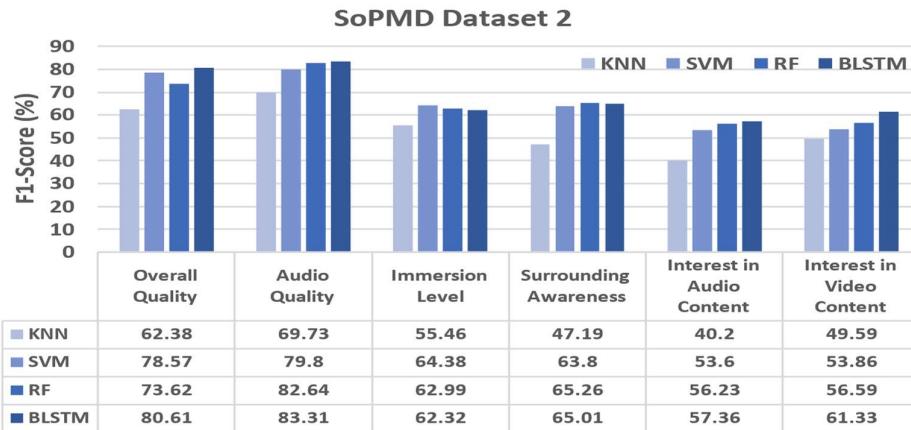
Fig. 1 outlines the methodology of physiology-based QoE prediction. The process involves steps including data pre-processing, segmentation, feature extraction, and classification.

The methodology is applied to two datasets, namely SoPMD Dataset 1 [Perrin et al., 2015] and SoPMD Dataset 2 [Perrin et al., 2016]. These datasets consist of physiological signals and various QoE factors, including perceived quality levels, user preference, and the sense of presence (SoP). The stimuli employed in these datasets were carefully designed to induce different levels of immersion by varying factors such as audio sound systems, compression levels, resolution, and device type. In the SoPMD Dataset 1, a QoE database was developed for the analysis of perceived Sense of Presence (SoP). The audio-visual stimuli were configured to induce low, middle and high levels of immersion based on the video quality (level of compression), resolution (UHD, HD and SD), and sound reproduction (mono, stereo and 5.1). On the other hand, the SoPMD Dataset 2 is a multimodal database that investigates QoE across three commonly used devices: iPad, iPhone and UHD TV. Each dataset includes a total of 540 data instances collected from 20 participants, including a set of 27 videos or trials. During each trial, three physiological signals, electroencephalogram (EEG), electrocardiogram (ECG), and respiration (RSP) were recorded at a sampling rate of 250 Hz. At the end of each trial, the participants provided QoE ratings using a 9-point scale, ranging from 1 to 9. In our study, ECG and RSP signals are used to predict QoE because, in contrast to EEG signals, they can be measured non-invasively using wearable devices, allowing for long-term ecologically valid monitoring. For classification, rating values from 1 to 3 were categorized as low class, 4 to 6 as middle class, and 7 to 9 as high class.

This study aimed to predict QoE subjective ratings as a low, middle, and high class using ECG and RSP signals. The signal from the 60-second stimulus period was processed for both ECG and RSP signals. The processed signal was then used for segmentation and feature extraction. For segmentation, a sliding window of 3 seconds with 50% overlap was applied to divide the 60-second signals into 40 frames. From each of these frames, a total of 19 features were extracted from various domains such as time, frequency, and non-linear. These features included mean, median, minimum, maximum, standard deviation, variance, first-degree difference, second-degree difference, normalized mean, normalized minimum, normalized maximum, normalized first-degree difference, normalized second-degree difference, skewness, kurtosis, power spectrum, an average of the gradients, sample entropy, and Hurst component. The models evaluated were bidirectional long-short-term memory (BLSTM) [Graves et al., 2005], support vector machine (SVM), k-nearest neighbour (KNN) and random forest (RF). These models were chosen based on our preliminary analysis [Vijayakumar et al., 2022], which indicated that they performed well for the QoE classification task. Since the classes in both datasets were imbalanced, the Synthetic Minority Oversampling TEchnique, SMOTE [Chawla et al., 2002] was applied to the training set to balance the class distribution, excluding



**Figure 2: Comparison of F1-Scores of shallow ML models and BLSTM model for 3-class classification of QoE subjective scores from the fusion of ECG and RSP signals of the SoPMD Dataset 1.**



**Figure 3: Comparison of F1-Scores of shallow ML models and BLSTM model for 3-class classification of QoE subjective scores from the fusion of ECG and RSP signals of the SoPMD Dataset 2.**

the test set. The training and testing sets were standardized using the mean and standard deviation of the training samples. To assess the performance of the models, stratified 10-fold cross-validation was employed. Each fold consisted of a 90% training set and a 10% testing set. The final performance was calculated by averaging the results from the ten folds. The performance metrics evaluated were accuracy, F1-score, precision, and recall.

### 3 Results

The study implemented ML and DL algorithms on SoPMD Dataset 1 to classify subjective ratings for five QoE factors, namely, level of immersion, perceived video quality, interest in audio content, interest in video content and surrounding awareness. The 3-class classification results of SoPMD Dataset 1 are shown in Figure 2. The BLSTM model outperformed the ML models, achieving classification accuracies and F1-scores ranging between 58% and 67% for classifying different QoE factors. The results were encouraging, indicating the effectiveness of the BLSTM framework for QoE assessment in this dataset.

Similarly, in SoPMD Dataset 2, the ML and DL algorithms were used to predict QoE ratings for six factors, namely, level of immersion, perceived overall quality, perceived audio quality, interest in audio content, interest in video content and surrounding awareness. The 3-class classification results of SoPMD Dataset 2, as depicted in

Figure 3, demonstrate the effectiveness of the BLSTM and RF classifiers in classifying QoE ratings based on physiological signals, with F1-scores ranging from 57% to 83%. Notably, the performance of perceived quality factors, such as overall and audio quality, was higher compared to SoPMD Dataset 1, achieving F1-scores of 80.61% and 83.31%, respectively. Similarly, the assessment of surrounding awareness in SoPMD Dataset 2 yielded comparable results to SoPMD Dataset 1. However, SOPMD Dataset 1 exhibited slightly better performance in predicting content preference ratings when compared to SoPMD Dataset 2.

Overall, the results demonstrate the effectiveness of the BLSTM algorithm in both datasets in classifying QoE ratings based on physiological signals, while also highlighting the differences in performance between the two datasets in terms of perceived quality factors and content preference rating prediction. The proposed framework in the study effectively reduces the number of parameters and improves classification performance through the use of a sliding window approach and feature extraction techniques. However, further investigation and critical analysis are necessary to fully evaluate the effectiveness, generalizability, and robustness of the proposed framework.

## 4 Conclusion

This study demonstrates the potential of using physiological signals and AI techniques to accurately predict user QoE. The results obtained from experiments conducted on two datasets, SoPMD Dataset 1 and SoPMD Dataset 2, highlight the effectiveness of incorporating physiological signals in predicting subjective QoE scores. These findings emphasize the importance of integrating physiological signals and ML and DL techniques to advance QoE prediction and enhance user experiences in multimedia applications.

## Acknowledgements

This research was funded by the Irish Research Council under grant number GOIPG/2021/357.

## References

- [Chawla et al., 2002] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321–57.
- [Engelke et al., 2017] Engelke, U. et al. 2017. “Psychophysiology-Based QoE Assessment: A Survey.” *IEEE Journal of Selected Topics in Signal Processing* 11(1): 6–21.
- [Graves et al., 2005] Graves, A., and Jürgen, S. 2005. “Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures.” *Neural Networks* 18(5): 602–10.
- [Möller et al., 2014] Möller, S., and Raake, A. eds. 2014. *Quality of Experience: Advanced Concepts, Applications and Methods*. Cham: Springer International Publishing. <http://link.springer.com/10.1007/978-3-319-02681-7>.
- [Perrin et al., 2015] Perrin, A. M. et al. 2015. “Multimodal Dataset for Assessment of Quality of Experience in Immersive Multimedia.” In *Proceedings of the 23rd ACM International Conference on Multimedia - MM '15*, Brisbane, Australia: ACM Press, 1007–10. <http://dl.acm.org/citation.cfm?doid=2733373.2806387>.
- [Perrin et al., 2016] Perrin, A. M. et al. 2016. “Towards Prediction of Sense of Presence in Immersive Audiovisual Communications.” *Electronic Imaging* 2016(16): 1–8.
- [Vijayakumar et al., 2022] Vijayakumar, S. et al. 2022. “BiLSTM-based Quality of Experience Prediction using Physiological Signals.” *QoMEX 2022*, Germany: IEEE, 1–4.

# Physiological Synchrony: A Novel Approach to Evaluating User Quality of Experience in Collaborative Distributed Virtual Reality Environments

\*Bhagyabati Moharana, Dr. Conor Keighrey, Dr Niall Murray

*Department of Engineering & Informatics, Technological University of Shannon, Athlone, Ireland*

## Abstract

Synchrony at the physiological level provides an objective measure that can be utilized to examine the collaboration between collaborating partners. When individuals collaborate on a task, their physiological signals can provide valuable insights into their cognitive states, emotions, and overall engagement levels. By analyzing and interpreting these signals synchrony, it becomes possible to gain a deeper understanding of the users' experiences and optimize the collaborative environment accordingly. In our proposed VR system, synchrony from the physiological data of individuals is used to infer the user Quality of Experience (QoE) of collaborators. We propose Physiological Synchrony (PS) of physiological signals as a new method to investigate QoE in a collaborative distributed Virtual Reality immersive environment. We investigate the potential effects of the proposed VR system to support synchrony during remote collaboration, as well as the design guidelines for building such systems.

**Keywords:** Virtual Reality (VR), Physiological Synchrony (PS), Quality of Experience (QoE).

## 1 Introduction

Virtual Reality (VR) emerges as a valuable tool for facilitating collaborative tasks in situations where physical proximity is restricted. Immersive VR enables individuals to engage in collaborative activities, such as meetings, conferences, training, and teamwork, while joining from different locations around the globe [Saffo et al., 2021]. Evaluating the Quality of Experience (QoE) of users in immersive VR helps understand how users perceive and experience the environment, which is crucial for assessing the effectiveness and usability of VR platforms [Vlahovic et al., 2022]. With the steady rise in the number of collaboration platforms providing rich interactive experiences like Spatial.io (<https://spatial.io/>), AltspaceVR (<https://altvr.com/>), Facebook Horizon(<https://www.oculus.com/facebook-horizon/>), Mozilla Hubs (<https://hubs.mozilla.com/>), evaluation of QoE of these systems is necessary. However, these systems lack the integration of collecting implicit metrics, which is the primary motivation behind the development. Our prototype VR collaborative system (see Figure 1) solves this problem by collecting eye-tracking data and physiological signals like the heart

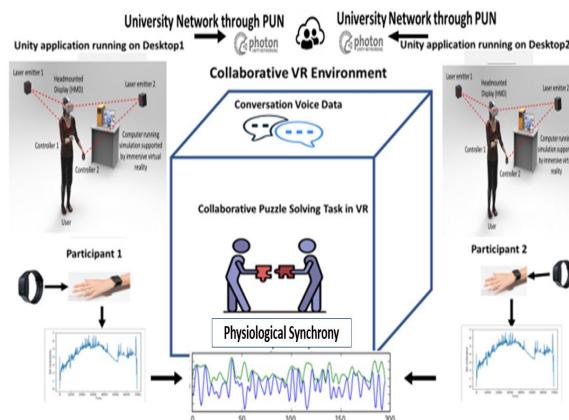


Figure 1: Physiological Synchrony in Virtual Reality

rate, cognitive load, and attention of the collaborators to enable a better mutual understanding. We have proposed a user study with synchrony of physiological signals as an additional metric, for predicting QoE between collaborators. Physiological synchrony serves as an important indicator of QoE in distributed VR collaborative tasks. Inspired by [Dich et al., 2018], we propose to use physiological synchrony as an indicator of collaboration quality, task performance, and learning. By analyzing users' physiological signals synchrony, researchers can optimize the VR experience, improve communication and collaboration, and ultimately provide a more satisfying QoE in an immersive collaborative VR environment.

## 2 State of the Art

In the field of evaluating Quality of Experience (QoE) in Social VR, various approaches are employed to capture users' perceptions, physiological responses, and performance outcomes. Traditionally, user surveys, questionnaires, and interviews utilizing rating scales, open-ended questions, and Likert scales are commonly used to gather subjective feedback on users' perceptions and overall QoE [Keighrey et al., 2020]. Additionally, physiological measures such as heart rate and skin conductance, along with behavioral indicators like body movements and gaze patterns, provide objective insights into arousal, cognitive load, and engagement levels. Objective measures like task completion time and error rates are utilized to evaluate performance in collaborative tasks. In addition to subjective, physiological, and task performance metrics, we propose the concept of physiological synchrony to evaluate QoE in Social VR. This synchrony metric can provide a better understanding of group Quality of Experience (QoE) in Social VR environments. By measuring synchrony in physiological signals and pupil dilation, a comprehensive understanding of group QoE in Social VR environments can be achieved. Integrating subjective assessments, physiological metrics, objective performance measures, and synchrony analysis offers a multi-faceted evaluation approach, enabling a more nuanced assessment of collaborative experiences and team performance in Social VR.

## 3 System Implementation

A collaboration VR system inspired by existing systems like Spatial and Mozilla Hubs was developed. Figure 1 shows an overview of our system. Participants on both sides wear the VR headset and E4 wristband. The prototype was built with Unity 3D Game Engine (2019.3.2f1) and tested on two HTC Vive pro eye (<https://www.vive.com/nz/product/vive-pro-eye/overview/>) VR headsets tethered to two Windows 10 computers. An Empatica E4 wristband for each participant. Physiological signals like EDA, HRV, IBI, BVP were obtained from Empatica Real-time cloud to which all data was uploaded from the wristband memory slot. To enable a framework that supports multiple physiological sensors to be integrated seamlessly, several SDKs were integrated as helper classes into Unity to get physiological data from Users. Photon Unity Networking (PUN) SDK with Photon Cloud [12] through the university network provided multi-user features. Eye tracking data was collected from HTC Vive Pro Eye, using Tobii XR SDK and SRanipal SDK. Audio communication was enabled by Photon Voice and was recorded using Audacity for future analysis as shown in Figure 2.

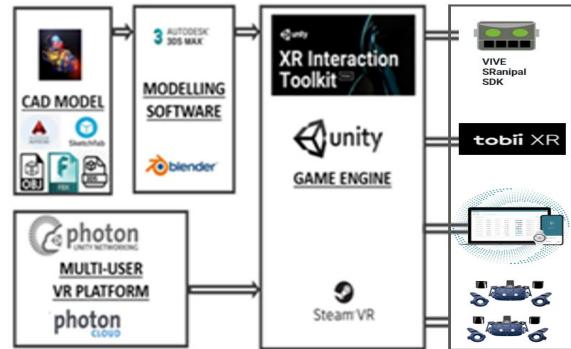


Figure 2: Components used to create the Virtual Collaborative Environment

**Collaborative Task Design in Virtual Environment:** The VR experience consisted of two users connected to each other virtually. Each participant was in a different physical room in the faculty building and was

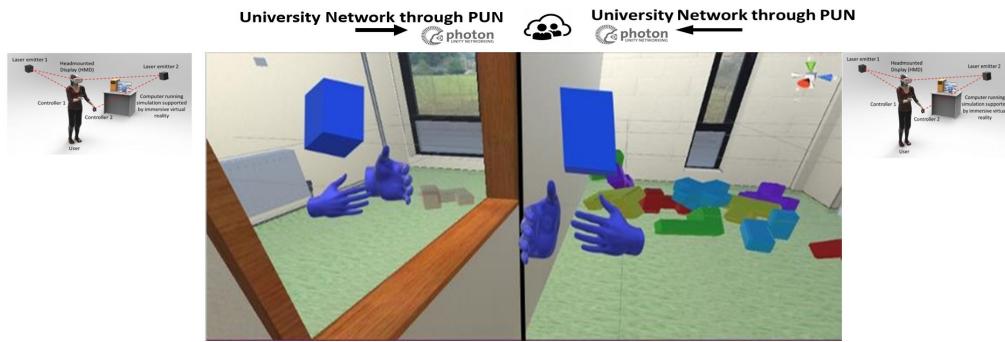


Figure 3: Task in Virtual Reality with Leader and Follower Participant

accompanied by either the PI or research assistant during the experiment. During the task, the follower works with the leader, following their commands and giving requested blocks to the leader. The leader must describe the color and shape to the follower. To correctly solve the task, they must move 10 shapes (in a specific order) from one room (where the follower is based) to another room (where the leader is based). In the leader's room, the leader completes the puzzle. Each user is restricted to moving around only in their respective virtual rooms. The follower can pass the objects through the window to the leader as shown in Fig 3. They move around wearing HTC Vive Pro Eye HMD with hand controllers helping them to manipulate the virtual objects with specific keys.

## 4 User study Evaluation

In our study, we plan to use physiological signals synchrony with pupil dilation synchrony, along with subjective metrics (Presence Questionnaire, SIM-TLX) and task-related metrics (Start time, end time, number of clicks, and number of teleports). In this experiment, there are two groups, one group of Leaders and another group of Followers task role that were shared between the collaborators with factors like Gender, technical expertise, demographics, and familiarity will be considered. In the user study, we would be interested to investigate the following two research questions: 1) Does Physiological Synchrony have any influence on Task performance? 2) How does the physiological synchrony affect groups of similar-gender participants? 3) What are the benefits of grouping experts with no experience users in VR remote collaboration? 4) How do Familiarity and Demographic groups influence synchrony in VR?

- Hypothesis I - Physiological Synchrony of the participant enhances the performance of the remote collaboration system.
- Hypothesis II – Grouping the same gender enhances the synchrony of the remote collaboration system.
- Hypothesis III – Grouping the same experience-level collaborators enhances the synchrony of the remote collaboration system.
- Hypothesis IV – Familiarity with the collaborator enhances the synchrony of the remote collaboration system.

### 4.1 Experiment Procedure.

The experiment would begin with the participants signing a consent form, answering demographic questions, describing their experience with VR/AR, and familiarity with collaborator. Participants would then be introduced to a short video on navigating the Virtual space and the tasks. At the beginning of the session, all participants completed the consent forms. The E4 was fitted to the wrist of the participant's non-dominant hand in accordance with standard practice [33], [34]. It was verified that the E4 was operating correctly, and 5 minutes of resting time was collected to get the baseline readings of all participants before exposing them to

the virtual environment or putting on the HMD. The HMD was fitted, and participants were provided an opportunity to familiarize themselves with the applications and controls (as part of a training phase in a bespoke training environment). Then, participants progressed to the shared virtual environment and worked together on the shared task. After completion, both E4 and the HMD were collected from the users, and post-questionnaires were given to collect subjective feedback from them.

## 4.2 QoE Metrics.

We plan to use a within-subject design between the experience of leader and follower roles during a Collaborative VR Task, we conducted an experimental study using a collaborative VR application. I. For each pair of participants, one would be assigned Leader the other will be assigned Follower. The role was randomly allocated one as Leader and the other as Follower. II. We would collect both objective and subjective measures from each condition. The time for completing the tasks would be recorded in a system log file to objectively measure task performance. At the end of each experiment, the participants would be asked to complete subjective questionnaires (from within the VR environment). We would use the questionnaires: SIM TLX [Harris et al., 2020] and Social Presence Questionnaire customized according to the task.

## 5 Conclusion

A collaborative VR solution that can be integrated with physiological signals synchrony and pupil dilations similarity to enhance the group QoE of collaborative VR applications was presented. Measuring the synchrony of physiological signals of the participants under different grouping conditions was proposed to enhance the QoE of collaborative VR. We plan to provide our findings from development, and user testing in the future.

## Acknowledgments

This research is funded by the Technological University of Shannon, Athlone Ireland with the Science Foundation Ireland CONFIRM Centre (grant number: 16/RC/3918) and is also partially supported via the Horizon Europe TRANSMIXR project (grant number: 101070109)

## References

- [Dich et al., 2018] Dich, Y., Reilly, J., and Schneider, B. (2018). Using physiological synchrony as an indicator of collaboration quality, task performance and learning. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19*, pages 98–110. Springer.
- [Harris et al., 2020] Harris, D., Wilson, M., and Vine, S. (2020). Development and validation of a simulation workload measure: the simulation task load index (sim-tlx). *Virtual Reality*, 24(4):557–566.
- [Keighrey et al., 2020] Keighrey, C., Flynn, R., Murray, S., and Murray, N. (2020). A physiology-based qoe comparison of interactive augmented reality, virtual reality and tablet-based applications. *IEEE Transactions on Multimedia*, 23:333–341.
- [Saffo et al., 2021] Saffo, D., Di Bartolomeo, S., Yildirim, C., and Dunne, C. (2021). Remote and collaborative virtual reality experiments via social vr platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- [Vlahovic et al., 2022] Vlahovic, S., Suznjevic, M., and Skorin-Kapov, L. (2022). A survey of challenges and methods for quality of experience assessment of interactive vr applications. *Journal on Multimodal User Interfaces*, 16(3):257–291.

# Will your Doorbell Camera still recognize you as you grow old?

Wang Yao<sup>1</sup>, Muhammad Ali Farooq<sup>1</sup>, Joseph Lemley<sup>2</sup>, and Peter Corcoran<sup>1</sup>

<sup>1</sup>School of Engineering, University of Galway, Ireland.

<sup>2</sup>Xperi Corporation, Galway.

## Abstract

Robust authentication for low-power consumer devices such as doorbell cameras poses a valuable and unique challenge. This work explores the effect of age and aging on the performance of facial authentication methods. Two public age datasets, AgeDB and Morph-II have been used as baselines in this work. A photo-realistic age transformation method has been employed to augment a set of high-quality facial images with various age effects. Then the effect of these synthetic aging data on the high-performance deep-learning-based face recognition model is quantified by using various metrics including Receiver Operating Characteristic (ROC) curves and match score distributions. Experimental results demonstrate that long-term age effects are still a significant challenge for the state-of-the-art facial authentication method.

**Keywords:** Synthetic Data, GAN, Age Effect, Face Recognition, Deep Learning

## 1 Introduction

Automatic face authentication/recognition is one of the active and long-standing research topics in the field of computer vision. Studies [Taskiran et al., 2020] have shown that factors such as age, lighting, and pose can have a significant impact on the performance of face recognition algorithms. Recent work [Yao et al., 2022] has studied the main factors such as illumination and pose that affect facial authentication, which shows the feasibility of implementing a robust authentication method for low-power consumer devices. In this work, we are focusing on another challenging factor, age, i.e., how age bias affects the state-of-the-art Face Recognition (FR) method.

Age progression as a basic demographic can be used for real-time biometric authentication applications such as doorbell cameras and FaceID for cell phones. It is necessary to obtain sufficient aging data from diverse identities to study the robustness of biometric authentication systems to age effects. Figure 1 shows an example of the aging effects of a celebrity. However, collecting data on a person from birth to old age is difficult, and existing age datasets rarely have large amounts of face data from multiple identities at different stages of age. In this work, a generative adversarial network (GAN) based aging technique has been used to create a synthetic aging dataset, which could generate different target ages of an identity. In this way, the robustness of the FR model to aging can be verified without a need to collect data from human subjects over long periods of time. The main contributions of this paper are as follows.

- (1) This work quantifies the long-term aging effect on state-of-the-art neural face recognition algorithm.
- (2) This work explores the feasibility of using synthetic aging data to augment real-world age data.
- (3) Experimental results show that large age intervals cause obvious degradation in the FR algorithm.

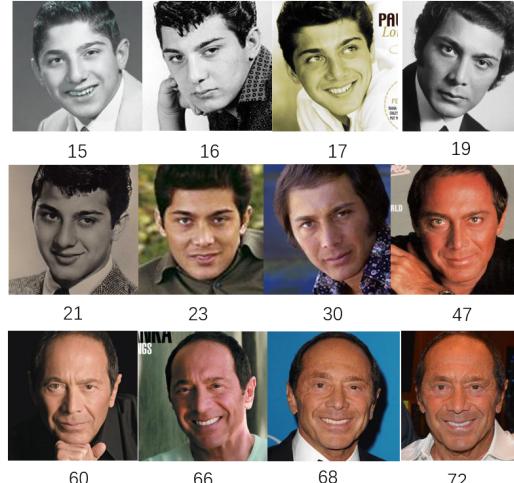


Figure 1: Example of an individual facial variation during the aging process (Images from [Moschoglou et al., 2017]).

## 2 Background

The development of face authentication technology in recent years has resulted in the emergence of many state-of-the-art FR algorithms based on deep learning. These algorithms have achieved remarkable performance improvements on large-scale data sets by optimizing deep neural networks and loss functions. Studies have shown that the performance of face recognition is affected by demographics including age, race, and gender which can produce bias. For example, [Deb et al., 2017] present a longitudinal study of FR and find a significant decrease in the accuracy of COTS-A and COTS-B face matchers over a time interval of 8.5 years. [Boussaad and Boucetta, 2021] find that the recognition rate becomes significantly low when age intervals are greater than 20 years. These studies always adopted feature extraction-based FR methods. This raises the following question in this study. How robust is the recent deep learning-based FR method to age bias?

## 3 Methodology

**Datasets:** Two publicly available age datasets AgeDB and Morph II are adopted in this work to study the age effect on the FR algorithm. (a) AgeDB [Moschoglou et al., 2017] contains 568 subjects with 16488 images. The average number of images per subject is 29. (b) Morph II [Ricanek and Tesafaye, 2006] contains 55134 images with 13618 identities taken from 2003 to 2007. Each subject has an average of 4 images.

**Synthetic Aging Method:** SAM [Alaluf et al., 2021], as one of the state-of-the-art synthetic aging techniques is adopted in this work to generate more aging samples. SAM treats the aging process as an image-to-image transformation problem by pairing a pre-trained fixed StyleGAN generator with an encoder. The task of the encoder is to encode real face images directly into the StyleGAN latent space to obtain the faces with expected age change. Some synthetic aging samples are shown in Figure 2.

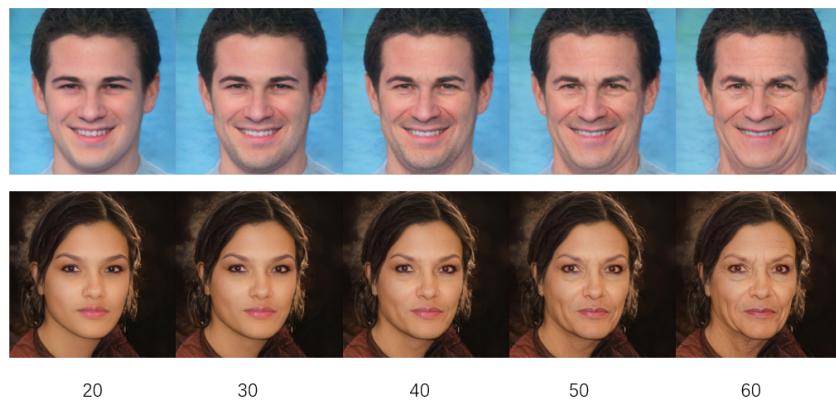


Figure 2: Samples of age variations using the synthetic aging method.

**Face Recognition Method:** ArcFace [Deng et al., 2019] is selected as the FR model to verify the validity of synthetic aging faces, which uses additive angular margin loss to obtain good intra-class compactness and inter-class dispersion. In this work, MTCNN is employed to detect and crop the faces. Then, these faces are fed into the FR network to obtain facial features. The identity similarity score is derived by calculating the cosine similarity of the two facial features.

**Evaluation Metrics:** Two evaluation metrics including the receiver operating characteristic (ROC) curve and match score distribution have been adopted in our experiment to quantify the effect of synthetic aging data on the high-performance deep-learning-based FR model.

## 4 Experiment

**Experimental Setting:** The process of this experiment is shown in Figure 3. First, all detectable face images from 20 to 30 years old are selected for pre-processing to obtain  $256 \times 256$  images. Then, these face images are taken as the original images to synthesize faces of different ages, and here we synthesize face images of 20, 30, 40, 50, 60, 70, and 80 years old respectively. Moreover, positive-identity pairs (PPs) and negative-identity pairs (NPs) are formed by using the original images. A positive-identity pair means an image pair from the same identity. A negative-identity pair means an image pair from different identities. For different age intervals, one of the images from PPs/NPs is replaced separately with the images of the target age. Finally, the images from PPs/NPs are fed into the FR model to calculate the similarity scores, which are used for evaluation.

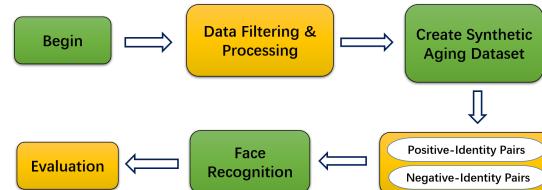


Figure 3: Experimental Process.

#### 4.1 ROC Comparison

The ROC curves in Figure 4 quantify the effect of different age intervals on the FR algorithm. Age intervals within 20 years have a weak effect on the performance of the state-of-the-art FR model. There is a slight degradation of the FR model performance for a 30-year time interval. Longer time intervals still cause significant degradation in the performance of the FR model.

#### 4.2 Match Score Comparison

Figure 5 shows the impostor and genuine distributions of different age intervals. The impostor distributions are similar across different age intervals, while the genuine distributions with a noticeable shift to the left as the age intervals get larger. As the age interval increases, the area where the impostor distribution intersects with the true distribution increases, and the recognition performance decreases. This indicates that the genuine distribution is the main reason leading to the degradation of the performance of the FR algorithm.

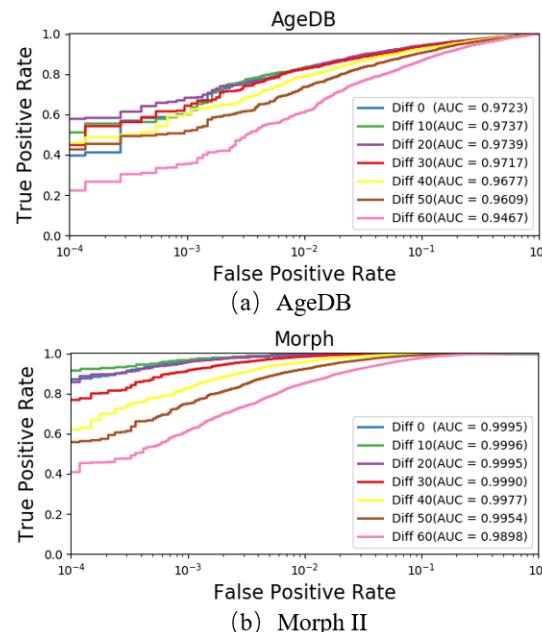


Figure 4: The ROC curves for ArcFace of AgeDB and Morph II.

### 5 Conclusion

This work qualifies the impact of age intervals on a SOTA FR algorithm and illustrates the potential value of synthetic age data in analyzing the robustness of face authentication systems. Initial experiments associated with synthetic age data have shown that it is valid to use synthetic data to enlarge the dataset for analysis of how aging affects FR models. The evaluation results from ROC curves and match score comparison show that age differences within 20 years do not have a noticeable impact on face recognition accuracy, while long-term age differences remain a significant challenge for the current facial authentication method. In future work, we will explore the quality of synthetic age data, and do a comparison of the real age data and the synthetic age data. We plan to expand this work to create a synthetic age dataset and to implement a privacy-secured robust face authentication solution on low-power neural accelerators.

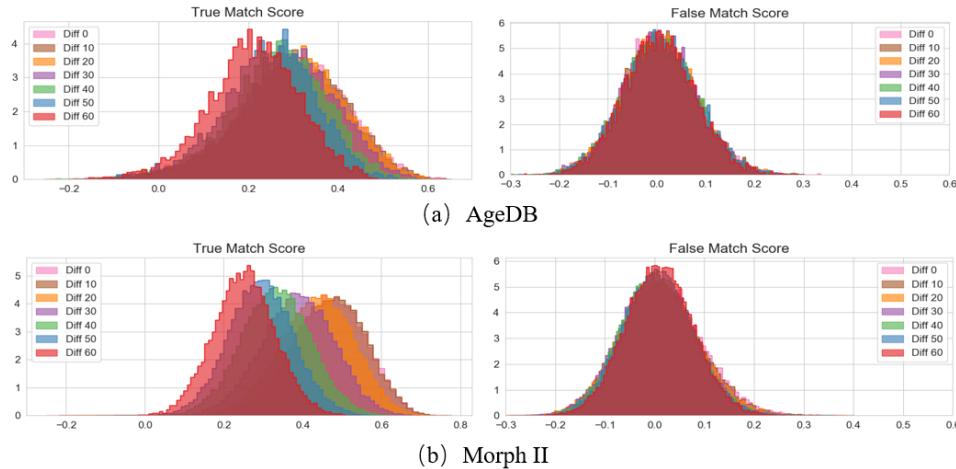


Figure 5: The match score distributions for ArcFace of AgeDB and Morph II.

## Acknowledgments

This research is supported by (i) Irish Research Council Enterprise Partnership Ph.D. Scheme (Project ID: EPSPG/2020/40), (ii) Xperi Corporation, Ireland, and (iii) the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF).

## References

- [Alaluf et al., 2021] Alaluf, Y., Patashnik, O., and Cohen-Or, D. (2021). Only a matter of style: Age transformation using a style-based regression model. *ACM Trans. Graph.*, 40(4).
- [Boussaad and Boucetta, 2021] Boussaad, L. and Boucetta, A. (2021). The aging effects on face recognition algorithms: the accuracy according to age groups and age gaps. In *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)*, pages 1–6.
- [Deb et al., 2017] Deb, D., Best-Rowden, L., and Jain, A. K. (2017). Face recognition performance under aging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–54.
- [Deng et al., 2019] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.
- [Moschoglou et al., 2017] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: The first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- [Ricanek and Tesafaye, 2006] Ricanek, K. and Tesafaye, T. (2006). Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345.
- [Taskiran et al., 2020] Taskiran, M., Kahraman, N., and Erdem, C. E. (2020). Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106:102809.
- [Yao et al., 2022] Yao, W., Varkarakis, V., Costache, G., Lemley, J., and Corcoran, P. (2022). Toward robust facial authentication for low-power edge-ai consumer devices. *IEEE Access*, 10:123661–123678.

# Defect Classification in Additive Manufacturing Using CNN-Based Vision Processing

Xiao Liu, Alessandra Mileo and Alan F. Smeaton

*Dublin City University, Glasnevin, Dublin 9, Ireland*  
*alan.smeaton@dcu.ie*

## Abstract

The development of computer vision and in-situ monitoring using visual sensors allows the collection of large datasets from the additive manufacturing (AM) process. Such datasets could be used with machine learning techniques to improve the quality of AM. This paper examines two scenarios: first, using convolutional neural networks (CNNs) to accurately classify defects in an image dataset from AM and second, applying active learning techniques to the developed classification model. This allows the construction of a human-in-the-loop mechanism to reduce the size of the data required to train and generate training data.

**Keywords:** Convolutional neural networks, additive manufacturing, defect classification, active learning

## 1 Introduction

Large and openly available datasets of annotated images containing up to millions of training examples such as Pascal VOC [Everingham et al., 2010] are available to machine learning researchers for many applications. This has enabled huge improvements in machine learning over recent years. By contrast, such openly available datasets are not available in the domain of Additive Manufacturing (AM) or 3D printing because labelled samples are difficult, expensive, and time-consuming to obtain as shown in [Qin et al., 2022] and [Wang et al., 2020]. As a result of poor data availability, researches in AM often have to use only a limited amount of labelled samples for training tasks before then leveraging a large number of unlabelled image data. Some researchers have called this the “small data challenge in the big data era” [Qi and Luo, 2020].

To overcome this challenge, we present a method that applies transfer learning and fine-tuning on a CNN-based neural network model to achieve accurate classification of manufacturing defects. This uses a dataset of images of the melt pool, created from the interaction between a laser and the materials used in manufacturing, taken during the AM process. Structural defects in the resulting output can sometimes be detected during manufacture from observations of the melt pool. Our technique involves using active learning algorithms to reduce the number of labelled samples required in the training process. We perform automatic labelling using the model to generate larger datasets of labelled images from unlabelled samples, for use in training.

## 2 Methods

*Transfer learning* is a method that performs training a neural network model using data from a source domain then later applying the trained model to a target domain that is different from the source. This allows rapid progress in re-training and significantly reduces the required number of training samples in the target domain. It is commonly used in computer vision tasks such as classification to support improved performance in domains which are data-poor. In recent years, transfer learning has proved to be effective in the task of defect classification in AM, such as the work presented in [Liu and Mileo, 2021] and [Westphal and Seitz, 2021] where transfer learning and fine-tuning were applied to the training of CNN based neural network architectures.

*Active learning* [Settles, 2009] is a technique for labelling data that selects and prioritises the most informative data points to submit to an annotator for labelling. Such prioritised data points have the highest potential impact on the supervised training of a machine learning model, thus accelerating the training process. The combination of transfer learning and active learning allows leveraging small amounts of labelled data to improve the performance of the training process of a deep learning model.

### 3 Classification Experiments in Additive Manufacturing

To investigate the potential for transfer learning and active learning in the task of defect detection in AM, a case study was carried out using the open image dataset from [Westphal and Seitz, 2021]. This contains 4,000 images, manually divided into 2 different defect detection classes in AM. The images in this dataset are clearly separated into 3 balanced subsets for training (2,000), testing (1,000) and validation (1,000).

To conduct experiments, we employed a VGG16 based classifier from previous work which proved to be accurate in the task of defect classification on images generated from emission monitoring during additive manufacturing [Liu et al., 2022]. This classifier relies on transfer learning in which 13 convolutional layers from a pre-trained VGG16 model are used for feature extraction and the weights in these layers had been trained using ImageNet data. After the convolutional layers, 2 dense layers with ReLU activation function are added followed by 1 dense layer as the output layer using Sigmoid as the activation function, since the targeted dataset are divided into 2 classes for binary classification. In the original paper [Westphal and Seitz, 2021], the best classification performance is generated using a VGG16 based CNN model which is the reason we do not use a more recent model such as ResNet. We consider that as a baseline for further investigation in this study.

The tuning of hyperparameters involves adjusting the optimiser, learning rate, batch size and training epochs. There are 3 optimisers in the test we use which are Adam, SGD and RMSprop in combination with learning rate in a range from  $10^{-2}$  to  $10^{-5}$ . We have also conducted training using different batch sizes (4, 32, 64) and training epochs (30, 60, 120). The cost function used in all tests is binary cross entropy. To reduce overfitting, weight regularisers are added to the 2 dense layers with the ReLU activation function mentioned above. The weight decay regulariser, also known as L2 regulariser which calculates the sum of the squared weights, is applied when initialising the keras model. The tuning of this hyperparameter is in a range from  $10^{-1}$  to  $10^{-4}$  and tested for multiple times until no obvious overfitting issue appears in the training and validation.

After tuning on hyperparameters for multiple combinations, the best performing combinations regarding the 3 types of optimisers are shown in Table 1 together with classification results on the validation dataset in comparison with the baseline from [Westphal and Seitz, 2021]. These initial tests were performed to check how adaptive our approach is on this dataset. The results show that all 3 optimisers can reach a value around 98% of the validation accuracy and our classification model is well-adapted to this dataset. The results also show that for this dataset a smaller batch size used in the training process such as 4, gives better performance and this can be explained as smaller batch sizes require more frequent weight updates during training. In turn this can help the model adjust its parameters more quickly and respond to changes in the data distribution which increases the model's ability to adapt to a new dataset. Finally, although not shown here, accuracy is stable throughout the training showing no overfitting.

### 4 Active Learning Experiments in Additive Manufacturing

Having developed a classifier which uses domain transfer across AM image datasets, we extended training to include active learning applied to further investigate classification performance during the progression of AL iterations. The second experiment was performed in a series of steps of (1) active sample section, (2) query for label, (3) train with queried sample, and (4) validate for current query iteration. The cycle iterates until a human supervisor decides to complete the training phase when validation accuracy achieves a target level.

Here we apply a pool-based sampling scenario and an uncertainty sampling query strategy [Settles, 2009]. This is the most commonly used query strategy to start generalised sampling on this particular AM dataset.

Table 1: Best performing hyperparameters for each optimiser, performance results on the validation set. Results marked ‘\*’ are updates provided directly to us by the authors of [Westphal and Seitz, 2021] in response to us pointing out errors in the original paper. An author correction to the copy of record is now underway.

Experiment: Optimiser, learning rate	Batch	Epochs	Confusion matrix	Accuracy	Precision	Recall	F1-Score	AUC				
Baseline	64	30	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>496</td><td>4*</td></tr> <tr><td>19</td><td>481</td></tr> </table>	496	4*	19	481	0.977*	0.992*	0.963*	0.977*	0.993
496	4*											
19	481											
SGD, lr=0.01	4	60	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>483</td><td>17</td></tr> <tr><td>4</td><td>496</td></tr> </table>	483	17	4	496	0.979	0.967	0.992	0.979	0.998
483	17											
4	496											
Adam, lr = 0.00001	4	120	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>490</td><td>10</td></tr> <tr><td>2</td><td>498</td></tr> </table>	490	10	2	498	0.988	0.980	0.996	0.988	0.998
490	10											
2	498											
RMSprop lr = 0.00001	4	60	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>485</td><td>15</td></tr> <tr><td>3</td><td>497</td></tr> </table>	485	15	3	497	0.982	0.971	0.994	0.982	0.997
485	15											
3	497											

The implementation of active learning uses Python 3 and Google Colab. During the experiment, a classifier model is initialised and the optimiser chosen is SGD as we found this gives more stabilised performance in the validation test and has minimal overfitting even when training is continued long after convergence. While Adam and RMSprop converge faster, there are larger fluctuations in the validation and minor overfitting after training reaches convergence. In addition, though SGD yields a result lower than the other 2 optimisers, it has slightly better potential that can be improved by applying active learning. During this experiment, 2,000 training samples were fed to the classifier with a total of 40 queries and for each query 50 samples were actively selected by the uncertainty sampling query strategy.

The selected and queried samples were assigned a label by an annotator after which the newly labelled samples were used to fine-tune the classifier to improve performance. This was evaluated using classification accuracy on the validation dataset at the end of each query iteration and later we show results on the test set.

Following the inclusion of active learning, validation accuracy in each query iteration is shown in Figure 1 where results show that with the aid of active learning, the model reaches convergence after the 13th query and the value of validation accuracy is around 98%. More specifically, the calculated mean value from the 13th to 40th queries is 0.981 with a standard deviation of 0.0246 and a peak of 0.990. This is slightly higher than the result of the SGD optimiser based model shown in Table 1 and 1% higher than the baseline. Overall performance after convergence is also relatively stable. Results also show that the model only needs the first 650 most informative samples to achieve the best performance which is only 37.5% of the total 2,000 labelled training data. This trained model was used to classify the labels on the testing dataset mentioned in Section 3,

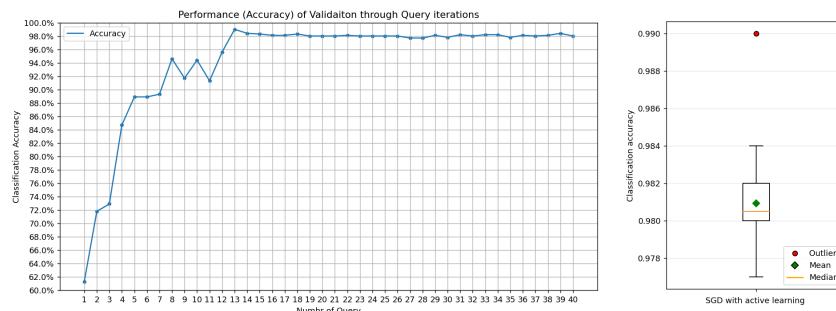


Figure 1: classification accuracy on the validation dataset at the end of each query iteration

which is a balanced dataset consisting of 1,000 samples and the results are shown in Table 2.

Table 2: Predicted results for auto-labeling on the test dataset with P, R and F1 calculated for each of 2 classes

Confusion matrix	Precision	Recall	F1-Score	AUC	Accuracy
487   13	0.994	0.974	0.984	0.998	0.984
3   497	0.975	0.995	0.984		

## 5 Conclusions

This paper presents an investigation into performance of a computer vision based classification task on a dataset from the additive manufacturing process. We use a CNN based classifier in combination with transfer learning and active learning strategies. We improved the overall validation accuracy to about 98%. We also conducted experiments to investigate the approximate minimum number of labelled samples needed to reach convergence in training. In future work we plan to further investigate the sampling strategies for active learning especially regarding class imbalance problems. We will involve approaches from semi-supervised learning to reinforce the labelling and self training as an extension to the current active learning mechanism.

**Acknowledgements:** XL is funded by SFI 16/RC/3872 at I-Form, the SFI Research Centre for Advanced Manufacturing and AM and AS are part-funded by SFI 12/RC/2289\_P2 at Insight the SFI Research Centre for Data Analytics at DCU.

## References

- [Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88:303–338.
- [Liu and Mileo, 2021] Liu, X. and Mileo, A. (2021). A deep learning approach to defect detection in additive manufacturing of titanium alloys. In *2021 International Solid Freeform Fabrication Symposium*. University of Texas at Austin.
- [Liu et al., 2022] Liu, X., Smeaton, A. F., and Mileo, A. (2022). An adaptive human-in-the-loop approach to emission detection of additive manufacturing processes and active learning with computer vision. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4021–4027. IEEE Computer Society.
- [Qi and Luo, 2020] Qi, G.-J. and Luo, J. (2020). Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2168–2187.
- [Qin et al., 2022] Qin, J., Hu, F., Liu, Y., Witherell, P., Wang, C. C., Rosen, D. W., Simpson, T. W., Lu, Y., and Tang, Q. (2022). Research and application of machine learning for additive manufacturing. *Additive Manufacturing*, 52:102691.
- [Settles, 2009] Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.
- [Wang et al., 2020] Wang, C., Tan, X., Tor, S., and Lim, C. (2020). Machine learning in additive manufacturing: State-of-the-art and perspectives. *Additive Manufacturing*, 36:101538.
- [Westphal and Seitz, 2021] Westphal, E. and Seitz, H. (2021). A machine learning method for defect detection and visualization in selective laser sintering based on convolutional neural networks. *Additive Manufacturing*, 41:101965.

# The Impact of Glare on End of Production Line Camera Calibration Algorithms: A Brief Analysis

Payal Bhattacherjee<sup>1</sup>, Anbucheziyan Selvaraju<sup>1</sup>, Sudarshan Paul<sup>1</sup>, Arindam Das<sup>1,2</sup>, Ishan Vermani<sup>3</sup>

<sup>1</sup>DSW, Valeo India, <sup>2</sup>University of Limerick, Ireland, <sup>3</sup>CDV, Valeo Germany

## Abstract

Camera calibration is vital for accurate measurement and analysis of visual data in computer vision and image processing, particularly in applications like automated driving. However, glare in images presents significant challenges, potentially leading to compromised performance and inaccurate results. Glare commonly occurs in production line environments due to factors like overhead lighting, sunlight, reflective surfaces, and polished materials. It manifests as overexposed regions, distorting image features and affecting calibration parameters such as camera intrinsic/extrinsic parameters, lens distortion coefficients, and image rectification parameters. Precise calibration algorithms relying on feature extraction and correspondence estimation face difficulties in the presence of glare. This study investigates the impact of glare on camera calibration, addressing specific issues and proposing potential solutions. The research focuses on camera extrinsic calibration and aims to mitigate the adverse effects of glare. Lighting conditions and quality control play crucial roles in ensuring optimal performance in production line environments.

**Keywords:** Camera Calibration, Glare, Gradient Filter, Line detection

## 1 Introduction

Camera calibration is essential for accurate measurements, object recognition, and 3D reconstruction in computer vision. Camera extrinsic calibration determines camera position and orientation relative to the ego-vehicle's surroundings. It is crucial for tasks like perspective transformation and object detection. Initial camera setups in factory-mounted vehicles have errors and require calibration near the end of the assembly line. Deviations in camera angles ( $dx$ ,  $dy$ ,  $dz$ ) and height are detected using target-based routines. However, varying lighting conditions and glare on the targets at the end of the assembly line pose challenges due to high dynamic range images. The impact of glare on camera calibration algorithms is multifaceted. Firstly, glare can cause the loss of crucial image features or introduce spurious features, which are essential for accurate correspondence estimation. These erroneous correspondences can result in distorted calibration parameters, leading to inaccurate 3D reconstructions or object measurements. Secondly, glare-induced overexposure can disrupt the distribution of pixel intensities. This, in turn, can lead to geometric distortions further propagating errors in subsequent image processing tasks. Finally, glare can distort or obscure calibration patterns, reducing

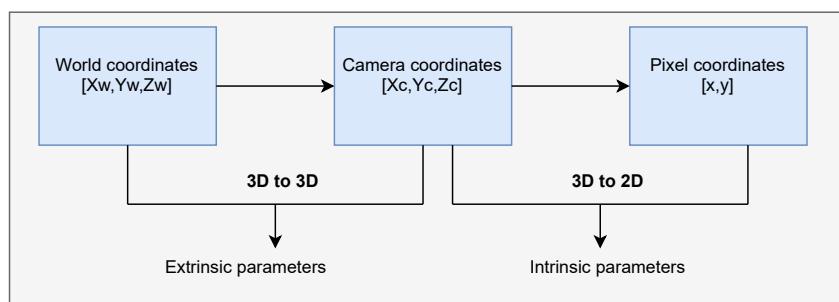


Figure 1: Extrinsic and Intrinsic Parameters for Calibration

the visibility of calibration markers or altering their shapes. This poses a significant challenge to calibration algorithms that rely on precise detection and localization of these patterns. When lighting conditions result in glare, it often manifests as bright spots or regions of high luminosity in the visual field. This can impede machine vision systems.

Mehran et al [Andalibi and Chandler, 2017] proposed a comprehensive system that combines photometric, geometric, and global positioning information to automatically detect glare. Singh et al [Singh et al., 2016] deploy a process involving selecting higher intensities in the image, creating a mask N, performing connected component analysis, and selecting the maximum connected components to isolate the glare region from the foreground. Ghulam Qadir et al [Qadir et al., 2011] proposed a novel approach for identifying unbalanced lighting in images using Benford's Law in conjunction with Discrete Wavelet Transform (DWT) to determine if an image exhibits unbalanced lighting.

## 2 Methodology

3D position and orientation of the camera optical axis with vehicle datum and ground plane vary from vehicle to vehicle and have to be estimated at the end of assembly line after the cameras have been mounted. Calibration routine for estimating the error in camera mounting is comprised of detecting a calibration pattern. In the current scope of the paper, a line based target pattern is used. Two individual Mats are placed parallel to the vehicle on the ground to calibrate surround view cameras. The pattern comprises of two parallel white lines and two vertical white lines on black background.

Calibration pattern detection gets highly impacted with glare which introduces additional intensity variations in the image that can impact the gradient information. The current algorithm used relies heavily on sobel edge detection, edge strength and edge orientation respectively. At first, edges whose magnitude is above the specified threshold and whose gradient direction is within the specified angular tolerance are extracted. Qualified edges of similar direction are grouped into labeled objects by 8-connectivity. Minimum and maximum number of pixels in a run is defined to qualify edges as part of the line segment. Thresholds are carefully tuned to qualify a detected endpoint as a line. Factory conditions causing glare, impact the above thresholds causing line detection failure leading to no calibration (no target found) or wrongly detecting a line segment as target(reflection on the car body) which lead to high deviations from the known nominal. This kind of behavior at the production plant causes wrong estimations of the extrinsic values. With wrong extrinsic real world positions of the detected objects will be shifted from the actual foot point of the object in the real world.

To compensate for the glare a methodology is added in the calibration pipeline. Glare detection involves the utilization of a combination of features that are computationally efficient yet highly effective in detecting both the location and size of glare regions. These features include the intensity, saturation, and local contrast of the input frame. Glare regions tend to be the brightest regions in an image I, and thus light intensity can be a useful feature. Value component from the HSV color-space is used to obtain the gray scale intensity of each pixel. Let V denote the intensity map. V is computed as follows:

$$V(x, y) = \max_{(x,y)} I_R(x, y), I_G(x, y), I_B(x, y) \quad (1)$$

where  $I_R$ ,  $I_G$ , and  $I_B$  denote the red, green, and blue channels of I.

Regions with low color saturation are indeed often indicative of glare regions. By extracting the saturation component(s) for each pixel, we can effectively identify areas in the image that lack vibrant colors and exhibit low saturation, which are potential candidates for glare regions.

$$S_{(x,y)} = \begin{cases} \frac{V(x,y) - \min_{x,y}\{I_R(x,y), I_G(x,y), I_B(x,y)\}}{V(x,y)} & \text{for } V(x, y) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Regions with very low luminance contrast are also good candidates of glare regions. To measure the local contrast, image is divided into  $(17 \times 17)$  overlapping macro blocks. Within each block, we quantify the root

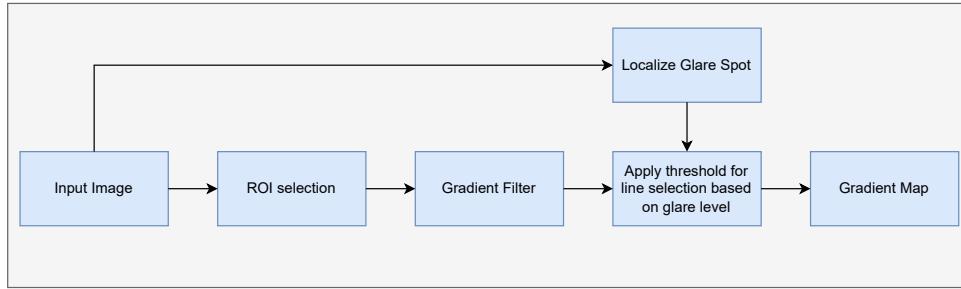


Figure 2: Our proposed method for obtaining Gradient Map

mean square (RMS) contrast by calculating the ratio of the standard deviation of the luminance values within the block to the mean of the luminance values within the same block. This measure provides an indication of the variation in luminance within the local region, relative to its average luminance level.

$$RMSContrast = std(L(x', y')) / mean(L(x', y')) \quad (3)$$

Here,  $L(x', y')$  represents the luminance values within the block ( $17 \times 17$ ) centered at  $(x, y)$ ,  $std()$  denotes the standard deviation, and  $mean()$  represents the mean value of the luminance within the same block. Dividing the standard deviation by the mean gives us a measure of the contrast within the local region.

$$Gmap = I(x, y) * 1 - s(x, y) * 1 - RMSContrast(x, y) \quad (4)$$

A glare map is generated and normalized within the range of [0,1], allowing for the localization of regions with high intensity, low color saturation, and low contrast. Once the glare spot is identified, the filtering thresholds for both glare and non-glare regions are calculated as follows.

$$Thresh = \begin{cases} 20 & \text{avg}(G_{mag}(x', y')) \text{ if } G_{map}(x, y) > 150 \\ 40 & \text{else} \end{cases} \quad (5)$$

$G_{mag}(x', y')$  is the  $5 \times 5$  block taken at the glare region centered at  $(x, y)$ , this  $5 \times 5$  block is moved across the entire glare region in an overlap manner. This  $5 \times 5$  window size can be selected based on the application. The weak edges are filtered by applying the threshold (Thresh) and gradient map is obtained.

$$\delta x(x, y) = I(x, y+1) - I(x, y-1)$$

$$\delta y(x, y) = I(x+1, y) - I(x-1, y)$$

$$Magnitude = abs(\delta x) + abs(\delta y)$$

$$Direction = [-\pi, \pi] = atan2(\delta y, \delta x)$$

Where,  $I$  represents the gray scale input image.  $\delta x$  represents the gradient in horizontal direction,  $\delta y$  represents the gradient in vertical direction.

### 3 Experimentation Details

The objective of the study was to evaluate the effectiveness of glare detection localization and dynamic threshold updates on line detection. The proposed methodology was tested on various lightning conditions causing line failures. As shown in Figure 3(i) highlighted a vertical line on the mirror right (MR) camera that was initially undetected. The line filtering thresholds were dynamically calculated and glare spots were detected which resulted in significant improvements. Figure 3(ii) demonstrates successful detection of vertical lines, resulting in reduced height deviations (dz mm) from the existing measurements to the proposed ones in the mirror right camera. The height deviation decreased from -1219mm to -26mm, and a similar reduction was observed in the x-degree rotation, decreasing from -35.98° to 0.24° as indicated in Table 1, Trail 1 column.

Camera	Trail 1						Trail 2						Trail 3										
	Existing Method			Proposed Method			Existing Method			Proposed Method			Existing Method			Proposed Method							
	$dx^\circ$	$dy^\circ$	$dz^\circ$																				
FV	-0.63°	-2.10°	-0.90°	-311	-1.53°	-2.03°	-1.12°	-62	1.80°	-0.46°	0.73°	-25	1.60°	-0.24°	0.97°	26	0.87°	-3.04°	0.43°	-30			
RV	1.20°	-2.20°	1.12°	-421	-0.33°	-2.17°	0.87°	-6	0.43°	-3.03°	0.07°	-145	1.07°	-2.93°	0.58°	-54	0.31°	-2.12°	0.07°	-28			
ML	-9.44°	2.28°	0.22°	-537	-0.96°	-0.24°	1.44°	-29	1.06°	1.90°	-0.13°	-165	0.31°	-1.41°	2.89°	5	-0.57°	1.13°	1.55°	24			
MR	-35.98°	1.45°	8.64°	-1219	0.24°	-0.28°	-0.03°	-26	1.11°	-0.81°	-0.76°	-183	1.61°	-1.31°	-0.41°	-4	1.77°	-0.26°	-1.22°	3			
																				1.74°	-0.24°	-1.21°	1

Table 1: Deviation in orientation ( $dx^\circ, dy^\circ, dz^\circ$ ) and height(dz) from Nominal - existing method vs. proposed method.

Figure 3(iii), images were collected from a different station with a distinct glare level, resulting in false line detection due to reflections on the car body. The failed horizontal line detection was clearly highlighted within a square red box. However, by detecting glare and updating the thresholds online, the false detection was eliminated. Moreover, the height deviations on the mirror left (ML) camera decreased from -165mm to 5mm, as demonstrated in Table 1, Trail 2. The orientation deviation includes  $dx^\circ$ ,  $dy^\circ$ , and  $dz^\circ$ , while  $dz$  represents the deviation in camera height from the ground in millimeters. Furthermore, Trail 3 in Figure 3(v) shows a different glare condition that affected horizontal line detection. Nevertheless, the deviations in height and rotation were effectively corrected. The study successfully addressed issues such as the loss of crucial image features, spurious changes in intensity variation, and obscuring of calibration patterns. As

a result of these interventions, the deviations in orientation from the CAD nominal values were reduced within a threshold of  $\pm 5$  degrees, and the height deviations were reduced to 100mm, similar to non-glare regions.

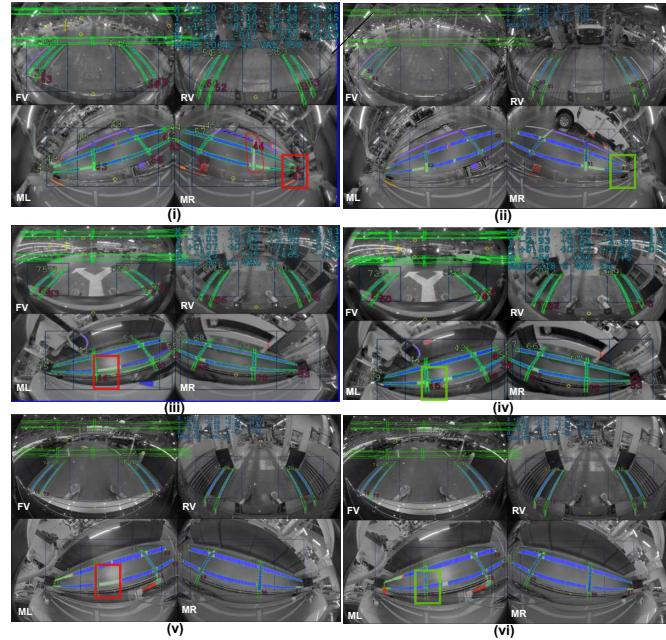


Figure 3: Left - detections impacted by glare, Right - improved detections

## 4 Conclusion

In this work, impact of glare on end of line camera calibration is studied. It was found that in factory conditions it is almost impossible to maintain homogeneous lightning conditions and avoid extreme gradients on the target. In order to mitigate end of line calibration errors like unable to detect target or excess deviation due to false line detections can be avoided by estimating the glare region in the image and apply a modified edge selection threshold in that area .The calibration accuracy is maintained to  $\pm 5$  deg from known CAD rotation and 100mm height is maintained which is similar to non glare scenarios.

## References

- [Andalibi and Chandler, 2017] Andalibi, M. and Chandler, D. M. (2017). Automatic glare detection via photometric, geometric, and global positioning information. *Electronic Imaging*, 29(19):77–77.
- [Qadir et al., 2011] Qadir, G., Zhao, X., Ho, A. T., and Casey, M. (2011). Image forensic of glare feature for improving image retrieval using benford’s law. In *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pages 2661–2664.
- [Singh et al., 2016] Singh, M., Tiwari, R. K., Swami, K., and Vijayvargiya, A. (2016). Detection of glare in night photography. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 865–870.

# Automatic Archery Scoring System Using Deep Learning and Image Processing

Haozhe Ma, Michael G. Madden

*School of Computer Science, University of Galway*

## Abstract

This paper proposes an automatic archery scoring system that uses smartphone to read and record scores, addressing the limitations of existing manual and hardware-based scoring systems. Utilizing advancements in computer vision technology, the system consists of two main components: arrow localisation and arrow scoring. The localisation model uses a modified U-Net architecture to accurately identify the landing positions of arrows on a target, generating a saliency map that is converted into pixel coordinates essential for scoring. A custom dataset, which is released to public, was created to train this model. The scoring component identifies the target face using colour masking and applies image processing techniques for noise reduction and skew correction. It then matches the coordinates from the localisation model to the parameters of each scoring ring to calculate scores. While the model shows progressive improvement during training, further diversification of the dataset, exploration of alternate neural network architectures, and optimisation of hyperparameters can enhance accuracy. Future work includes improving front-end development for seamless functionality integration and adopting augmented reality (AR) for improved perspective correction.

**Keywords:** Deep learning, computer vision, image analysis, target recognition.

## 1 Introduction and Related Work

Archery has a long history as a weapon in combat and has evolved into a popular sport. In modern target archery, archers aim to hit a standardized circular target. The target faces, standardized by the World Archery Federation (WA or FITA), consist of concentric circles with assigned point values. When scoring arrows, if an arrow lands on the line between two rings, it is scored with the higher point. Many archers still manually record scores on paper or existing apps, which can be time-consuming and distracting. The goal of this work to develop an automatic archery system using smartphones to read and record scores.

Existing hardware-based archery scoring systems include FalcoEye [FalcoEye, 2012] and the RyngDyng system [RyngDyng Technology, n.d.]. The FalcoEye system utilizes laser technology and optical methods for accurate detection and visualization of hit points. The RyngDyng system uses cameras for precise arrow position measurement, requiring hardware setup and calibration. These systems are relatively heavy and expensive.

Previous projects have explored computer vision techniques for real-time archery scoring. Parag's approach in 2017 used a colour-based method and frame difference, but it struggled with lighting variations and object distinction [Parag, 2017]. His project employed the Hough line transform and contour testing to detect arrow shaft lines and intersections more accurately. Morphological operations were also proposed to obtain arrow outlines [Parag, 2017]. Other researchers have proposed a smartphone-based archery analysis system that utilizes machine learning-based object detection to locate arrows [Peng et al., 2021]. The image processing methods [Zin et al., 2013; Parag, 2017] relied on frame difference in videos to isolate arrows, which required fixing the camera position and implementing frame extraction algorithms. Furthermore, intricate geometric shape analysis was performed to detect arrow locations. However, our project aimed to devise a lightweight solution for automatic scoring, eliminating the need for supplementary tripod setups and utilizing users' mobile phone cameras. Nevertheless, we found the

approaches for target sheet detection and arrow scoring to be valuable, employing a similar image processing methodology.

## 2 Implementation

The system comprises two main components: arrow localisation and arrow scoring.

In the localisation component, pure image processing approaches show inconsistencies when there is a change in lighting condition and potential movements of camera. Parag [2017] requires a camera record video in fixed position to extract arrows from other objects. We decided to employ deep learning approach to make the localization simpler in implementation part and potentially able to handle more generalized real-world conditions. After conducting tests and experiments, we selected the architecture proposed by Ribera et al. [2019] for tasks such as crowd counting. They made slight modifications to the U-Net architecture to accommodate a single class object as input. Notably, their approach does not require bounding boxes during training data labelling and allows for a flexible number of inputs and outputs. These features make Ribera et al.'s architecture well-suited for this project's requirements, as the number of arrows shot onto the target can vary during training and competitions. Compared to generic object detection architectures like YOLO, Ribera et al.'s approach offers greater adaptability.

Through training, the neural network acquires the capability to accurately identify the exact location where each arrow lands on the target, which is commonly referred to as the intersection between the arrows and the target paper. Following the completion of training, the localisation neural network model is able to generate predicted locations for new images in the form of a saliency map, which can subsequently be converted into sets of pixel coordinates, as illustrated in the example in Figure 1. These coordinates play a pivotal role in furnishing the essential arrow positions for the scoring component of the system, thereby enhancing the accuracy of the mask for arrow scoring.

Creating a diverse and extensive training dataset was crucial for training the deep learning arrow localisation model. As there was no publicly available dataset containing archery target bosses, we captured 480 distinct images over training sessions, featuring various scores, arrow colours, perspectives, target backgrounds, and levels of target paper tear. Data labelling involved assigning point labels using the Label Studio tool, labelling each image with corresponding location coordinates and classes representing score values ranging from 1 to 10, along with a class for missed arrows. Data augmentation techniques, including rotation and random flipping, were applied to the original image set, resulting in an additional 86,400 images. Our dataset, consisting of accurately labelled FITA standard indoor targets, along with their scores, is publicly available on Google's Kaggle platform in JSON format under the CC BY-NC-SA 4.0 license [Ma, 2023].

The scoring component utilizes the standardised colours associated with different scoring rings on the target face. To identify the target face, the system employs a colour range to mask out specific colours, such as blue, from the target as shown in figure 2. Subsequently, upon locating one or more rings, the system applies a series of image processing techniques to reduce noise and correct any skew in the circles. This crucial step enables precise determination of the centroid corresponding to the bull's eye on the target face. With knowledge of the exact ring colour for detection,

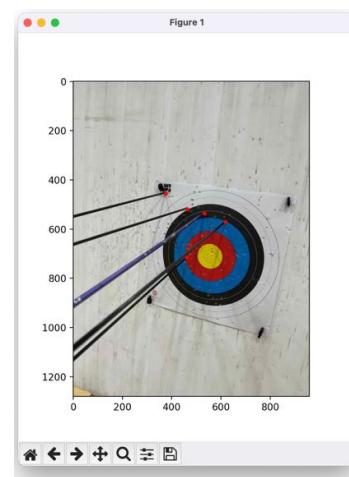


Figure 1: Visualisation of the labelled points

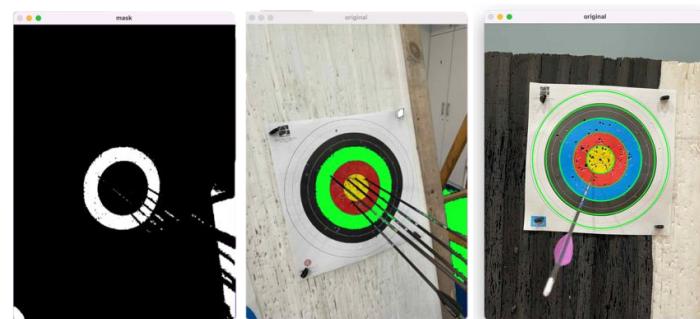


Figure 2: Illustration of segmenting other rings relative to the outer ring ellipse. The centroid is marked with a blue dot.

the system then calculates the parameters of each ring and matches the coordinates predicted by the localization neural network to obtain the score for each detected arrow.

Finally, the system generates an output image that visualizes the predicted arrow locations and provides the corresponding scores in the user interface.

### 3 Experiments and Evaluation

Ribera et al. [2019] proposed a novel modified version of the average Hausdorff distance and called it the weighted Hausdorff distance (WHD):

$$d_{WH}(p, Y) = \frac{1}{S + \epsilon} \sum_{x \in \Omega} p_x \min_{y \in Y} d(x, y) + \frac{1}{|Y|} \sum_{y \in Y} M_\alpha [p_x d(x, y) + (1 - p_x) d_{max}]$$

where:

$$S = \sum_{x \in \Omega} p_x, \quad M_\alpha [f(a)] = \left( \frac{1}{|A|} \sum_{a \in A} f^\alpha(a) \right)^{\frac{1}{\alpha}}$$

Here,  $d_{WH}(p, Y)$  is the weighted Hausdorff distance. During training, as shown in Figure 3, there are several loss terms displayed. Term 1 and Term 2 are the two terms to be summed for calculating weighted Hausdorff distance, respectively. Term 3 is the smoothed L1 loss for the regression of the object count, this is related to the number of ground truth points and the number of predicted points on the image, calculated by first subtract the estimated point number from ground truth number then perform regression. The final training loss is a combination of Term 1, Term 2 and Term 3, as the sum of the weighted Hausdorff distance and the regression term:

$$\mathcal{L}(p, Y) = d_{WH}(p, Y) + \mathcal{L}_{reg}(C - \hat{C}(p))$$

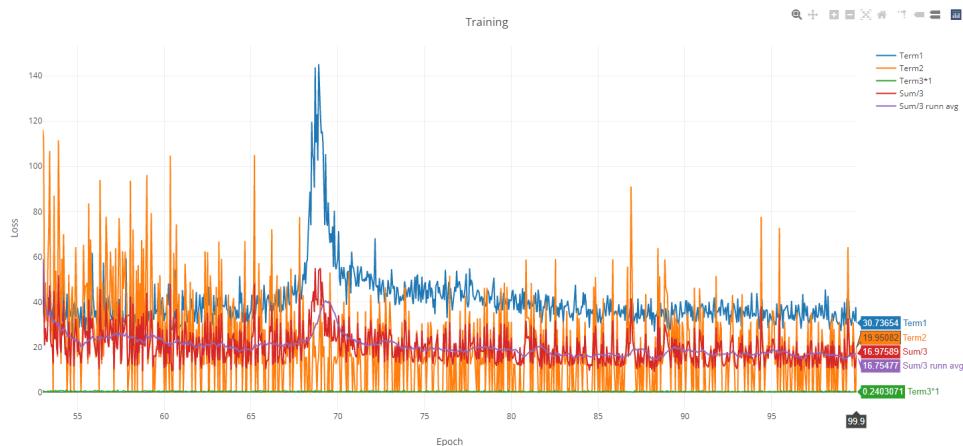


Figure 3: Training loss using original 480 images set

As can be seen in Figure 3, as the training progresses, the model exhibits increased confidence in accurately identifying the intersection points. This trend can be seen qualitatively in the saliency map shown in Figure 4, which shows that the model's ability to locate the intersection points on the target improves as training proceeds from 400 epochs to 1000 epochs. Throughout the training process, the training loss gradually decreased and reached a value below 10 at approximately 1000 epochs. At the point where we halted the



Figure 4: (Left) Saliency map produced during validation after around 400 epochs, the white points are ground truth. (Right) Localisation after around 1000 epochs of training.

training process, the model's performance was evaluated on a validation set that consisted of randomly torn target papers. The achieved recall and precision rates on this set were approximately 40%. Additionally, a separate test was conducted using five images with clean papers, similar to the illustration shown in Figure 4 (right). In this test, the model achieved a recall and precision rate of 100%, accurately identifying the positions of the arrows.

## 4 Conclusions, Limitations and Future Work

While the evaluation of the training process with existing dataset indicates progressive improvement in the detection results as training advances, it remains uncertain whether the current architecture can ultimately attain a level of accuracy that allows athletes to reliably substitute manual reading and counting tasks with this tool.

To improve the accuracy of the detection model, strategies include diversifying the training dataset by gathering a wider range of samples, exploring alternate neural network architecture like anchor-free YOLOv5, and further optimising hyperparameters such as the learning rate and optimizer choice.

Further front-end development is vital for the seamless integration of image capture, detection, and scoring functionalities. This necessitates additional research and training, particularly for enabling trained model accessibility across scripts and refining perspective correction to enhance bullseye detection accuracy.

Existing solutions, like Microsoft Office Lens, struggle with perspective correction when the target paper's colour resembles the background. An alternative could be Augmented Reality (AR) surface detection. Using frameworks like ARCore for Android and ARKit for iOS, the system could detect the target board surface and overlay a virtual grid. This could offer a reference for perspective correction without needing to identify specific points, like target paper corners.

## References

- [Ma, 2023] Ma, H. (2023). Archery target FITA 60cm, *Kaggle.com*. Retrieved 28 June 2023 from <https://www.kaggle.com/datasets/haozhema/archery-target-fita-60cm>
- [FalcoEye, 2012] FalcoEye (2012, 27 Sep). *FalcoEye Promotion video* [video]. YouTube. [https://www.youtube.com/watch?v=moAM26mwvWc&ab\\_channel=ArcheryScoringSystem](https://www.youtube.com/watch?v=moAM26mwvWc&ab_channel=ArcheryScoringSystem)
- [Zin et al., 2013] Zin, T. T., Oka, I., Sasayama, T., Ata, S., Watanabe, H., & Sasano, H. (2013). Image processing approach to automatic scoring system for archery targets. In *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 259-262). <https://doi.org/10.1109/IIH-MSP.2013.73>
- [Parag, 2017] Parag, R. (2017). Sequential recognition and scoring of archery shots. <https://www.semanticscholar.org/paper/Sequential-recognition-and-scoring-of-archery-Parag/4a6c29fb0a7962ff8c3a8c87f28ad6575615bcbb>
- [Peng et al., 2021] Peng, J.-S., Chen, Y.-J., Lin, W.-Y., Chen, H.-C., & Liao, C.-N. (2021). The Development and Implementation of a Smartphone Based Archery Analysis System. *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 1–2. <https://doi.org/10.1109/ICCE-TW52618.2021.9602923>
- [Ribera et al., 2019] Ribera, J., Güera, D., Chen, Y., & Delp, E. J. (2019). Locating Objects Without Bounding Boxes (arXiv:1806.07564). arXiv. <https://doi.org/10.48550/arXiv.1806.07564>
- [RyngDyng technology, n.d.] RyngDyng technology. (n.d.). Retrieved 28 June 2023, from <https://www.archery-electronics.com/en/public/ryngdyng/technology>

# Deep Learning enabled Computer Vision in Remanufacturing and Refurbishment applications: Defect Detection and Grading for Smart Phones

Reenu Mohandas, Martin Hayes, Colin Fitzpatrick, Mark Southern

*University of Limerick*

## Abstract

This project combines the use of Deep Learning using Computer Vision for Remanufacturing End of Life consumer electronics products, like smart phones, wearables like fitness trackers and smart watches which could be used in a secondary market alongside new ones to meet consumer demand. The process of refurbishment of these devices is a growing industry but heavily dependent on manual labour which brings in subjectivity of decisions, especially in grading. Deep Learning based computer vision in this work achieves high accuracy, greater than 80% in detecting defects on phones and determining phone grades, to ensure high consistent throughput rate in the remanufacturing process, thereby more sustainability.

**Keywords:** Deep Learning, Defect Detection, Remanufacturing, Refurbishment, Computer Vision,

## 1 Introduction

Consumer Electronics and IT & telecommunication equipment including mobile phones, iPads and tablets, wearables like activity trackers and smart watches are the few sectors that benefitted from the autonomous mass production in Industry 4.0. On reaching their End-Of-Life (EOL), these products are commonly referred to as E-waste or WEEE (Waste Electrical and Electronic Equipment). Shorter Product Life Cycle, fast changing customer attitudes, and ceaseless roll-out of newer versions into the market, all contributed to WEEE becoming the fastest growing stream of waste. In 2017, the United Nations University published a report, The Global E-waste Monitor Report, stating that only 20% of 44.7 million tonnes (Mt) E-waste generated was properly recycled [Islam, M.T. and Huda, N., 2018] in the year 2016.



Figure 1: Smart Phones for Recycling

Refurbishment and remanufacturing are ways to reduce carbon emissions, by restoring EOL products to reusable states without loss in function. To address the conflict of interest between economic goals and ecological ones, manufacturer's profit in remanufacturing sector should also be taken into account[Cheng, P., et al, 2022]. High variability and difference in availability are challenging factors in this sector, where there is a need to maintain high throughput rate to combat for the lower price of final products to make the process profitable[Schlüter, M et al., 2021]. This is where AI and Human-in-the-Loop concepts could be of assistance, to ensure faster and more accurate processing of objects and accommodate the wide variability of objects presenting for processing.

## 2 Purpose of Research

Companies rely heavily on manual labor for sorting, inspection, grading and identification of products and this needs experience, but also introduces subjectivity into the process. To maintain consistency in sorting and identification of defects, determined observation is needed for extended periods of time. This subjective nature adds more pressure to the workers as they can directly impact the economic value of the reverse logistics company[Schlüter, M et al., 2021]. The core problem in this scenario is the human-centric nature of sorting, identification, and defect detection. This is particularly important for determining what polishing process should be chosen for the screens. One of the possible ways to help this is to take humans a step further away and introduce automated systems, AI based identification/detection systems and computer vision to detect, identify and evaluate the products and human operator/personnel can assist by a confirmation or dismissal of suggestions generated by the AI. This will lead way to a more automated re-manufacturing/refurbishment setting with Human-in-the-Loop for further training and updating the initial trained model to be more accurate and faster.

## 3 Methodology

The problem under consideration is cosmetic damage on a glass screen of an electronic equipment, consider smart phone as a subset. Detection and analysis of scratches will help to determine if they could be sent to rework for processes like polishing which could remove the scratches and upgrade the final value of the equipment in the secondary market. Identification of more severe damage leads to efficient use of time in more valuable products sending the damaged ones to disassembly or eventually even to just recycle the product.

In the experiment, 100 images from smart phone samples were collected and annotated for segmentation of scratch defects, single class detection using Mask-RCNN[He, K. et.al 2017] one of the highly accurate instance segmentation models available in TensorFlow. Segmentation is the process of partitioning an image into constituent parts/regions or objects[E Woods et. al, 2008]. Considered as a one-class problem at this stage, every instance of scratch is separately identified and counted to determine the severity of defect/damage. Semantic segmentation is the process of classifying the pixels in the image into one identified category, without differentiating different object instances[He, K. et.al 2017]. Instance segmentation is in itself a very challenging task that there is detection of each object and then segmenting each instance separately.

Transfer Learning is the machine learning technique to address the problem of low number of training data. The deep learning model could be trained on a generic dataset and the final layers could be fine-tuned using the smaller available dataset for the detection/classification task[Tan, C. et. al, 2018]. Segmentation of scratch and detection of scratches are both performed in the input image. Pixel-wise segmentation helps measure the pixel-wise length of the scratch, detection creates bounding boxes to locate the scratches and obtain the x-min, y-min, x-max and y-max coordinates of the scratch.



**Figure 2:** phone screen scratch images taken using a digital microscope

**Experiment setting and Dataset:** Annotation of the image data for instance segmentation was carried out using software ‘*labelme*’. The Mask-RCNN model is trained on GPU GeForce RTX 2080 Ti using TensorFlow 1.13.1 and the detection function is programmed using Python 3.6 and OpenCV 4.4.0. The reason for using instance segmentation is the dynamic nature of variability in the type and location of scratches. Image data is collected using a digital microscope with (1920 x 1080 camera resolution and 2.5 - 68 (Optical) and 69d - 136.5d (Digital) magnification range(Figure 3).



Figure 3: Digital Microscope

## 4 Results and Discussion

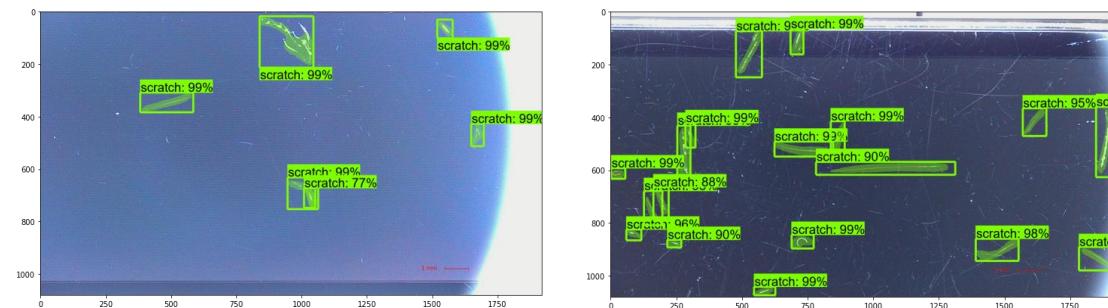


Figure 4: Segmentation results of scratches on the smart phone screen

The model gives promising results with very high accuracy with the limited dataset used for finetuning. The aim of this work is to grade the smart phone screen based on the cosmetic defects. At this stage of the project, the defect under consideration is scratches on the phone screen. Figure 3 shows the detection results with the number of scratches detected. Currently, any phone with less than 10 count of scratches, is considered Grade 1, 11-30 count of scratches is considered Grade 2 and more than 30 count of scratches is Grade 3. The importance of this classification is to determine the amount of time that need to be spent on the phone(device in general) which determines if the value of product could be improved by less expensive rework or indeed to discard as the rework is expensive and which drives the cost for a secondary product.

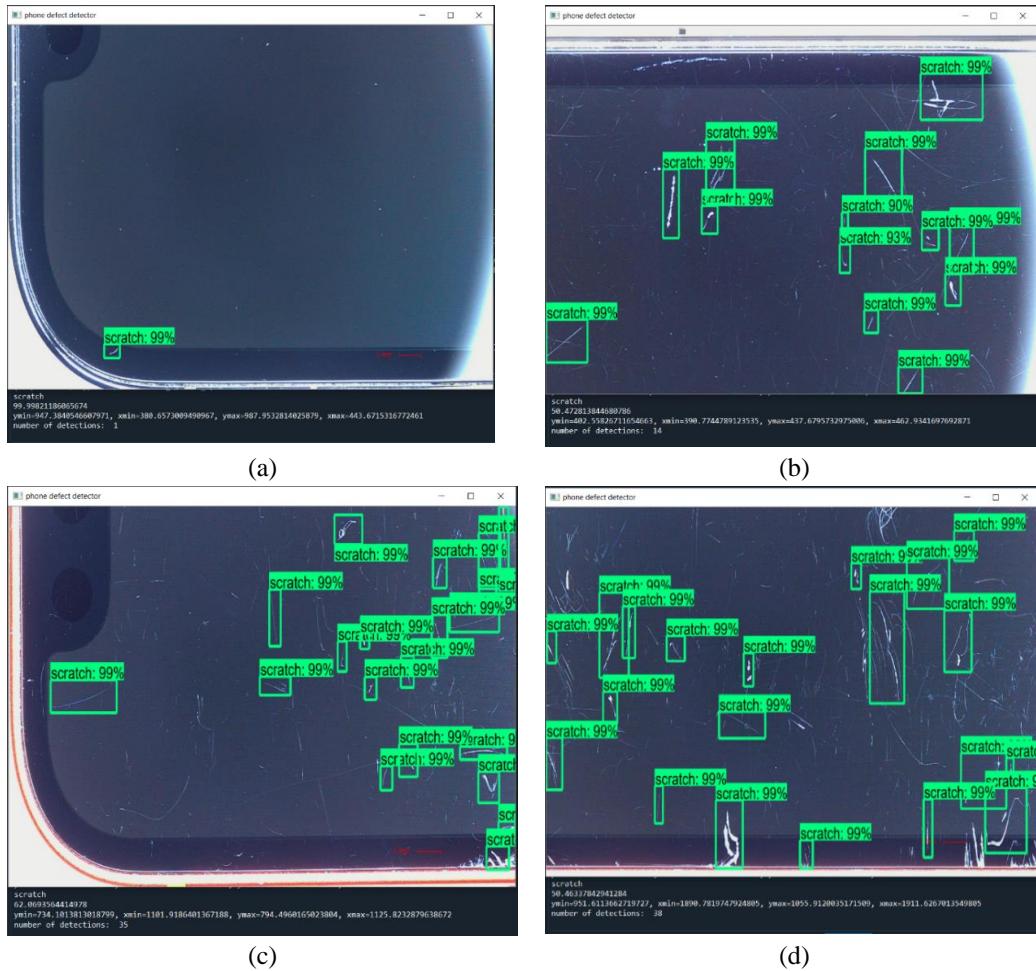
Detection model	IOU value	Precision	Accuracy
Mask CNN	0.8314	0.8	0.8

Table 1: Evaluation of results

The Accuracy of Mask CNN model is measured using the IOU score value which is the ratio of Intersection over union of the predicted mask to the ground truth mask of the object under consideration. The models shows high precision and Accuracy values where Precision depicts the rate of True positives per total detection and Recall shows the total True positives against all ground truth instances.

## 7 Conclusions

The currently trained models give high accuracy in detecting and segmenting the scratches on the phone screen and classify the screen using the number of scratches detected. The experiment is currently in progress and will be further developed into a multi-class experiment with consideration of the different types of scratches the phones are presented with. Future work includes strategies to automatically detect and differentiate cracks from scratches on phone screen and further improve automatic grading to refurbished smart phones.



**Figure 5:** (a) shows a Grade1 phone high value for secondary use (b) Grade2 phone and (c)&, (d) Grade3 phones.

## References

- [Islam, M.T. and Huda, N., 2018] Islam, M.T. and Huda, N., 2018. *Reverse logistics and closed-loop supply chain of Waste Electrical and Electronic Equipment (WEEE)/E-waste: A comprehensive literature review*. Resources, Conservation and Recycling, 137, pp.48-75..
- [Cheng, P., et al, 2022] Cheng, P., Ji, G., Zhang, G. and Shi, Y., 2022. *A closed-loop supply chain network considering consumer's low carbon preference and carbon tax under the cap-and-trade regulation*. Sustainable Production and Consumption, 29, pp.614-635..
- [Schlüter, M et al., 2021] Schlüter, M., Lickert, H., Schweitzer, K., Bilge, P., Briese, C., Dietrich, F. and Krüger, J., 2021. *AI-enhanced identification, inspection and sorting for reverse logistics in remanufacturing*. Procedia CIRP, 98, pp.300-305.
- [E Woods et. al, 2008] E Woods, R. and C Gonzalez, R., 2008. Digital image processing.
- [He, K. et.al 2017] He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. *Mask r-cnn*. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969)
- [Tan, C. et. al, 2018] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C., 2018. *A survey on deep transfer learning*. In Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27 (pp. 270-279).

# Improving GMM registration with class encoding

Solmaz Panahi, Jeremy Chopin, Matej Ulicny & Rozenn Dahyot

*Department of Computer Science, Maynooth University, Ireland*

## Abstract

Point set registration is critical in many applications such as computer vision, pattern recognition, or in fields like robotics and medical imaging. This paper focuses on reformulating point set registration using Gaussian Mixture Models while considering attributes associated with each point. Our approach introduces class score vectors as additional features to the spatial data information. By incorporating these attributes, we enhance the optimization process by penalizing incorrect matching terms. Experimental results show that our approach with class scores outperforms the original algorithm by [Jian and Vemuri, 2011] in both accuracy and speed.

**Keywords:** Point set registration, Graph matching, Gaussian mixture models (GMMs)

## 1 Introduction

Registration of point sets, that involves aligning multiple sets of points in a common coordinate system, is a fundamental task in computer vision and pattern recognition. Gaussian Mixture Models (GMMs) are a powerful representation for point distributions and the Euclidean distance between GMMs is a robust cost function to minimize for estimating the transformation between point sets to perform registration [Jian and Vemuri, 2011]. Point sets can be understood as graphs with points as nodes but without edges linking the nodes. Consequently graph matching algorithms have also been proposed to register graphs affected by spatial deformation [Zhou and De la Torre, 2013]. In this paper, we show experimentally that adding class scores as attributes benefit registration with GMMs [Jian and Vemuri, 2011], providing a more accurate registration while reducing computational time.

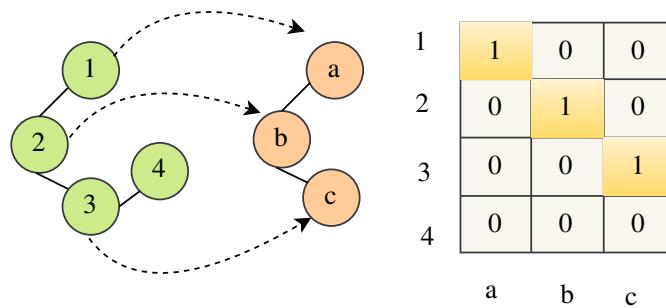


Figure 1: Example between two graphs with nodes  $\mathcal{V}_1 = \{1, 2, 3, 4\}$  and  $\mathcal{V}_2 = \{a, b, c\}$  (left). Graphs Matching aims at estimating binary matrix X of correspondences (shown right) between  $\mathcal{V}_1 = \{1, 2, 3, 4\}$  and  $\mathcal{V}_2 = \{a, b, c\}$  given a global affinity matrix K.

## 2 State of the art

[Chopin et al., 2023] recently introduced a deep learning pipeline for image segmentation augmented with a graph matching technique for post-processing to correct errors of over-segmentation. Their proposed graph encoding uses class scores as features on the vertices (or nodes) of a graph that are associated with segmented regions. The graph matching is performed between a graph model of perfect segmentation and a new observed one assuming that both are spatially aligned, hence the approach is sensitive to rotation effects for instance. To address this problem, graph matching can be extended with registration methods such as ICP (Iterative Closest Point) for matching graphs subject to global rigid and non-rigid geometric constraints [Zhou and De la Torre, 2013]. As an alternative to ICP registration, GMM-reg have been proposed as a robust approach able to estimate spatial transformations even with outlier correspondences [Jian and Vemuri, 2011]. Here we investigate extending GMM registration using class scores as node attributes and relate it to graph matching.

**Notations.** We define two graphs noted  $\mathcal{G}_1 = \{\mathcal{V}_1, \mathcal{E}_1\}$  and  $\mathcal{G}_2 = \{\mathcal{V}_2, \mathcal{E}_2\}$  such that (ignoring subscripts)  $\mathcal{V}$  is the set of vertices (a.k.a. nodes), and  $\mathcal{E}$  the set of edges. Graph matching aims to estimate a binary matrix  $X = [X_{ij} \in \{0, 1\}]$  of correspondences between nodes  $\mathcal{V}_1$  and  $\mathcal{V}_2$ , given a global affinity matrix  $K$  encoding node and edge similarities between the two graphs (cf. Fig 1).

**Graph Matching.** For each pair of nodes  $(i, j) \in \mathcal{V}_1$  and pair  $(l, s) \in \mathcal{V}_2$  the given affinity  $K_{is, jl}$ , graph matching finds a mutual assignment between elements of the sets  $\mathcal{V}_1$  and  $\mathcal{V}_2$  to maximize the total score for all pairs of assignments [Zhou and De la Torre, 2013]:

$$\hat{X} = \arg\max_X \sum_{i,j \in \mathcal{G}_1} \sum_{s,l \in \mathcal{G}_2} K_{is, jl} X_{is} X_{jl} \quad (1)$$

subject to some constraints [Huang et al., 2021]  $\forall i \in \mathcal{G}_2 : \sum_{s \in \mathcal{G}_1} X_{is} \leq 1$  and  $\forall s \in \mathcal{G}_1 : \sum_{i \in \mathcal{G}_2} X_{is} \leq 1$ . When nodes capture spatial information, graph matching can be extended to take into account a deformation  $T$  allowing to register graph  $\mathcal{G}_1$  on graph  $\mathcal{G}_2$  [Zhou and De la Torre, 2013].

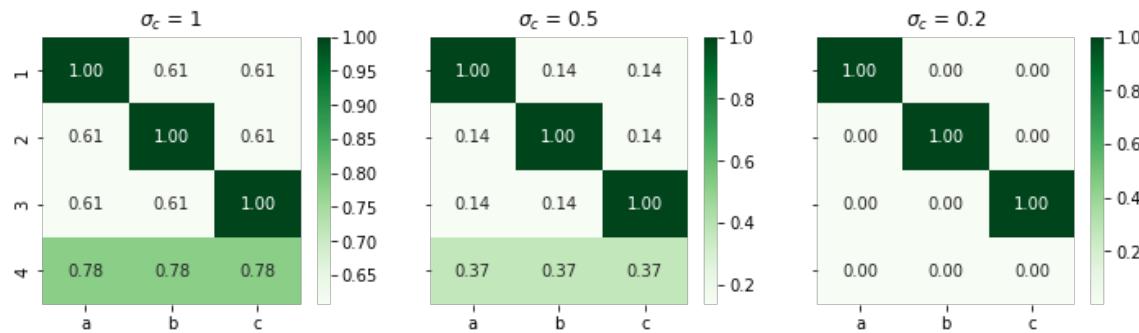


Figure 2:  $\sigma_c$  effect when computing weights  $w_{i,s}$  between nodes of graphs shown Fig. 1. Selecting smaller values for  $\sigma_c$  imposes a strict constraint on the optimization process by penalizing the matching of nodes that belong to different classes. The matrix of weights converges towards the binary matrix  $X$  of correspondences shown in Fig. 1.

**Registration with GMMs.** [Jian and Vemuri, 2011] have proposed to estimate deformation  $T$  between point sets  $\mathcal{V}_1$  and  $\mathcal{V}_2$  by minimising the Euclidian distance between two Gaussian Mixtures Models (GMMs) fitted on

each point set. Here rigid transformation (rotation, translation) is considered in  $\mathbb{R}^2$ , in which case the estimation of  $T$  is performed as:

$$\hat{T} = \arg \max_T \sum_{i=1}^{|\mathcal{V}_1|} \sum_{s=1}^{|\mathcal{V}_2|} \mathcal{N}(0; T(v_1^{(i)}) - v_2^{(s)}, \Sigma) \quad (2)$$

where  $T$  is a transform function of parameter  $\theta = [t_1, t_2, \phi]$  representing the translation and rotation,  $\mathcal{N}(x; \mu, \Sigma)$  indicates the normal distribution for random vector  $x$  with mean  $\mu$  and covariance  $\Sigma$ . For simplicity, we have chosen isotropic covariance  $\Sigma = \sigma^2 I_2$  in this work ( $I_2$  identity matrix in  $\mathbb{R}^2$ ).  $v_1^{(i)}$  is the spatial coordinate in  $\mathbb{R}^2$  used as attribute for the node  $i$  in  $\mathcal{V}_1$  (resp.  $v_2^{(s)}$  is the spatial coordinate in  $\mathbb{R}^2$  used as attribute for the node  $s$  in  $\mathcal{V}_2$ ).

### 3 Registration with class attributes

Following [Chopin et al., 2023], we propose to extend Equation (2) by concatenating a class vector (noted  $c$ ) to the spatial coordinate ( $v$ ) as part of the attribute describing the nodes such that the estimation becomes:

$$\hat{T} = \arg \max_T \sum_{i=1}^{|\mathcal{V}_1|} \sum_{s=1}^{|\mathcal{V}_2|} \exp \left( \frac{-\|T(v_1^{(i)}) - v_2^{(s)}\|^2}{4\sigma^2} \right) \times \exp \left( \frac{-\|c_1^{(i)} - c_2^{(s)}\|^2}{4\sigma_c^2} \right) \quad (3)$$

In formula (3), the term with class scores can be interpreted as weights  $w_{is}$  that does not depend on the transformation  $T$  to be estimated:

$$w_{is} = \exp \left( \frac{-\|c_1^{(i)} - c_2^{(s)}\|^2}{4\sigma_c^2} \right) \quad (4)$$

We note that if  $c_1^{(i)} = c_2^{(s)}$  the weight is equal to one  $w_{is} = 1$ . When  $c_1^{(i)} \neq c_2^{(s)}$  (nodes with different class scores), then the weight is less than one  $w_{is} < 1$ . This implies that incorrect matches are penalized. We show experimentally (Section 4) that this leads to better estimates  $\hat{T}$  with faster convergence of the algorithm.

**Effect of  $\sigma_c$ .** When we choose  $\sigma_c$  small enough, the weights  $w_{is}$  for  $i \neq s$  become close to zero. Figure 2 shows the effect of different values of  $\sigma_c$ .

### 4 Experimental results

We use the fish data from [Jian and Vemuri, 2011] in our experiments<sup>1</sup>. Fish data has 98 points for which 2D coordinates are provided and these are augmented with a unique class score vector  $c = [0, 0, \dots, 1, \dots, 0]^T \in \mathbb{R}^{98}$  and  $\sum_{i=1}^{98} c_i = 1$ . Class scores used for  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are the same while 2D coordinates are rotated from  $\pm 24^\circ$  to  $\pm 96^\circ$  with  $\pm 24^\circ$  intervals in  $\mathcal{G}_2$ . The experiments are done with  $\sigma = \{2; 1; 0.5\}$  and  $\sigma_c = 0.2$ . The error is computed as the Euclidian distance  $\|\hat{T} - T_{GT}\|$  between the estimated transformation  $\hat{T}$  and ground truth  $T_{GT}$ . Figure 3 shows the experiments categorized by different values of  $\sigma$ . The results yields notable improvements in terms of accuracy and computational efficiency. Estimations on each experiments has been improved as well as decreasing the convergence time across different experiments.

### 5 Conclusion

Class scores have proven useful for graph matching for improving image segmentation results from deep learning [Chopin et al., 2023] and this work shows that class information also can provide better performance of registration with GMMs [Jian and Vemuri, 2011]. Future work will investigate if edge information can also be used efficiently for registration.

<sup>1</sup>The code will be available at [https://github.com/solmaki97/GMMReg\\_Extension](https://github.com/solmaki97/GMMReg_Extension)

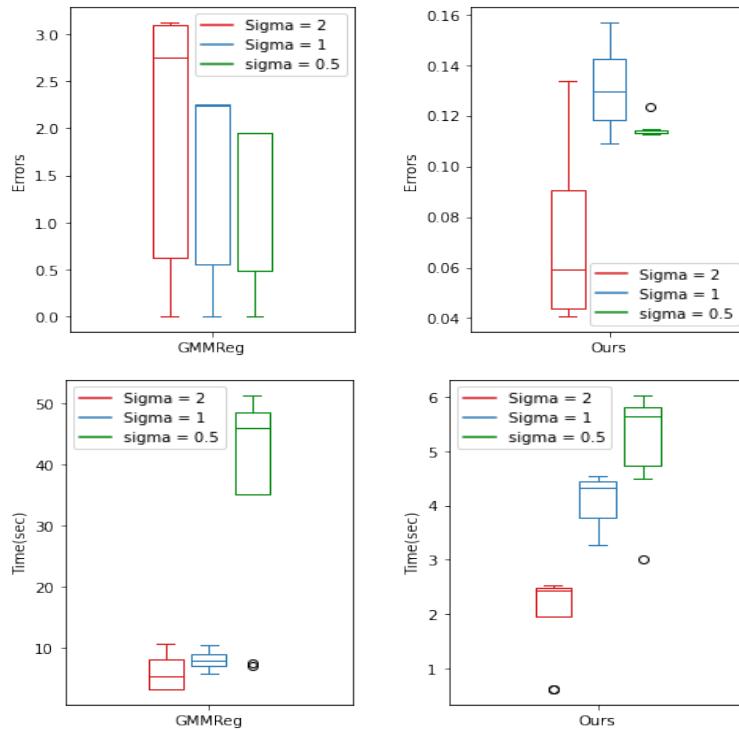


Figure 3: Experiment results on fish data with  $\sigma_c = 0.2$ . Distribution of errors  $\|\hat{T} - T_{GT}\|$  are shown as box plot for  $\sigma = \{2, 1, 0.5\}$  for both [Jian and Vemuri, 2011] algorithm (top left) and our approach (top right): note the difference of the order of amplitude in the error report on y-axis. Time for computations are likewise reported (bottom plots) showing our algorithm efficiency.

## Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049/, the SFI Research Centres ADAPT (13/RC/2106 P2) and IFoM (16/RC/3872), and is co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Chopin et al., 2023] Chopin, J., Fasquel, J.-B., Mouchère, H., Dahyot, R., and Bloch, I. (2023). Model-based inexact graph matching on top of cnns for semantic scene understanding. *Computer Vision and Image Understanding (CVIU) journal*.
- [Huang et al., 2021] Huang, X., Mei, G., Zhang, J., and Abbas, R. (2021). A comprehensive survey on point cloud registration. *CoRR*, abs/2103.02690.
- [Jian and Vemuri, 2011] Jian, B. and Vemuri, B. C. (2011). Robust point set registration using Gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1633–1645.
- [Zhou and De la Torre, 2013] Zhou, F. and De la Torre, F. (2013). Deformable graph matching. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2922–2929.

# Calculating Breathing Rates from Remote PPG Signals Using Machine Learning Methods

Adara Andonie, Timothy Hanley, Dara Golden, Robyn Maxwell, Joe Lemley, and Ashkan Parsi

*OCTO Sensing Team, Xperi Inc., Galway, Ireland*

## Abstract

Remote detection of health metrics, such as heart rates or breathing rates, has grown in interest. There is numerous amount of literature on calculating breathing rates using RGB cameras or thermal cameras but extracting the respiratory signal from NIR cameras is not straightforward. Thus, different ideas must be implemented. It is possible however to extract the PPG signal from an NIR camera, and from PPG signals breathing rates can be calculated. This research presents the first step in being able to predict the breathing rates from NIR cameras, which is to be able to predict the respiratory signal based on a remotely acquired PPG signal. Using machine learning, models were trained to predict respiratory signals from PPG signals. Results show that the models are predicting the respiratory signals with fair accuracy.

**Keywords:** NIR Cameras, PPG Signals, Breathing Rates, Machine Learning

## 1 Introduction

Breathing rate (BR) is an important aspect of health monitoring and can provide various information on an individual. The ability to accurately monitor health metrics remotely has many advantages in situations where devices cannot be used, for example, while driving, monitoring the elderly in their own homes [Aoki et al., 2001] [Martinez and Stiefelhgen., 2012] and monitoring infants [Scalise et al., 2011].

It is possible to obtain the BR indirectly through an ECG signal [Bao et al., 2020] [Kim et al., 2007], and less commonly done, the PPG signal as well. The PPG signal holds information for calculating health metrics, such as extracting the heart rate (HR) from the signal. However, the PPG signal contains information that makes it possible to calculate the BR from this signal [Karlen et al., 2013].

Often when remotely calculating BR, cameras such as RGB and thermal cameras are used. However, extracting the breathing rate from a Near-Infrared (NIR) camera is quite difficult as breathing is not so easily seen on these cameras which is why this paper is important because it provides a steppingstone in that direction.

Work combining machine learning to calculate the BR from health signals has been done [Ganser et al., 2019]. However, this research considered other metrics such as ECG, SPo2, HR, and others from a device to predict the respiratory rate.

Currently, there is little literature on calculating the BR from only PPG signals and even less on predicting the BR from only the PPG signals obtained from NIR cameras. This paper will introduce a method to predict the respiratory rate based on remotely detected PPG signals from NIR cameras using machine learning, present current findings, and finally expand on future developments that can be employed.

## 2 Data Processing and Training Methods

This section will cover information on the data used for training, validation, and testing of the machine learning models, the data processing, and the background of the models used for training.

## 2.1 Materials and Data Processing

The training and validation dataset used is the BICMC dataset [Pimentel et al., 2017]. BICMC is a large public dataset, extracted from the MIMIC II dataset, containing signals (RESP, PLETH (PPG), V, AVR, II) and numeric (HR, PULSE, RESP, and SpO2). There was a total of 53 recordings done and each recording lasted 8 minutes. The signals were sampled at 125 Hz and the numeric were sampled at 1 Hz.

The data was downloaded and all the signals/numeric for each recording were combined into one large data frame for each recording. From these data frames, only the PPG signals, RESP signals, and times were extracted and then down-sampled to 30 Hz to match the sampling frequency of the test data, and the signals were then normalized between 0 and 1. These signals and times were saved as data frames for each recording. The data was then split 80% to training and 20% to validation.

The testing data was from the public dataset MR-NIRB [Nowara et al., 2022] which contains a series of videos of subjects in different driving scenarios filmed using both RGB and NIR cameras centered on the faces of the subjects. The subjects wore pulse ox to measure the PPG signal. From the NIR videos two PPG signals were considered for this paper: the ground truth PPG signal from the pulse ox and remotely estimated PPG signals calculated as a separate project.

Due to the lack of publicly available datasets containing both PPG signals and respiration signals, a different approach had to be taken to validate the results. As the PPG indicates HR, we have data with PPG signals in different HR ranges. For example, we have PPG signals that fall into the HR range of 50-60 beats per minute (bpm), 60-70 bpm, and 130-140 bpm. We validated the predicted signals with the idea that HR correlates with BR [Bahmed et al., 2016], as the lower the HR the lower the BR; the same applies for higher HR and higher BR. Thus, the number of peaks detected should correspond to the HR ranges.

## 2.2 Machine Learning Models

Several machine learning models for regression were used to predict the respiratory signal [Ganser et al., 2019]: OLS, Elastic Net, Bayesian Ridge, Lasso, Ridge, MLP, KNN, Random Forest Regressor, ADA Boost Regressor, and XGB Regressor. Out of these models, KNN and MLP proved the best for the task at hand. The qualifications that determined which models were better suited included the number of peaks of the actual respiratory signal and the number of peaks of the predicted respiratory signal from the test data and the clarity of the predicted signal. If a model had a poor signal, the peak detection would still consider weak points that would otherwise be invalid. This explains why even if the number of peaks for the actual respiration signal and predicted signal were the same, the model was disregarded.

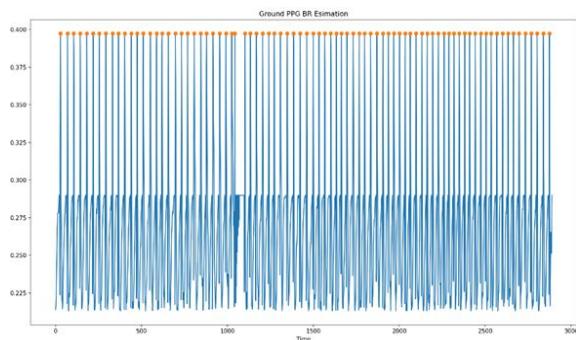
Model	Actual Resp Signal Peaks	Predicted Resp Signal Peaks
OLS	17	17
Elastic Net	17	15
Bayesian Ridge	17	17
Lasso	17	15
Ridge	17	17
MLP	17	15
KNN	17	16
RFF	17	16
Ada	17	15
XGB	17	17

**Table 1.** Table comparing the results of the different machine learning models between the number of peaks in the ground truth respiration signal and the number of peaks in the predicted respiration signal.

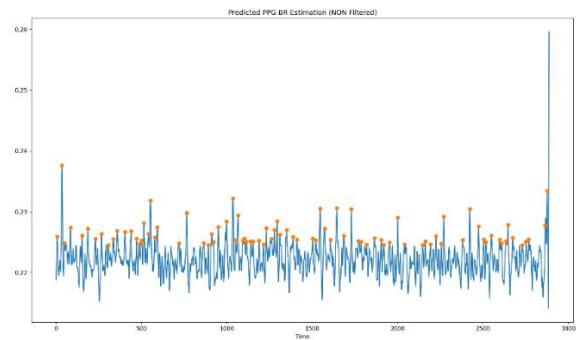
### 3 Results and Discussion

This paper will focus on the results of the MLP model as an example. Figure 1 shows the non-filtered predicted respiration signal of the ground truth PPG signal with an HR of 50-60 bpm with a total number of peaks of 83. Figure 2 shows the non-filtered predicted respiration signal of the estimated PPG signal from a NIR camera with the same HR range of 50-60 bpm with a total number of peaks of 86. Figure 3 shows the non-filtered respiration signal of the ground PPG signal with a HR range of 120-130 bpm with a total number of peaks of 194 and Figure 4 is the non-filtered respiration signal from the estimated PPG signal in the same range of 120-130 bpm and a total number of peaks of 159.

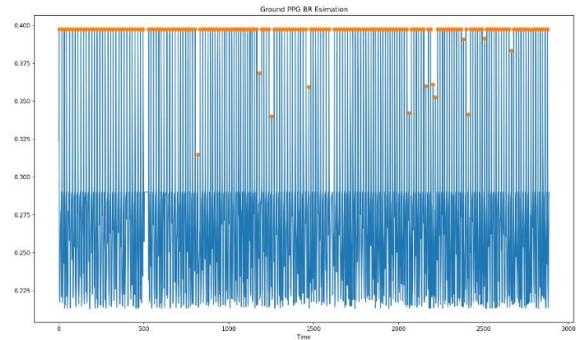
Results demonstrate that the model can predict the respiration signal for various ranges of PPG signals. The number of peaks detected for an HR range of 120-130 bpm is 194 and more than halved (86) for an HR range of 50-60 bpm on the ground truth ppg signal. This pattern continues for the estimated PPG signal from the NIR camera. Another important detail to take notice of is the number of peaks from the ground truth signal prediction and its respective estimated PPG signal prediction. Even with no post-processing done, these numbers are relatively close to each other (83 and 86 number of peaks from lower HR and 194 and 159 for higher HR). These observations further indicate the accuracy of the model in the situation that there is currently no ground truth respiratory signal or BR to compare results due to the lack of available databases.



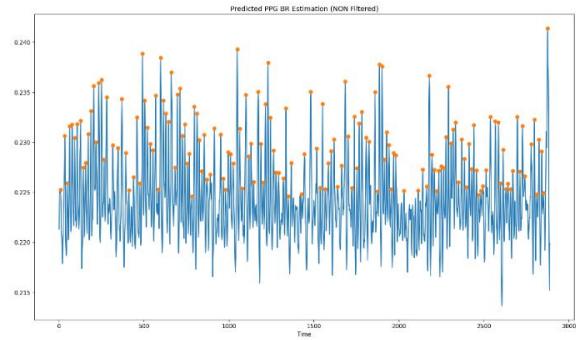
**Figure 1. MLP Respiratory signal of ground truth PPG with HR range 50-60 bpm**



**Figure 2. MLP Respiratory signal of estimated PPG with HR range 50-60 bpm**



**Figure 3. MLP Respiratory signal of ground truth PPG with HR range 120-130 bpm**



**Figure 4. MLP Respiratory signal of estimated PPG with HR range 120-130 bpm**

## 4 Conclusion and Future Work

Results from the machine learning models demonstrate that they can predict the respiratory signal based on provided PPG signals, with significant differences in the number of peaks for PPG signals with an HR range of 50-60 bpm and those within the range of 120-130 bpm for both the ground truth PPG signal and the remotely estimated PPG signal from a NIR camera. The next step in this research is to post-process the predicted signals in a way that doesn't overfit and skew the results such that the signals are clearer and easier to extract an accurate BR. Another important step would be to continue searching for available datasets with ground PPG and respiratory signals or start designing an experiment to collect the necessary information to test the training of the models more accurately. The exploration of different models will continue to be worth researching and trying. Deep Learning methods are currently being explored, such as LSTM and RNN models, to possibly improve the results of the machine learning counterparts.

## References

- [Aoki et al., 2001] H. Aoki, Y. Takemura, K. Mimura, and M. Nakajima, "Development of non-restrictive sensing system for sleeping person using fiber grating vision sensor," MHS2001. Proceedings of 2001 International Symposium on Micromechatronics and Human Science (Cat. No.01TH8583), Nagoya, Japan, 2001, pp. 155-160, doi: 10.1109/MHS.2001.965238.
- [Bahmed et al., 2016] Bahmed F, Khatoon F, Reddy B R, Relation between respiratory rate and heart rate – A comparative study. Indian J Clin Anat Physiol 2016;3(4):436-439
- [Bao et al., 2020] Bao X, Abdala AK, Kamavuako EN. Estimation of the Respiratory Rate from Localised ECG at Different Auscultation Sites. Sensors (Basel). 2020 Dec 25;21(1):78. doi: 10.3390/s21010078. PMID: 33375588; PMCID: PMC7796076.
- [Ganser et al., 2019] J. Ganser, H. Dashwood, and A. Al-Khafaji, "Predicting Respiratory rate using data available on a smartwatch". Apr. 2019.  
[https://github.com/JoeGanser/Predicting\\_Respiratory\\_Rate#readme](https://github.com/JoeGanser/Predicting_Respiratory_Rate#readme)
- [Karlen et al., 2013] W. Karlen, S. Raman, J. M. Ansermino and G. A. Dumont, "Multiparameter Respiratory Rate Estimation From the Photoplethysmogram," in IEEE Transactions on Biomedical Engineering, vol. 60, no. 7, pp. 1946-1953, July 2013, doi: 10.1109/TBME.2013.2246160.
- [Kim et al., 2007] Kim, J.M., Hong, J.H., Kim, N.J., Cha, E.J., Lee, TS. (2007). Two Algorithms for Detecting Respiratory Rate from ECG Signal. In: Magjarevic, R., Nagel, J.H. (eds) World Congress on Medical Physics and Biomedical Engineering 2006. IFMBE Proceedings, vol 14. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-540-36841-0\\_1030](https://doi.org/10.1007/978-3-540-36841-0_1030)
- [Martinez and Stiefelhagen., 2012] M. Martinez and R. Stiefelhagen, "Breath rate monitoring during sleep using near-ir imagery and PCA," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 2012, pp. 3472-3475.
- [Nowara et al., 2022] E. M. Nowara, T. K. Marks, H. Mansour and A. Veeraraghavan, "Near-Infrared Imaging Photoplethysmography During Driving," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 4, pp. 3589-3600, April 2022, doi: 10.1109/TITS.2020.3038317.
- [Pimentel et al., 2017] M. A. F. Pimentel et al., "Toward a Robust Estimation of Respiratory Rate from Pulse Oximeters," in IEEE Transactions on Biomedical Engineering, vol. 64, no. 8, pp. 1914-1923, Aug. 2017, doi: 10.1109/TBME.2016.2613124.
- [Scalise et al., 2011] L. Scalise, I. Ercoli, P. Marchionni and E. P. Tomasini, "Measurement of respiration rate in preterm infants by laser Doppler vibrometry," 2011 IEEE International Symposium on Medical Measurements and Applications, Bari, Italy, 2011, pp. 657-661, doi: 10.1109/MeMeA.2011.5966740.

# The Xperi 3D Full Body Photogrammetric Scanner

Arpad Zoldi, Stefan Bigoi, Jakub Pawelec, Padraig Toomey, Victor Vlad and, Bogdan Basuc

Xperi Corporation

## Abstract

In this paper we present an overview of a high-resolution 3D full-body photogrammetric scanner developed by Xperi corporation. This custom scanner captures a 360-degree set of HQ images of a person with c.140 DSLR cameras and a multi-spectral flash illumination system. Some of these DSLR cameras are modified to capture 850 nm and 940 nm NIR (near infrared) images and the flash system provides associated NIR flashes before and after the visible flash. The resulting HQ images are processed by custom software to build a dense finite element model with corresponding texture maps and additional 3D metadata. The resulting models can be subsequently rigged to generate 3D movement and gesture sequences. The primary use for this data is to build large data libraries to train state-of-the-art neural algorithms for Automotive and Consumer Technology applications. Such models are more flexible and can often provide better ground truth and repeatability than data collected with human subjects. A number of example use-cases in the context of Driver Monitoring systems are discussed.

**Keywords:** Image Processing, Machine Vision, Photometric Scanner, 3D models, Neural Algorithms

## 1 Introduction

Collecting data for tasks specific to machine vision is challenging, often requiring substantial effort and resources. One of the challenges that persists, is obtaining labelled data that accurately represents real world scenarios that the model will encounter. In recent years, synthetic data generation has emerged as a valuable technique to tackle this problem (Lu et al., n.d.). Synthetic data can be used to supplement real-world data. The advantage of this is the possibility to generate data at scale.

There are currently many techniques available to generate this data such as using generative adversarial networks (GANs) (Gilbert et al., 2021). 3D models of people can be animated to mimic real-world actions and behaviour in the context of driver monitoring systems (DMS). Furthermore, generating synthetic data can drastically reduce the time needed to acquire and annotate data, enabling the creation of labelled data with ground truth information, which simplifies the training process for neural algorithms.

To address this fundamental problem with data acquisition, Xperi has developed a scanner which utilises its own corresponding toolchain that generates realistic 3D models from a series of overlapping, still images of real people. The scanner is supported by customised software, which is designed for high volume capture and parallel processing using photometry to capture as many details as possible. The final 3D model can then be used to simulate human actions, gestures, expressions, and activities in all manner of contexts. Moreover, there are applications in building 2D image datasets that are used to train neural algorithms running on the AI platform systems found in driver monitoring systems where we can potentially save lives.

This paper presents a comprehensive overview of Xperi's high-resolution 3D full-body photometric scanner, detailing its development, technical components, and the associated software for data processing. Furthermore, it explores the potential applications and benefits of the generated 3D models in the Automotive and Consumer Technology sectors, with a specific focus on Driver Monitoring systems.



Figure 2 Participant in Scanner

The scanning is based on imagery taken by 117 DSLR-cameras mounted within the 3D scanner. The cameras are fixed on 11 static poles, and another 12th pole is configured to be swung inward as a means of entry into the scanner. The isometric view in figure 2.1 below shows the entrance into the scanner. The cameras are triggered synchronously via their internal trigger mechanism. Data is collected by specifically modified cameras in the VIS, and NIR, which is then stored on the scanner's control PC. The images are then processed to produce high resolution meshes with VIS and NIR textures, to create the final 3D model. To capture NIR images, all scans require short flashes of light in different wavelengths (850nm and 940nm NIR). To obtain a high-resolution mesh, the 117 cameras are each positioned at slightly different angles, giving coverage of the entire body. Each of

these cameras also has a high-quality lens with a fixed focal length. While this does limit the lenses' field of view, it comes with the benefits of more stable optical properties, less distortion in the images, and higher level of detail. Specular highlights from flash photography can cause artifacts on the generated mesh, which is minimized by using diffuse and indirect studio flashes.

## 2.1 System Hardware Architecture

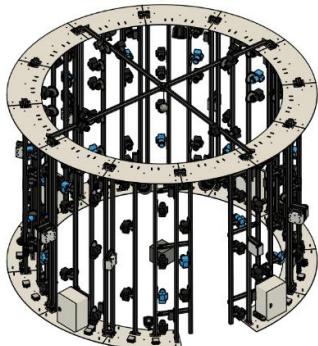


Figure 2.1 CAD drawing

The entire system consists of multiple interconnected subsystems that streamline data acquisition and processing. It includes 36 modified cameras capable of capturing NIR frequencies, and 81 regular VIS cameras. A central control box stores the main hardware interface, as well as the microcontroller, the circuitry, and power supplies. The microcontroller is responsible for triggering the cameras, and controls NIR-LED lighting frequencies during data capture. A scanner-control PC manages the administration, data storage, collection, and 3D mesh generation. The brief flashes during every capture provide lighting in visible light, and two NIR frequencies (850nm and 940nm). The microcontroller is powered through a direct connection to the scanner control PC, which operates the system via the Xperi Scan Queue software. All collected data is stored on the scanner-control PC through USB hubs, which also enables administration and setup. All the cameras are connected to three master hubs located in the scanner, which are then connected to four master hubs that are connected directly to the scanner-control PC.

## 2.2 System Software Architecture

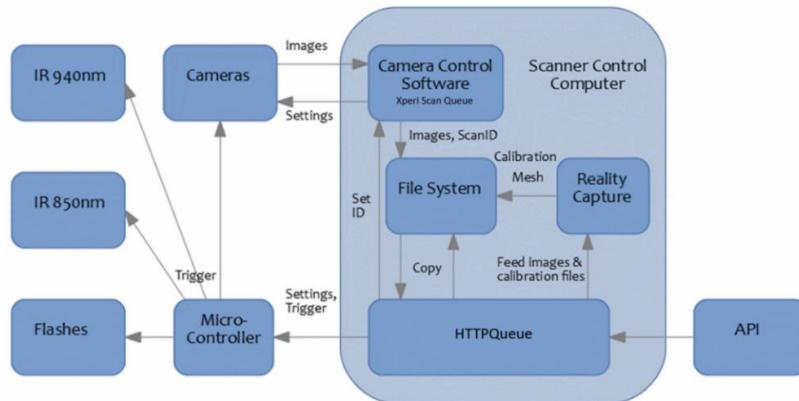


Figure 2.2 Block Diagram of the System Software Architecture

The software toolchain assists and streamlines the acquisition of 3D models, and processing of related data. 3D models are first obtained as a sequence of still images, in the VIS and NIR spectrum, which are then processed to yield the final output. The optimization of the capturing process focuses on visible, NIR 850nm, NIR 940 nm, image generation. The data capture procedure is handled on the microcontroller. The scanner control PC manages data collection. Xperi Scan Queue checks whether a scan is complete, handles metadata collection, and 3D reconstruction queuing. An API enables microcontroller scanning adjustments, and any camera setting changes. The 3D reconstruction phase takes place in Reality Capture, where the software generates a .obj format mesh using the captured images, which is textured with visible and NIR imagery. The procedure is controlled through batch processes, which ensure efficient and systematic handling of the data.

### 2.3 Data Capture Procedure

Scans are initiated in the Xperi Scan Queue app, signalling the microcontroller in the control box to run through the entire scan operation. Firstly, all cameras are prepared for synchronized capture. Following that, the LED lighting in the 850nm frequency is switched on, shortly thereafter a picture is taken with the IR cameras only, and the 850nm frequency light is switched off again. A second capture is triggered for all cameras. Due to the time it takes for the cameras to release the shutter, the corresponding flash is triggered at a slightly later point in time. This interval is referred to as the flash gap.

The flash gap is set at a default value (e.g., 100ms), and the shutter speed is set to a value that ensures sharp images. After each capture, the microcontroller waits a short time interval to trigger the next capture. This interval is referred to as the capture gap. This procedure is repeated with the 940nm frequency LEDs, and another scan is done by triggering just the IR cameras. The initial trigger current is switched off, the cameras return to their normal state. Immediately after the first scan, the images will be sent from the cameras to the scanner-control-PC via the USB connection and the hubs.

## 3 Example demonstrators

### 3.1 3D Model Generation:



Figure 3.1.1 VIS Texture



Figure 3.1.2 Mesh

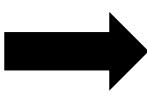


Figure 3.1.3 Completed Model

The scanner generates a 3D mesh for each participant scanned (Fig. 3.1.2). Textures are created by combining and overlapping the captured 2D Images, and separate textures are created for VIS, NIR 940nm and 850nm (Fig. 3.1.1 shows us the visible light texture), which then wraps around the generated mesh, giving us a high-resolution virtual model of the participant. The models may be sent to a 3D modelling expert who can modify each model as needed, cleaning up details and adding motion to various body parts, such as the limbs, head, eyelids and mouth, depending on the context. The finished models are used in simulations to create synthetic data, as can be seen in figure 3.3.1.

### 3.2 Scanner Performance



Figure 3.2 3D Model Close-up with Errors

### 3.3 Driver Monitoring Systems use case:



Figure 3.3.1 Model in Simulated Environment

Scanner performance is affected by various factors and is very sensitive to less-than-ideal circumstances, where model generation becomes prone for error. Participants wearing dark or baggy clothes, as well as anything reflective or glossy such as jewellery, glasses, and other accessories, often encounter errored models.

Figure 3.2.1 shows a case where baggy clothes caused the participant's model to have incomplete arms. Slight movements and jitters can also cause the model to be errored. The speed at which photos are taken minimises the risk of this happening. Ideal models are generated when participants wear more tightly fitting clothes, with no black or reflective surfaces, and no accessories, as seen in figure 3.1.3.



Figure 3.3.2 Real-life Environment

A relevant use case for these models exists in the context of Driving Monitoring Systems (DMS) - The EU passed a law mandating driver drowsiness and attention warning systems in all newly manufactured vehicles from July 2022 for new types and July 2024 for all new vehicles (Europa.eu, 2019). Realistic driving scenarios can be created by incorporating the 3D models into simulated environments. These generated models offer precise control over the simulation environment, which ensures repeatability and facilitates rigorous testing. By leveraging virtual environments, Xperi can accelerate its software development cycles, reduce safety risks associated with physical testing, and iterate quickly to enhance the accuracy and performance of their DMS.

## References

- [Xperi, 2019] Xperi. (2019). 3D Full-body Scanner System Design Document [Internal document]. Unpublished.
- [Europa.eu, 2019] Europa.eu. (2019). Available at: [https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=PI\\_COM:Ares\(2021\)1075107&rid=11](https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=PI_COM:Ares(2021)1075107&rid=11)
- [Corcoran, P., 2022] Corcoran, P. (2022). Flash Photography - Risk Analysis Report. [online] Available at: <https://www.universityofgalway.ie/c3i/datasets/datacollectionactivities/ethicalreviewsupport/>
- [Gilbert et al., 2021] Gilbert, A., Marciniak, M., Rodero, C., Lamata, P., Samset, E. and Mcleod, K. (2021). Generating Synthetic Labeled Data From Existing Anatomical Models: An Example With Echocardiography Segmentation. IEEE Transactions on Medical Imaging, [online] 40(10), pp.2783–2794. doi: <https://doi.org/10.1109/TMI.2021.3051806>.
- [Lu et al., n.d.] Lu, Y., Tech, V., Wang, U., Wei, W. and Wang, H. (n.d.). Machine Learning for Synthetic Data Generation: a Review. [online] Available at: <https://arxiv.org/pdf/2302.04062.pdf>

# An Introduction to the Xperi Driving Simulation Environment: Hardware, Software and Data Acquisition

Rachel Corcoran, Paul Kielty, Padraig Toomey, and Joe Lemley,  
*Xperi , Galway, Ireland*

## Abstract

In this paper details are given on the Driving Simulation Environment system developed by Xperi and how it is used for a large-scale data acquisition for driving drowsiness studies. The data collected includes RGB, NIR and thermal video data, audio data (speech and paralinguistic elements such as yawns), and several biosignals from a human participant in a simulated driving environment. This set of biosignals contains electroencephalography (EEG), heart rate, blood oxygen level, breathing rate, eye-motion and blink data, temperature, and galvanic skin response. The experimental protocol involves completing a number of tests and simulated driving sessions at fixed times over a 24-hour period, with rest periods where participants can eat, watch movies, read, or otherwise entertain themselves. However, they are not allowed to sleep or consume any stimulants or depressants during the data acquisition. A key goal is to develop machine vision algorithms that can detect human cognitive, emotional and drowsiness states from non-contact video data. Operational details and procedural aspects of the data acquisition are given. This paper complements the training workshop run at IMVIP 2023.

**Keywords:** Driving Simulation, Image Processing, Machine Vision, Driver Monitoring, Data Acquisition.

## 1 Introduction

It is estimated that tired drivers are the cause of 1 in 5 road collisions every year in Ireland (Pires et al. 2020; Commission 2021). When a driver starts to tire their reactions slow and eventually, they begin to exhibit microsleeps or even fall asleep at the wheel. The introduction of driver monitoring systems (DMS) will mitigate the associated risks but requires a better understanding of how drowsiness manifests in different individuals and can be identified via non-contact solutions such as image/video-based sensing (Albadawi, Takruri, and Awad 2022; Cardone et al. 2021; Jahan et al. 2023). Xperi is a world leader in DMS technologies and is currently undertaking a large-scale data study on 500 human participants to improve the sensing capabilities of DMS technologies. A training workshop will be presented at IMVIP 2023 for delegates who wish to learn about the data acquisition techniques and learn from the experience of the Xperi team in this field.

This short paper provides complimentary information on the rationale and basis for this study and some details on the driving simulation environment (DSE) that is used at Xperi to conduct these studies. It will also provide an overview of the software, sensors and protocols used to gather data in the DSE. The importance of different data types is explained as well as the process of relating certain data to a participant's drowsiness state, which is detected using neuromorphic vision cameras. The data collected includes heart rate, neural signals, blood oxygen levels, body temperature, eye movements, self-assessment tests and respiration. These act as indicators of the cognitive and emotional state of a participant and confirm their level of drowsiness as detected by the driver monitoring software through event cameras. This allows for precise training of a range of advanced neural-network based machine vision algorithms.

## 2 The Driving Simulation Environment

### 2.1 Hardware and Control Subsystems

The Xperi DSE consists of a car interior set on an actuated platform and installed with a specialized driving simulator wheel, pedals, and gearstick. The steering wheel and pedals have force feedback for immersion. Three 49-inch monitors are configured as a continuous display and offer a wide field of view for rendered scenes in front of the simulator. For a display to run tests in parallel with the driving, a small additional monitor is mounted in front of the driver (but elevated for minimal obstruction of the display). The DSE equipment and software are controlled with one computer, with a separate computer used for collecting all video, audio, and physiological data. Other features of the DSE system that may need to be adjusted by technical staff are controlled using a keyboard/mouse interface. Note that the DSE has a complete seating arrangement to allow for studies with passengers and rear-seat occupants as well as driver-only studies. The simulator frame and dashboard provide mounting points for cameras and other equipment.



Figure 1(a): View of seating and immersive tri-screen display; blinds are drawn when the simulation is active.



Figure 1(b): Driver view with instrument cluster, steering wheel, pedal assembly and video acquisition systems

### 2.2 Driving Simulator Software

For an immersive driving experience, the DSE supports realistic driving video games *Assetto Corsa* and *EuroTruck Sim 2*. These offer highly configurable simulation environments providing customizability in terms of time of day, vehicle size, engine power, driving dynamics, braking capabilities, and other aspects of vehicle response. The software parameters are pre-set according to each driving scenario and are designed to provide study participants with a range of challenges of varying difficulty during each driving session. As their drowsiness level increases, they become more likely to make errors or fail in the assigned tasks. Custom software was developed for the simultaneous recording and monitoring of the camera feeds and biosignals.

## 3 Study Rationale and Data Type Selection

This study collects these multiple, complimentary video data streams, together with sensing data from a range of on-body biophysical sensors. The on-body sensor signals provide biophysical markers and biometric data that can characterise more specifically the cognitive, or emotional states of a human participant. These data are then used to annotate video data to provide a ground truth for training advanced machine vision models that may be incorporated into future generations of Driver Monitoring technology. The DMS in today's vehicles are often based on a near Infrared (NIR) video camera. This captures light across a broader range of wavelengths (up to 1500 nm) than a conventional video camera and thus performs better in the variable lighting conditions of a vehicle cabin. It can operate well both in strong sunlight and in night-time environments. However, there are several emerging imaging technologies which can provide different information that may be useful to tomorrow's

DMS. These include thermal imaging cameras based on uncooled bolometer technology which have become more competitive cost-wise in recent years, and event-cameras, sometimes referred to as neuromorphic vision systems, which instantly capture light-changes from individual pixels rather than accumulating a full image frame. Thermal imaging is based on the thermal emissivity of the human body and can provide very useful information on breathing patterns, facial blood flow and other physiological signals and operates independently from the lighting levels in the vehicle cabin. Event cameras detecting changes in individual pixel intensity provide much higher motion sensitivity and can track facial and eye-behaviours with higher temporal resolution than a video camera.

### 3.1 Data Types

**Video and Audio Data:** RGB, NIR, thermal and neuromorphic vision data are collected via a dedicated PC running custom acquisition software. Among other features, this software offers simultaneous displays of all cameras feeds and assigns a timestamp to each recorded frame. Capturing time-synchronized data is important to allow accurate annotation across data streams for the training of machine vision algorithms.

**Biosignal Data and equipment:** Biosignal data are captured primarily via the PLUX Biosignal Sensor Hub (Cardone et al. 2021; Jahan et al. 2023) supporting simultaneous acquisition across eight sensor channels. The following sensor types were selected for this study.

- Blood oxygen saturation levels are measured using SPO2 sensor.
- Heart rate is measured using a 3-electrode electrocardiography sensor.
- Eye motion and blink signals are measured using electrooculography.
- Temperature is measured with a thermistor secured directly on the forehead.
- Breathing rate is measured with an inductive respiration sensor.
- The galvanic skin response is measured with an electrodermal activity sensor on the palm

EEG across 32 channels is recorded with a [Neuroelectrics Enobio 32 EEG](#). EEG readings can be used to detect, for example, the onset of eye blinks, microsleeps and drowsiness levels. The change in electrical activity that occurs due to an eye blink is known as the Berger Effect, and it presents as an increased frequency of alpha waves in the prefrontal cortex (Kirschfeld, Kuno, 2005).

**Measuring Drowsiness Level:** The participant completes 1-hour simulated driving sessions at 5 points throughout the 24 hours. Accompanying tests to quantify drowsiness level are performed directly before and after the driving blocks. These include the Psychomotor Vigilance Task (PVT) (Huang et al. 2020; Trott et al. 2022), the Alpha Attenuation Test (AAT) (Stampi, Stone, and Michimori 1995), the Karolinska Sleepiness Scale (KSS) questionnaire, and observations of how well the participant is driving in the simulator (O'Callaghan et al. 2022). The PVT evaluates alertness and attentiveness by the participants reaction time. It is based on the participant's reaction time to visual stimuli that occur at random intervals. The AAT requires the participant hold their eyes closed for 2 minutes, then eyes open for 2 minutes, and evaluates their drowsiness by the difference in EEG activity in the 8-12Hz (alpha) band between the two states. (Hussain et al. 2022; Jing et al. 2020; O'Callaghan et al. 2022). Self-annotations also describe the participant's drowsiness level and are produced by the participants through KSS testing. When requested, the participant uses a tablet to indicate their drowsiness level on a scale from 1-9. Each value has a description, starting with "extremely alert" at 1 and finishing with "very sleepy, great effort to keep awake, fighting sleep" at 9. A value of 10 can also be assigned by the acquisition staff if the participant has fallen asleep when they should be completing the KSS questionnaire.

## 4 Data Acquisition Protocol

Participants for this study are recruited via an external agency. No later than 1 day before a participant's scheduled acquisition, they are contacted to ensure full understanding of the acquisition and verify all consent forms have been read and signed. At any time in the 24-hour acquisition the participant can request to finish the study and will be compensated based on the time completed. An overview the first 6 hours of the study protocol is provided in table 1. The recording sequences are repeated at 3pm, 11pm, 3am and 4am.

Summary of Activity	Approximate Duration	Goals
Introduction to the Lab environment, technical staff; explanations and Q&A	30 minutes	Allow participant to get familiar with the lab, staff and to confirm understanding of study protocols and expectations.
Sensor application, signal quality checks, and DSE instruction.	1 hour	All health sensors are applied and the quality of all biosignals and camera outputs is checked. The participant is instructed on the DSE and given time to practice driving.
Recording sequence #1: Drowsiness tests and 1 hour of driving; run remotely by technicians.	1.5 hours	Drowsiness level measured with AAT, PVT, and KSS before and after 1 hour of simulated driving. All cameras and sensors are recording throughout.
Driving without visible sensors. run remotely by technicians.	30 minutes	Sensors visible to the cameras are removed for another 15 minutes of driving. This generates data to verify camera-based algorithms are not using visible sensor features.
Break	2.5 hours	Remaining sensors removed and participant has a break.

Table 1: Initial 6-hour cycle of the study protocol

## References

- Albadawi, Yaman, Maen Takruri, and Mohammed Awad. 2022. “A Review of Recent Developments in Driver Drowsiness Detection Systems.” *Sensors* 22 (5): 2069.
- Cardone, Daniela, Chiara Filippini, Lorenza Mancini, Antonella Pomante, Michele Tritto, Sergio Nocco, David Perpetuini, and Arcangelo Merla. 2021. “Driver Drowsiness Evaluation by Means of Thermal Infrared Imaging: Preliminary Results.” In *Infrared Sensors, Devices, and Applications XI*, edited by Ashok K. Sood, Priyalal Wijewarnasuriya, and Arvind I. D’Souza, 25. San Diego, United States: SPIE. <https://doi.org/10.1117/12.2594504>.
- Commission, European. 2021. “Road Safety Thematic Report—Fatigue.” European Road Safety Observatory; European Commission, Directorate General ....
- Huang, Ying, Steve Hennig, Ingo Fietze, Thomas Penzel, and Christian Veauthier. 2020. “The Psychomotor Vigilance Test Compared to a Divided Attention Steering Simulation in Patients with Moderate or Severe Obstructive Sleep Apnea.” *Nature and Science of Sleep*, 509–24.
- Hussain, Iqram, Md Azam Hossain, Rafsan Jany, Md Abdul Bari, Musfik Uddin, Abu Kamal, Yunseo Ku, and Jik-Soo Kim. 2022. “Quantitative Evaluation of EEG-Biomarkers for Prediction of Sleep Stages.” *Sensors* 22 (8): 3079.
- Jahan, Israt, KM Aslam Uddin, Saydul Akbar Murad, M. Saef Ullah Miah, Tanvir Zaman Khan, Mehedi Masud, Sultan Aljahdali, and Anupam Kumar Bairagi. 2023. “4D: A Real-Time Driver Drowsiness Detector Using Deep Learning.” *Electronics* 12 (1): 235.
- Jing, Difei, Dong Liu, Shuwei Zhang, and Zhongyin Guo. 2020. “Fatigue Driving Detection Method Based on EEG Analysis in Low-Voltage and Hypoxia Plateau Environment.” *International Journal of Transportation Science and Technology* 9 (4): 366–76.
- Kirschfeld, Kuno. 2005. “The Physical Basis of Alpha Waves in the Electroencephalogram and the Origin of the ‘Berger Effect.’” *Biological Cybernetics* 92 (3): 177–85.
- O’Callaghan, David, Cian Ryan, Ashkan Parsi, and Joseph Lemley. 2022. “An EEG-Based Method for Drowsiness Level Estimation.” In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 1–4. <https://doi.org/10.1109/BHI56158.2022.9926820>.
- Pires, Carlos, Katrien Torfs, Alain Areal, Charles Goldenbeld, Ward Vanlaar, Marie-Axelle Granié, Yvonne Achermann Stürmer, Davide Shingo Usami, Susanne Kaiser, and Dagmara Jankowska-Karpa. 2020. “Car Drivers’ Road Safety Performance: A Benchmark across 32 Countries.” *IATSS Research* 44 (3): 166–79.
- “PLUX Biosignals.” n.d. Accessed July 7, 2023. <https://www.pluxbiosignals.com/>.
- Stampi, Claudio, Polly Stone, and Akihiro Michimori. 1995. “A New Quantitative Method for Assessing Sleepiness: The Alpha Attenuation Test.” *Work & Stress* 9 (2–3): 368–76.
- Trotti, Lynn Marie, Prabhjyot Saini, Erin Bremer, Christianna Mariano, Danielle Moron, David B. Rye, and Donald L. Bliwise. 2022. “The Psychomotor Vigilance Test as a Measure of Alertness and Sleep Inertia in People with Central Disorders of Hypersomnolence.” *Journal of Clinical Sleep Medicine* 18 (5): 1395–1403.

# Neuromorphic Seatbelt State Detection for In-Cabin Monitoring with Event Cameras

Paul Kielty<sup>1</sup>, Cian Ryan<sup>2</sup>, Mehdi Sefidgar Dilmaghani<sup>1</sup>, Waseem Shariff<sup>1</sup>, Joe Lemley<sup>2</sup>, and Peter Corcoran<sup>1</sup>

<sup>1</sup>*University of Galway, Galway, Ireland*

<sup>2</sup>*Xperi Corporation, Parkmore Indl. Estate, Galway, Ireland*

## Abstract

Neuromorphic vision sensors, or event cameras, differ from conventional cameras in that they do not capture images at a specified rate. Instead, they asynchronously log local brightness changes at each pixel. As a result, event cameras only record changes in a given scene, and do so with very high temporal resolution, high dynamic range, and low power requirements. Recent research has demonstrated how these characteristics make event cameras extremely practical sensors in driver monitoring systems (DMS), enabling the tracking of high-speed eye motion and blinks. This research provides a proof of concept to expand event-based DMS techniques to include seatbelt state detection. Using an event simulator, a dataset of 108,691 synthetic neuromorphic frames of car occupants was generated from a near-infrared (NIR) dataset, and split into training, validation, and test sets for a seatbelt state detection algorithm based on a recurrent convolutional neural network (CNN). In addition, a smaller set of real event data was collected and reserved for testing. In a binary classification task, the fastened/unfastened frames were identified with an F1 score of 0.989 and 0.944 on the simulated and real test sets respectively. When the problem extended to also classify the action of fastening/unfastening the seatbelt, respective F1 scores of 0.964 and 0.846 were achieved.

**Keywords:** CNN, Driver Monitoring, Event Camera, Neuromorphic Sensing, Seatbelt

## 1 Introduction

Neuromorphic vision describes a class of sensors designed to mimic biological perceptual functions. One such sensor is an event camera, which differs from a conventional camera in that each pixel records data asynchronously. Whenever one of these pixels detects a relative change in brightness above a set threshold an ‘event’ is logged. Each event is comprised of a timestamp, the coordinate of the pixel that reported the event, and a polarity to indicate whether an increase or decrease in brightness occurred. The event camera does not output images, but a list of events generated by motion or lighting changes in the scene. The event data has no intrinsic framerate, however, its time resolution exceeds that of video captured at 10,000 frames per second. Event cameras also offer higher dynamic range and lower power consumption than most conventional shutter cameras [Gallego et al., 2022].

A 2018 meta-analysis found that a fastened seatbelt reduces the risk of injury in road collisions by 65% [Fouda Mbarga et al., 2018], and in the United States, seatbelt use was shown to reduce mortality by 72% [Crandall et al., 2001]. Existing seatbelt alert systems in modern vehicles rely on pressure sensors in the seat to determine occupancy and simply detect if the seatbelt tongue is inserted in the buckle. This can easily be spoofed by buckling and sitting in front of the seatbelt, and has no ability to determine if a seatbelt has been fastened correctly. Also, it is often only implemented in the front seats of the vehicle. Camera-based seatbelt detection systems have the potential to rectify these flaws. With the ever-increasing demand for safer, more intelligent vehicles, there have been remarkable developments in camera-based DMS. At this stage they have been fully implemented in many modern consumer vehicles. With the camera systems already in place, it is possible to add new DMS features with minimal additional cost. Recent research has revealed how event cameras hold many advantages over standard shutter cameras in for driver monitoring tasks, particularly when it comes to face and

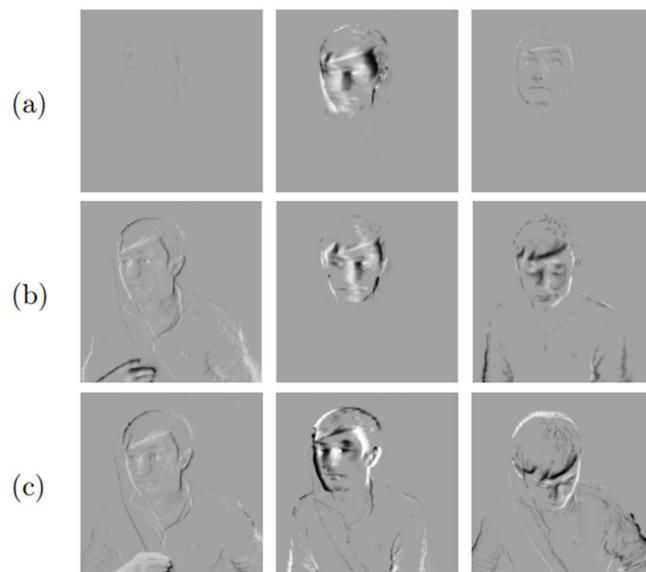
eye motion analysis [Ryan et al., 2021, Chen et al., 2020]. In this paper, we demonstrate the viability of another feature in an event-based DMS by creating the first event-based seatbelt state detector.

## 2 Event Data Simulation and Collection

An obstacle regularly faced in event camera research is the lack of publicly available large-scale datasets. This has driven the development of event simulators such as V2E [Delbrück et al., 2020], which enables the synthesis of realistic events from NIR or RGB videos by analysing the differences between consecutive frames. Most of the event data used for this research was simulated with V2E from a non-public industry dataset of NIR videos. Using a wide field of view camera on the rear-view mirror of a car, various subjects were recorded fastening and unfastening their seatbelts repeatedly. The video frames were labelled according to the following classes: (0) The subject's seatbelt is fastened. (1) The subject's seatbelt is unfastened. (2) The subject is fastening their seatbelt. (3) The subject is unfastening their seatbelt. A set of real event data was also collected for testing the network. A Prophesee EVK4 event camera was mounted beside the rear-view mirror of a driving simulator and focused on the driver's seat. Six subjects were asked to fasten and unfasten their seatbelt at random intervals throughout each recording. These videos were labelled manually with the same 4 classes as the NIR dataset.

## 3 Pre-processing of Event Data

The event data, both simulated and real, are saved as lists of events in text format. To use this data in CNNs and other image-based systems, it must first be represented in a 2D array. This is typically achieved by accumulating a group of events and summing the positive and negative events at each pixel location to create a 2D frame [Gallego et al., 2022]. When transforming an event recording into frames with this technique, the decision of how many events should be accumulated per frame must be carefully considered. The two most common approaches are to accumulate events over a fixed duration or accumulate a fixed number of events for each frame. The former method of grouping the events by a fixed duration is useful in tasks that could benefit from the temporal information in a sequence of frames as the generated frames will have fixed time spacing, much like conventional video formats. However, this approach is prone to generating frames with few events if there is little motion in the scene over the fixed duration. The alternative approach of forming each frame from a fixed number of events gives some assurance of a minimum amount of spatial information in each frame, at the loss of much of this temporal information. This works better for keeping the seatbelt visible when there is minimal motion in the frame, but motion of the head or background can quickly saturate the event count and generate many frames where the seatbelt is absent. A custom accumulation approach was developed for the final iteration of the dataset. Each frame was defined by a fixed number of events, however, only events within a rectangle bounding the subject's torso were counted. This maintained seatbelt visibility in more frames than the original two methods, as demonstrated in Fig. 1. where the fixed counts/duration were specified so each method generates 75 frames of the same "Seatbelt Fastened" clip. In the full 75 frames, the seatbelt



**Figure 1.** Frames from a "Seatbelt Fastened" event clip generated using (a) a fixed time period, (b) a fixed event count, and (c) a fixed event count over the torso region.

was visible in (a) 27%, (b) 71%, and (c) 93%. The final dataset contained 108,691 synthetic event frames and 8,317 real event frames. The simulated videos were randomly separated into training, validation, and test sets. The real event videos were all reserved for the testing.

## 4 Network Architecture

It is difficult to distinguish individual fastening/unfastening frames, but it becomes obvious when the whole sequence of frames is considered. Additionally, for the static classes with unreliable seatbelt visibility, using a sequence of frames can provide a more reliable result. For these reasons we used a recurrent CNN architecture which takes a frame sequence as the input for each prediction. Fig. 2 gives a high-level overview of the structure. The MobileNetV2 network is used as an efficient, lightweight backbone for initial feature extraction [Sandler et al., 2018]. Recent years have seen self-attention introduced to many CNN tasks for its ability to contextualize and apply a weighting to input features, with only a small computational cost. The self-attention module in our proposed network is implemented according to [Zhang et al., 2018]. When attended feature maps have been generated for every frame of the input sequence, they are stacked and passed to the recurrent head of the network. This is comprised of a 2 stacked bi-directional LSTM layers [Hochreiter and Schmidhuber, 1997].

## 5 Training

In this work, two models were trained. The first was for binary classification of frame sequences using the static fastened/unfastened classes only. For the second model, all classes were included to determine if the 4 states could be reliably identified, as they must all be handled in a real-world implementation. To train the network, the videos were split into single-class sequences of 15 frames, before randomized cropping and downsampling to a resolution of 256x256. Using cross-entropy loss and a batch size of 15 sequences, the network was trained for 30 epochs. The initial learning rate of  $1 \times 10^{-4}$  was halved every 5 epochs.

An added benefit of the self-attention layer is allowing us to visualize the areas in each frame that are more heavily weighted by the network. This can be helpful to verify that the network is utilizing appropriate features. Fig. 3 shows these weighted regions tracking the seatbelt when visualized on the real event videos in the test set.

## 5 Results and Conclusion

The results of the 2-class model and 4-class model on both the simulated and real test sets are compared in Table 1. As expected, the

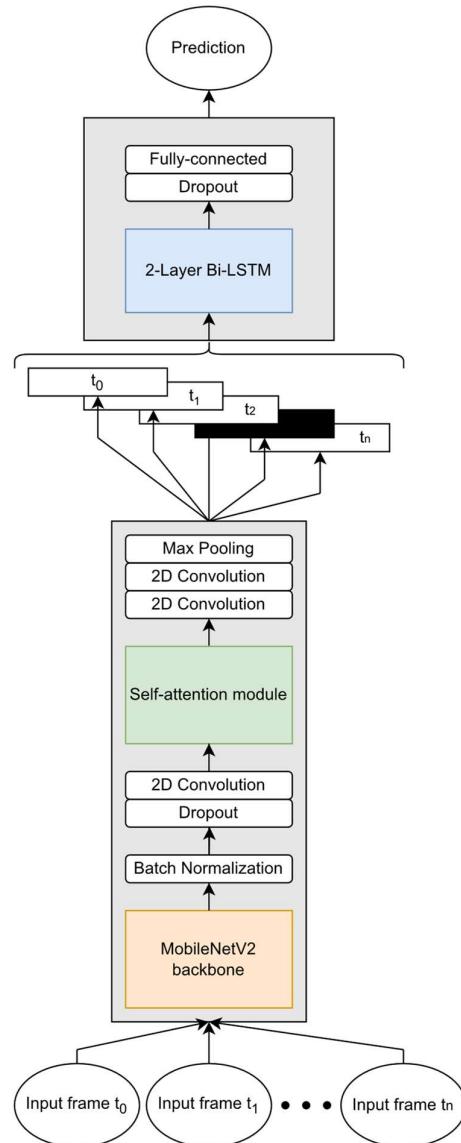


Figure 2: Proposed network architecture.

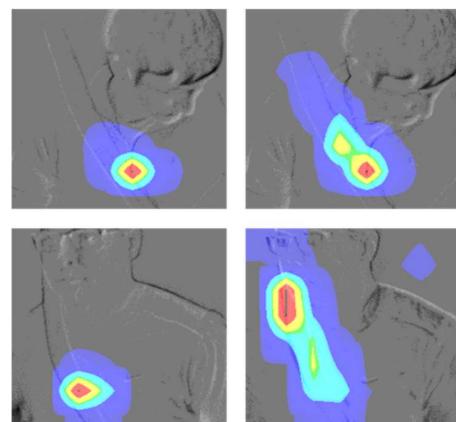


Figure 3. Visualized attention maps on test frames generated from real events.

2-class model was more accurate, but the 4-class model demonstrates that handling of all classes is possible without a dramatic reduction in performance. This model treats the 4 classes as independent, but we know they can only transition in a fixed sequence. Future work will leverage this fact for improved accuracy.

Model	Test set	F1
2-class	Simulated	0.989
	Real	0.944
4-class	Simulated	0.964
	Real	0.846

**Table 1. Summary of model performance.**

## Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at the University of Galway [13/RC/2106\_P2]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Chen et al., 2020] Chen, G., Hong, L., Dong, J., Liu, P., Conradt, J., and Knoll, A. (2020). Eddd: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor. *IEEE Sensors Journal*, 20(11):6170–6181.
- [Crandall et al., 2001] Crandall, C. S., Olson, L. M., and Sklar, D. P. (2001). Mortality reduction with air bag and seat belt use in head-on passenger car collisions. *Am. J. Epidemiol.*, 153(3):219–224.
- [Delbrück et al., 2020] Delbrück, T., Hu, Y., and He, Z. (2020). V2E: from video frames to realistic DVS event camera streams. *CoRR*, abs/2006.07722.
- [Fouda Mbarga et al., 2018] Fouda Mbarga, N., Abubakari, A.-R., Aminde, L. N., and Morgan, A. R. (2018). Seatbelt use and risk of major injuries sustained by vehicle occupants during motor-vehicle crashes: a systematic review and meta-analysis of cohort studies. *BMC Public Health*, 18(1):1413.
- [Gallego et al., 2022] Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., and Scaramuzza, D. (2022). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.
- [Jain, 1989] Jain, A. K. (1989). *Fundamentals of Digital Image Processing*. Englewood Cliffs NJ: Prentice-Hall.
- [McCarthy, 1960] McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine. *Communications of the ACM*, 7:184–195.
- [Ryan et al., 2021] Ryan, C., O’Sullivan, B., Elrasad, A., Cahill, A., Lemley, J., Kielty, P., Posch, C., and Perot, E. (2021). Real-time face eye tracking and blink detection using event cameras. *Neural Networks*, 141:87–97.
- [Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.
- [Zhang et al., 2018] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks

# Sign Language Recognition: Can depth cameras be used to correct Mediapipe errors?

Frank Fowley<sup>1,3</sup> and Anthony Ventresque<sup>2</sup>

<sup>1</sup>School of Computer Science, University College Dublin & SFI Lero

<sup>2</sup>School of Computer Science and Statistics, Trinity College Dublin & SFI Lero

<sup>3</sup>Centre for Research Training in Digitally-Enhanced Reality (D-REAL)

## Abstract

Mediapipe is increasingly deployed in Sign Language Recognition (SLR) pipelines. This paper highlights errors in Mediapipe that adversely affect the performance of these models. The use of depth cameras has been intuitively proposed for the correction of depth estimation errors. We propose a Mediapipe input pre-processing technique using depth sensor data specifically tailored for SLR systems. However, our experimental results show that such techniques are problematic owing to the inherent distortion in depth sensor output.

**Keywords:** Mediapipe, Depth Camera, Image Pre-processing, Sign Language Recognition, Computer Vision

## 1 Introduction and background

Sign Language Recognition (SLR) models are used to identify the phonetic elements of signs, one of which is the articulated hand shape [Rastgoo et al., 2021]. Mediapipe is a pose estimation model that outputs the 3D coordinates of the hand finger-tips and skeletal joints from an RGB image. These hand pose vectors can be used as feature descriptors in deep learning SLR models. Any errors in the Mediapipe output are propagated through the rest of the recognition pipeline. This paper describes an experimental approach to correct Mediapipe pose keypoint errors by using images from a depth sensor. The paper is structured as follows. Section 2 introduces the two main types of error in Mediapipe output. Section 3 describes a technique to correct such errors and Section 4 summarises the results and gives a conclusion.

### 1.1 Related work

Classification of Sign Language articulations is a difficult task because of the complex handshapes and signing speed of native signers [Bragg et al., 2019]. Pose models have been used successfully to extract features as part of Sign Language Recognition (SLR) pipelines [Wadhawan, A. & Kumar, P., 2021]. The top performing model in the Kaggle Isolated American Sign Language (ASL) Recognition competition achieved an accuracy of 89% using Mediapipe for feature extraction [Kaggle, 2023]. The performance limit due to the inherent class confusion of visually similar signs has been reported [Fowley et al., 2022].

## 2 MediaPipe errors

**Errors due to self-occluded hand poses:** The matrix in figure 1 displays the average cosine similarity between classes of poses abstracted by Mediapipe from a dataset of fingerspelling signs signed by native signers. Each class sample is represented as a 63-element vector made up of the 3D coordinates of the 21 skeletal keypoints of the pose. The values circled in red represent visual similarities between some hand shapes, for example, between “M” and “N”, and “U” and “V”. The figure shows additional class confusion, not evident in the latent space of fingerspelling handshapes, due to Mediapipe prediction errors. They are highlighted in yellow and are class confusion between “S” and “M” or “N”, and between “E” and “O”. On closer inspection of the actual sample poses, it is seen that

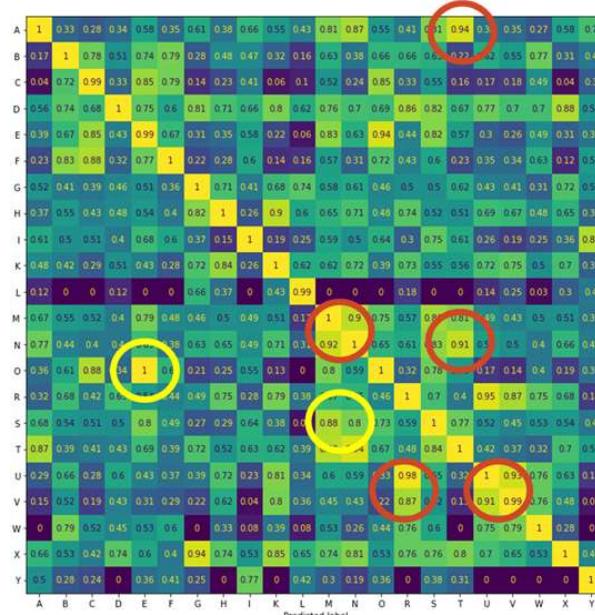


Figure 1: Sign Language class similarity matrix

Mediapipe mis-calculates the Z coordinate of these samples. This has the effect of “squeezing” the fingers of the “O” pose into an “E” pose and has brought the thumb backwards in the “S” pose to become an “N” pose. The above are errors in the YZ plane or depth plane in the 3D pose predictions. There are also errors in the frontal elevation or XY plane of the poses.

**Errors due to other skin texture behind the hand:** The output poses in figure 2 show that the skin texture of the face behind the hand is causing Mediapipe errors, with some fingertips appearing at the cheek and ear locations. In the case of the ‘N’ and ‘R’ examples, their errored poses are those of ‘A’ and ‘U’ and would result in this misclassification by any subsequent SLR model.

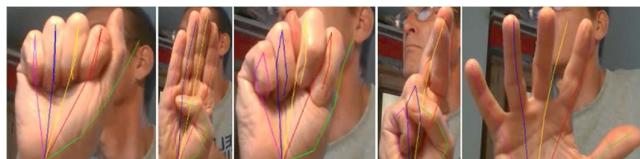


Figure 2a: ‘A’,‘B’,‘N’,‘D’,‘5’ Errors

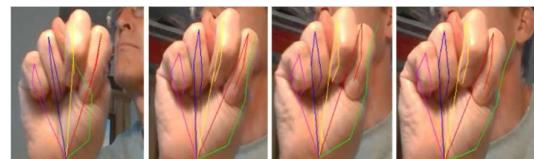


Figure 2b: ‘N’ Errors

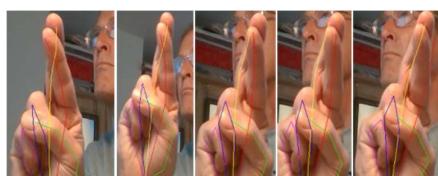


Figure 2c: ‘R’ Errors

Figures 2b and 2c show the erroneous effect for the ‘N’ Sign handshape as it moves in front of the face. The Mediapipe pose is correct before it is in front of the face. The errored thumb location gets progressively worse as it moves across the skin texture of the face. The three incorrect ‘N’ poses would be confused for the ‘A’ handshape whilst the errored ‘R’ pose would be mis-classified as a ‘U’.

### 3 Using depth sensors to correct Mediapipe errors

In order to correct such Mediapipe errors in the Z coordinate of the pose keypoints, an obvious solution is to use a depth camera to obtain the Z coordinate value directly at the computer vision capture stage of the SLR pipeline.

**Intel RealSense depth sensor:** The Intel RealSense D435 device is a structured-light stereo-vision infra-red laser depth camera [Intel, 2023]. It incorporates a colour camera with up to 1920 x 1080 RGB pixel resolution, an infrared (IR) 850 nanometre laser projector and two Full-HD resolution depth sensors with up to 1280 x 720 stereo depth resolution. It operates at up to 90 frames-per-second streaming with a wide depth diagonal field-of-view of over 90 degrees and a range of 0.2 metres to over 3 metres. The depth image pixel value represents the distance in metres from the camera to the object surface in the camera view. The image processing on the device allows captured depth and colour images to be pixel-wise aligned and their streams synchronised. This allows for the Mediapipe keypoint pixel coordinates in the colour image to be “de-projected” into world-space 3D coordinates using the depth image.

**Depth sensor noise and de-noising techniques:** Depth sensors are susceptible to the following types of image distortion [Grunnet-Jepsen & Tong, 2018]. *Shadow* is due to the triangulation technique used to accurately determine the depth when using dual-receivers. *Holes* result in the appearance of black pixel values caused by low-confidence depth calculations, camera occlusion and over-exposure. *Edges* are due to reflectivity blur around edges. *Jitter* is caused by surface reflectivity differences across the textures of IR-illuminated objects in the view. The above can be seen as black edge distortion and noise on the unfiltered examples shown in figures 3 and 6. De-noising is achieved by filtering which smooths out isolated incorrect pixel values [Tadic et al, 2020]. Edge-preserving low-pass filters can remove the high-frequency noise appearing at edges while maintaining edge definition. Hole-filling techniques use nearest neighbour pixels to fill the missing hole value. Figure 3 shows the effectiveness of de-noising when applied to the process of hand segmentation. The de-noising filtering and smoothing functions in the Realsense SDK were applied to the sensor output for all our experiments.



Figure 3: The effect of de-noising on hand segmentation for the ‘5’ and ‘S’ shapes showing the raw depth image and segmented hand image. The clearer images are de-noised.

## 4 Proposed image processing method

The proposed pre-processing method uses two techniques to correct Mediapipe errors using aligned depth images.

**To correct the Mediapipe depth error (Y-Z plane error):** The Realsense sensor incorporates an RGB camera, the frames from which can be used as input to Mediapipe to provide 3D pose data. The corresponding aligned depth frame is used to improve the Z value of these landmark estimates.

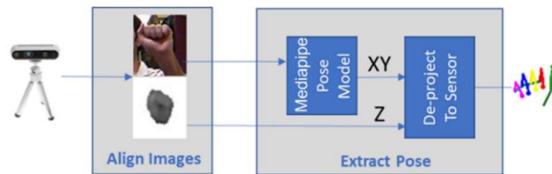


Figure 4: Depth correction pre-processing method.

The technique is to firstly “de-project” the keypoint XY pose estimate from the Mediapipe pixel coordinates to real-world units and use the depth map value as the Z coordinate instead of the pose model estimated depth value. This is intended to “correct” the Z coordinate of the pose model predicted value with the actual depth value.

**To correct the Mediapipe skin texture error (X-Y plane error):** We propose a novel hand segmentation technique to reduce confusion from other body parts appearing behind the hand in the camera view. This “depth-segmentation” assumes that the image pixel point with the minimum depth, will be on the “dominant” signing hand which will be the body part nearest to the camera.

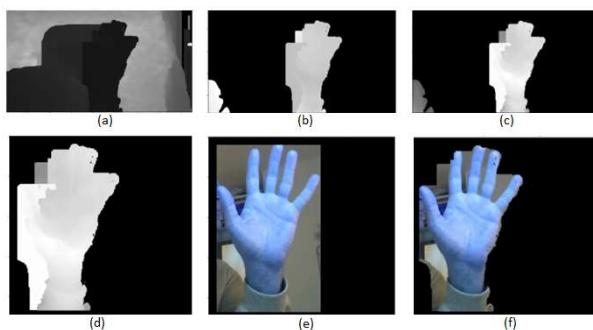


Figure 5: The image processing stages (a) through (f)

denoised raw depth image (a) is first depth-segmented by zeroing or blackening the pixels not within the signing space depth field (b). The min-depth of the image is assumed to be on the hand. Using a field-of-depth from this min-depth, the hand is segmented in the depth image (c). The largest contour in this image is then abstracted (d) and its bounding box is used to crop the hand in the corresponding colour image (e) which is aligned with the depth image. The contour in (d) is also used to segment the hand in the colour image (f).

## 5 Experimental results

While there were some improvements in specific hand shape pose estimation by Mediapipe, the overall results demonstrated that the use of depth images to improve Mediapipe performance is problematic. Two main obstacles in the success of our proposed pre-processing are the following.

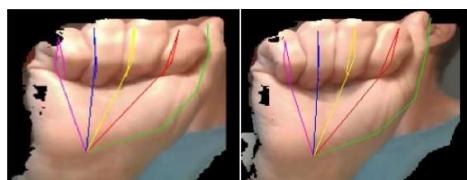


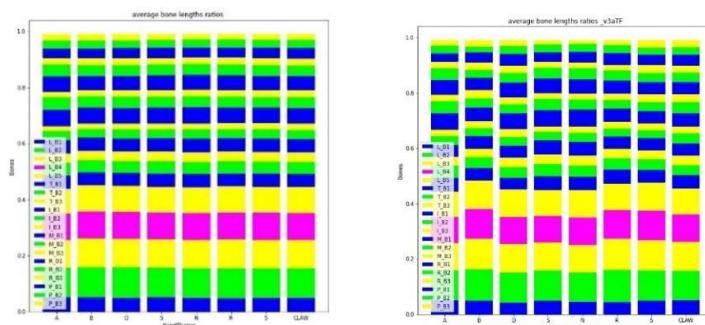
Figure 6: Residual depth image noise.

**Residual depth image distortion:** Despite de-noising and pre-processing methods applied, some images contained distortion at the edges of the segmented hand. For any predicted keypoints on XY pixel locations that are close to the edge of the hand in such an image, the de-projected depth value in that noise area will be a corrupted or zero valued. This effect can be seen in figure 6. On the left of the

figure, one of the pinky bone’s end is located in edge noise which will have an incorrect de-projected Z (depth) value. Similarly for the thumb tip in the right of the figure.

**Mediapipe tracking issue:** Mediapipe ‘tracking’ enables the pose estimation model to use previously streamed frames to temporally improve its accuracy. In order to evaluate whether our pre-processing technique results in improved Mediapipe pose estimations, we measured the resulting skeletal bone lengths of the corrected poses and compared them with the actual bone lengths of the tester. The tests were conducted on eight ASL handshape phonemes and are summarised in the charts in figure 7. The results show that the cropping and segmentation do not

lead to improvement in pose estimation. The bone lengths are more accurate when using the full colour images with Mediapipe. The change in Mediapipe input image size and resolution due to the cropping and segmentation counteract the Mediapipe tracking and results in a decrease in Mediapipe accuracy.



**Figure 7: Average bone lengths for 8 pose-sets. Left: cropped hand images. Right: segmented hand images**

texture pixels. However, the noise in the depth images causes the contour segmentation of the depth images to be unstable.

## 6 Conclusion

We have conducted an in-depth analysis of Mediapipe errors which cause mis-classifications in Sign Language Recognition models specifically. We have carried out initial investigation into mitigation measures that could be used to correct such errors. The results show that further research and experimentation is needed to provide an effective pre-processing pose correction pipeline.

## Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224, and supported, in part, by Science Foundation Ireland grant 13/RC/2094.

## References

- [Bragg et al., 2019] Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). *Sign language recognition, generation, and translation: An interdisciplinary perspective*. ASSETS '19, page 16–31.
- [Fowley et al., 2022] Fowley, F., Rushe, E., & Ventresque, A. (2022). *A Data Augmentation and Pre-processing Technique for Sign Language Fingerspelling Recognition*. In 24th Irish Machine Vision and Image Processing Conference.
- [Grunnet-Jepsen & Tong, 2018] Grunnet-Jepsen, A & Tong, D. (2018), *Depth Post-Processing for Intel® RealSense™ D400 Depth Cameras*, New Technologies Group, Intel Corporation, 2018, Rev 1.0.2.
- [Intel, 2023] Intel (2023), *Intel® RealSense™ Product Family D400 Series Datasheet, Revision 015, March 2023, Document Number: 337029-013*.
- [Kaggle, 2023] Kaggle, <https://www.kaggle.com/competitions/asl-signs/>. Last checked: 30th June, 2023.
- [Rastgoo et al., 2021] Rastgoo, R., Kiani, K., and Escalera, S. (2021). *Sign language recognition: A deep survey*. Expert Systems with Applications, 164:113794.
- [Tadic et al, 2020] Tadic, V., Burkus, E., Odry, A., Kecskes, I., Kiraly, Z., Odry, P., *Effects of the post-processing on depth value accuracy of the images captured by RealSense cameras*, Contemporary Engineering Sciences, Vol. 13, 2020, no. 1, 149-156, doi: 10.12988/ces.2020.91454
- [Wadhawan & Kumar, 2021] Wadhawan, A. & Kumar, P. (2021). *Sign language recognition systems: A decade systematic literature review*. Archives of Computational Methods in Engineering, 28, 785-813.

Figure 7 shows the average hand bone lengths for 8 pose recordings of the ‘A’, ‘B’, ‘D’, ‘S’, ‘N’, ‘R’, ‘5’, ‘CLAW’ hand shapes. The vertical values are the stacked average lengths for each of the 20 hand bones. The bone lengths for the ‘5’ handshape approximates the ground truth. The results show that the Mediapipe output for full colour image input is the most consistent and accurate. The cropped segmented hand images have blackening out behind-the-hand skin

# Investigating the Interplay Between Cervical Spine Sagittal Balance and Lower Back Pain Using Computational Biomechanics and Biomedical Imaging

Katterine Nery Rios Peralta<sup>1</sup>, David MacManus<sup>2</sup>, Michaela Davis<sup>1</sup>, Kathleen M. Curran<sup>1</sup>

<sup>1</sup>*School of Medicine, <sup>2</sup>BRAIN Lab, School of Mechanical & Materials Engineering, University College Dublin, Dublin 4, Ireland.*

## Abstract

Understanding cervical sagittal balance has emerged as a crucial diagnostic criterion due to the relationship between disability, horizontalization of the gaze, and multiple neuropathologies. While numerous studies have investigated the entire cervical complex, the distinctive morphology, and biomechanics of the first two cervical vertebrae necessitate further analysis regarding weight distribution and behavior in relation to typical vertebrae. To investigate this phenomenon, we developed a morphologically simplified finite element model of the C1-C5 segment to examine the stresses and strains experienced by the cervical spine under multi-modal loading, to simulate the influence of the upper cervical region on the other cervical vertebrae spanning from C1 to C5. Our preliminary results demonstrate the effects of spine position on the distribution of stresses and strains in the intervertebral discs which may play a key role in the causation of pathological cervical balance. Emulating the superior and anterior compression, the effective stress trend was shown in the postero-superior and the deformation in C1-C2-C3 and antero-inferior for C4-C5. Based on these findings, it can be compared with various degenerative pathophysiological processes of the cervical spine. This novel approach opens up new perspectives for investigating the interplay between cervical sagittal balance and lower back pain.

**Keywords:** Upper Cervical, Finite Element, Sagittal Balance, Biomechanics, Forward Head Posture.

## 1 Introduction

### 1.1 Cervical Sagittal Balance

The sagittal profile to achieve horizontal vision is important in all ages, and the deterioration of spinal alignment due to aging is compensated by the supportive function of the spine, pelvis and lower extremities to maintain the horizontal gaze (Hasegawa et al., 2017). The sagittal balance of the cervical spine is correlated with disability and neck pain when the forward head posture is present due to a smaller craniocervical angle considered less than 48-50° (Ruivo et al., 2014). Conversely, the craniocervical angle is often used for posture measurements, but it is rarely analysed with the general radiological parameters of sagittal balance (Le Huec et al., 2019). Nevertheless, the behavior of the cervical spine defines the position because it is part of the sagittal balance system that involves the equilibrium of the body through physiological alignment of the spine in the most efficient manner via muscular forces (Le Huec et al., 2019). To fully understand the effects of the cervical spine alignment on the thoracolumbar spine it is important to understand the underlying biomechanics and interplay between cervical and lumbar spine biomechanics. Several studies have reported that the cervical sagittal balance is related to degenerative cervical spondylosis (Paholpak, 2017; Wang et al., 2021), lumbar pain (Arima, 2021), scoliosis (Obeid et al., 2015), pelvis position (Ferrero et al., 2021), muscle degeneration (Tamai et al., 2018), risk of postoperative sagittal spinal pelvic malalignment (Passias, 2015), global spinal alignment, and horizontal gaze (Diebo et al., 2016) and amongst others. To establish a connection between these pathologies and cervical sagittal balance, it is crucial to comprehend the effects starting from the fundamental level and progressing towards a comprehensive understanding of its biotensegric mechanism. This mechanism explores the morphological complexity of the cervical spine's architecture through its geometry, recognizing it as an energy-efficient mechanism (Scarr, 2020). However, despite the fact that cervical sagittal balance has been shown to be related to a wide variety of pathophysiological processes, it has not been investigated in depth using

the finite element method.

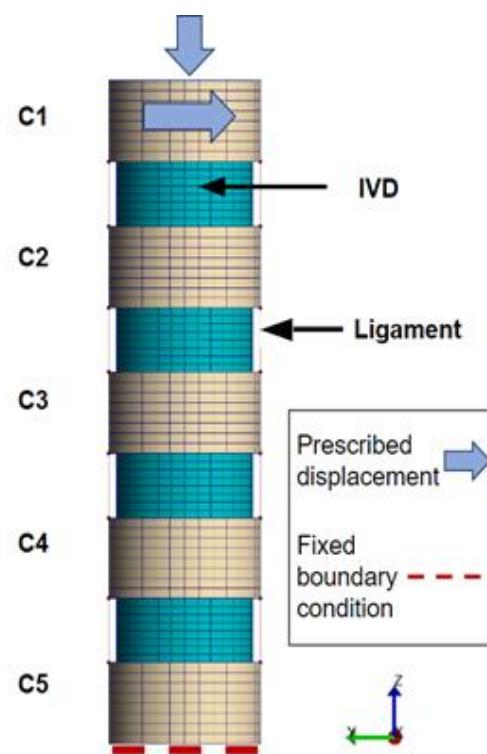
## 1.2 Biomechanics

In order to understand the cervical spine behavior it is important to recognize that the forward head posture moves the center of the body mass forward (Kim, 2022) through the particular anatomy and biomechanics of C1-C2 comprising approximately 40% of the total cervical flexion and 60% of the total cervical rotation (Frankel, 2021, Camp, 2016). Additionally, the complexity of the upper cervical region is not limited by mechanical parameters alone, it also affects the central nervous system. The head-forward individuals exhibited abnormal sensorimotor control and autonomic nervous system dysfunction compared to those with normal head alignment (Moustafa et al., 2020, Khan et al., 2020). In surgical practice, it is critical to understand that the biomechanics of the cervical spine plays an important role in maintaining correct head position after corrective surgeries of the thoracic spine (Cecchinato et al., 2014; Liu et al., 2022) and in the design of physical rehabilitation programmes. Considering the above, it is essential to understand the relationship

Between the stresses and strains that develop in the spinal components arising from cervical sagittal imbalance and how this contributes to pain in the lumbar spine region.

## 1.3 Finite Element Model of the Cervical Spine

Computer models are powerful tools to analyze the biomechanics of the human body (Crawford et al., 2003; Wang et al., 2014). The interpretation of the cervical sagittal balance effects on other spinal elements can be analyzed meticulously by calculating the stresses and strains that are produced in vertebrae, intervertebral disks, ligaments, etc. Furthermore, the joints and mechanical components of the upper cervical spine are complex and difficult to understand conventionally. Here, we used FEBio to develop a morphologically simplified model of the C1-C5 spine (Maas et al., 2012). Modeling of the cervical lordosis has been shown to be closely modeled with a circle (Harrison et al., 2004), so it is worth starting with basic geometric shapes for a simplified model of the cervical spine. The aim of this project was to develop a simple finite element model of the C1-C5 spine segment to predict the stresses and strains that develop in the cervical spine under multi-modal loading and inform future development of anatomically accurate and biofidelic models.



**Figure 1** Boundary condition and Prescribed displacement of the finite element model of the C1-C5 sagittal cervical balance

## 2 Methods

Computational modeling has been used to simulate the influence of the upper cervical spine with respect to the other cervical vertebrae from C1 to C5. This was achieved by developing a morphologically simplified finite element model of the C1-C5 spine in FEBio Studio 2.1.2 ([www.febio.org](http://www.febio.org)). A review of the literature provided information on the essential components that should be incorporated into the basic modeling of the cervical spine, including the vertebral bodies, ligaments, and intervertebral discs. In this case, simple geometric elements have been used to emulate the form and function of these components.

### 2.3 Finite Element Modelling

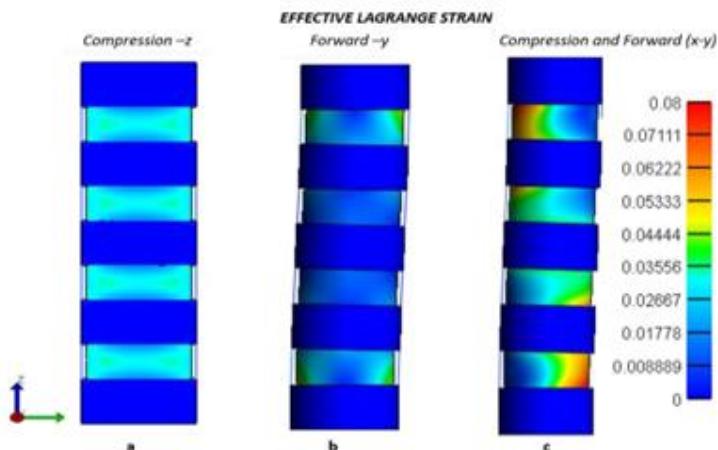
The cervical spine geometry was discretized using a 8-node hexahedral mesh. An 8-node hexahedral mesh was used as it provides excellent balance between computational accuracy, cost, and geometrical accuracy. Hexahedral elements are also more robust against element locking compared to tetrahedral elements.(Tadepalli et al., 2011) Contact interaction properties were assigned using Facet-to-Facet parameters to the superior and inferior endplate emulating the linking between endplate and intervertebral disc (IVD) surfaces. Four ligaments were modelled representing Posterior Longitudinal Ligaments and Four ligaments representing Anterior Longitudinal Ligaments.

### 2.4 Material properties

The vertebrae were modelled as an isotropic linear elastic material with a Young's modulus ( $E$ ) = 17 GPa and Poisson's ratio = 0.3. The IVDs were modelled as neo-Hookean materials with a Young's modulus = 1.2 MPa and Poisson's ratio = 0.4. Ligaments were modelled as linear springs with  $E$  = 2 MPa.

### 2.3 Boundary conditions

In order to emulate a simple cervical spine model, a fixed boundary condition was applied to the bottom vertebrae (C5). **Figure 1** A prescribed displacement of 0.7 mm in the negative Z-direction was applied to C1 to simulate compression of the spine. A second simulation was performed simulating a 2 mm lateral displacement of C1. Finally, combined compression and lateral displacement of C1 was simulated.



**Figure 2** Effective strain **a** displacement (Horn et al.), **b** displacement ( $y$ -), **c** displacement ( $z$ -y) of the finite element model of the C1-C5 sagittal cervical spine element model of the C1-C5 sagittal cervical spine.

## 3 Results

Once the model was developed the Effective stress and Effective Lagrange strain were predicted under three different loading conditions. First *compression only*, then *forward displacement only* and finally *combined compression-forward displacement*. In the top displacement (Compression) the stress distributed throughout the spine is slightly reflected in the vertebral bodies with but more intensely in the ligamentous insertions. Whereas strain is more pronounced in the IVDs, especially in the outer circumference. For the forward displacement case that are emulating the forward head position; the stress is distributed in the ligamentous insertion of the lower edge of C1 vertebrae and the upper edge of C2 vertebrae. The same is repeated for the lower edge of C4 vertebrae and upper edge of C5. The strain is observed attenuated in the IVDs of C1-C2 and C4-C5 and a minuscule deformation in the IVDs C2-C3 and C3-C4. After testing the effect separately, the analyze of both effects was performed at the same time, which would be the emulation of the C1-C2 articulation compressing and forward. Where the stress is pronounced in the posterior-inferior ligamentous insertion of C1, the complete posterior block of C2 and the posterior-superior ligamentous insertion of C3. In the lower part, the postero-inferior ligamentous insertion of C3 is detailed, the complete anterior block of C4 and a moderate part in anterior C5. Analyzing the figure of combined loads can be seen how stress and deformation are higher under a combined compression and forward displacement. This can provide information on the pathophysiological processes that lead to a process of wear due to microtraumas that can develop over time. This simplified model serves as a preliminary investigation of our hypothesis with further investigations planned with anatomically accurate FE models.

## 4 Conclusions

The simple finite element of the cervical spine produced general results on how multimodal loading might affect adjacent vertebrae. Emulating superior and anterior compression, the effective strain trend was shown in the postero-superior and the deformation in C1-C2-C3 and antero-inferior for C4-C5. Based on these findings, it can be compared with various degenerative pathophysiological processes of the cervical spine using different machine/vision imaging according to the type of pathophysiology. In addition, it can be predicted how the intervertebral discs deform in response to the anterior displacement of the upper segment. In this case, it is representing the upper cervical in the forward head posture position. This is a preliminary work where C1-C5 was analyzed to obtain information about its behavior and to be able to analyze it in the future with more precision. The model will be expanded to include the facet structure that is a load transmitter, the spinous processes that contact the stronger posterior ligaments, and the absence of C6-C7 for comprehensive analysis. Low back pain is one of the highest rates of disability and this model will provide us with new information that can be used to understand the biomechanical phenomenon that afflicts the lumbar spine and adjacent structures.

## References

- ARIMA, H. 2021. Characteristics affecting cervical sagittal alignment in patients with chronic low back pain. *Journal of Orthopaedic Science*, 26
- BOISSIE'RE, L. 2015. Cervical spine balance: postoperative radiologic changes in adult scoliosis surgery. *European Spine Journal*, 24.
- CAMP, C. L. 2016. Epidemiology, Treatment, and Prevention of Lumbar Spine Injuries in Major League Baseball Players. *The American Journal Orthopedics*, 137-143.
- CECCHINATO, R., LANGELLA, F., BASSANI, R., SANSONE, V., LAMARTINA, C. & BERJANO, P. 2014. Variations of cervical lordosis and head alignment after pedicle subtraction osteotomy surgery for sagittal imbalance. *Eur Spine J*, 23 Suppl 6, 644-9.
- CRAWFORD, R. P., ROSENBERG, W. S. & KEAVENY, T. M. 2003. Quantitative Computed Tomography-Based Finite Element Models of the Human Lumbar Vertebra Body: Effect of Element Size on Stiffness, Damage, and Fracture Strength Predictions. *Journal of Biomechanical Engineering*, 125, 434-438.
- DIEBO, B. G., CHALLIER, V., HENRY, J. K., OREN, J. H., SPIEGEL, M. A., VIRA, S., TANZI, E. M., LIABAUD, B., LAFAGE, R., PROTOPSALTIS, T. S., ERRICO, T. J., SCHWAB, F. J. & LAFAGE, V. 2016. Predicting Cervical Alignment Required to Maintain Horizontal Gaze Based on Global Spinal Alignment. *Spine (Phila Pa 1976)*, 41, 1795-1800.
- FERRERO, E., GUIGUI, P., KHALIFE, M., CARLIER, R., FEYDY, A., FELTER, A., LAFAGE, V. & SKALLI, W. 2021. Global alignment taking into account the cervical spine with odontoid hip axis angle (OD-HA). *Eur Spine J*, 30, 3647-3655.
- FRANKEL, M. N. A. V. 2021. *Basic Biomechanics of the Musculoskeletal System*, Wilters Kluwer
- HARRISON, D. D., HARRISON, D. E., JANIK, T. J., CAILLIET, R., FERRANTELLI, J. R., HAAS, J. W. & HOLLAND, B. 2004. Modeling of the sagittal cervical spine as a method to discriminate hypolordosis: results of elliptical and circular modeling in 72 asymptomatic subjects, 52 acute neck pain subjects, and 70 chronic neck pain subjects. *Spine (Phila Pa 1976)*, 29, 2485-92.
- HASEGAWA, K., OKAMOTO, M., HATSUSHIKANO, S., SHIMODA, H., ONO, M., HOMMA, T. & WATANABE, K. 2017. Standing sagittal alignment of the whole axial skeleton with reference to the gravity line in humans. *Journal of Anatomy*, 230, 619-630.
- KIM D, D. D., MENGER RP. 2022 Spine Sagittal Balance. . StatPearls [Internet] Treasure Island (FL); 2022 Jan: StatPearls Publishing.
- KHAN, A., KHAN, Z., BHATI, P. & HUSSAIN, M. E. 2020. Influence of Forward Head Posture on Cervicocephalic Kinesthesia and Electromyographic Activity of Neck Musculature in Asymptomatic Individuals. *J Chiropr Med*, 19, 230-240.
- LE HUEC, J. C., THOMPSON, W., MOHSINALY, Y., BARREY, C. & FAUNDEZ, A. 2019. Correction to: Sagittal balance of the spine. *Eur Spine J*, 28, 2631.
- LIU, Y., LIU, J., LUO, D., SUN, J., LV, F. & SHENG, B. 2022. Focusing on the amount of immediate changes in spinopelvic radiographic parameters to predict the amount of mid-term improvement of quality of life in adult degenerative scoliosis patients with surgery. *Arch Orthop Trauma Surg*.
- MAAS, S. A., ELLIS, B. J., ATESHIAN, G. A. & WEISS, J. A. 2012. FEBio: finite elements for biomechanics. *J Biomech Eng*, 134, 011005.
- MOUSTAFA, I. M., YOUSSEF, A., AHBOUCH, A., TAMIM, M. & HARRISON, D. E. 2020. Is forward head posture relevant to autonomic nervous system function and cervical sensorimotor control? Cross sectional study. *Gait Posture*, 77, 29-35.
- PAHOLPAK, P. 2017. Kinematic evaluation of cervical sagittal balance and thoracic inlet alignment in degenerative cervical spondylolisthesis using kinematic magnetic resonance imaging. *The Spine Journal*, 17.
- PASSIAS, P. G. 2015. Magnitude of preoperative cervical lordotic compensation and C2-T3 angle are correlated to increased risk of postoperative sagittal spinal pelvic malalignment in adult thoracolumbar deformity patients at 2-year follow-up. *The Spine Journal*, 15.
- RUIVO, R. M., PEZARAT-CORREIA, P. & CARITA, A. I. 2014. Cervical and shoulder postural assessment of adolescents between 15 and 17 years old and association with upper quadrant pain. *Braz J Phys Ther*, 18, 364-71.
- SCARR, G. 2020. Biotensegrity: What is the big deal? *Journal of Bodywork & Movement Therapies*, 24.
- TAMAI, K., ROMANU, J., GRISDELA, P., PAHOLPAK, P., ZHENG, P., NAKAMURA, H., BUSER, Z. & WANG, J. C. 2018. Small C7-T1 lordotic angle and muscle degeneration at C7 level were independent radiological characteristics of patients with cervical imbalance: a propensity score-matched analysis. *Spine Journal*, 18, 1505-1512.
- TADEPALLI, S. C., ERDEMIR, A. & CAVANAGH, P. R. 2011. Comparison of hexahedral and tetrahedral elements in finite element analysis of the foot and footwear. *J Biomech*, 44, 2337-43.
- WANG, W., BARAN, G. R., BETZ, R. R., SAMDANI, A. F., PAHYS, J. M. & CAHILL, P. J. 2014. The Use of Finite Element Models to Assist Understanding and Treatment For Scoliosis: A Review Paper. *Spine Deform*, 2, 10-27.
- WANG, Z., XU, J. X., LIU, Z., WANG, Z. W., DING, W. Y. & YANG, D. L. 2021. Spino Cranial Angle and Degenerative Cervical Spondylolisthesis. *World Neurosurg*, 151, e517-e522.

# Towards a performance analysis on pre-trained Visual Question Answering models for autonomous driving

Kaavya Rekanar<sup>1</sup>, Ciarán Eising<sup>1</sup>, Ganesh Sistu<sup>2</sup>, Martin Hayes<sup>1</sup>

<sup>1</sup>*University of Limerick*, <sup>2</sup>*Valeo Vision Systems*

<sup>1</sup>*firstname.lastname@ul.ie*, <sup>2</sup>*firstname.lastname@valeo.com*

## Abstract

This short paper presents a preliminary analysis of three popular Visual Question Answering (VQA) models, namely ViLBERT, ViLT, and LXMERT, in the context of answering questions relating to driving scenarios. The performance of these models is evaluated by comparing the similarity of responses to reference answers provided by computer vision experts. Model selection is predicated on the analysis of transformer utilization in multimodal architectures. The results indicate that models incorporating cross-modal attention and late fusion techniques exhibit promising potential for generating improved answers within a driving perspective. This initial analysis serves as a launchpad for a forthcoming comprehensive comparative study involving nine VQA models and sets the scene for further investigations into the effectiveness of VQA model queries in self-driving scenarios. Supplementary material is available on the Github page.

**Keywords:** Visual Question Answering, Transformers, Performance Analysis, Multi-modal Models

## 1 Introduction

Visual Question Answering (VQA) is the process of generating natural language responses to open-ended questions by leveraging visual information derived from an image. This task encompasses the generation of textual answers to queries expressed in natural language. Visual question answering (VQA) holds significant importance for self-driving cars due to the requirement for enhanced perception and decision-making in autonomous vehicles. By incorporating VQA systems into the framework of self-driving cars, key benefits like Contextual Understanding, Enhanced Human-Machine Interaction, Adaptive Decision-Making, and Safety and Error Handling can be realized [Xiong et al., 2022]. By integrating VQA capabilities, autonomous vehicles can enhance their perception, communication, and decision-making processes, ultimately leading to safer and more efficient driving experiences.

This paper provides an introductory overview of the analysis conducted on three select models, focusing specifically on their performance in the domain of Visual Question Answering (VQA) with a strict focus on driving scenarios. It is part of a research study aimed at identifying the most effective VQA model for answering questions related to driving. Although there are review papers available on VQA models [Zhong et al., 2022], there is a notable research gap, as none of these studies has conducted a model evaluation in the context of common driving scenarios. A survey has been conducted to observe how pretrained available models respond to questions and how similar or different the answers are when compared to humans. The comparative analysis done has led us to the result that the available models are not as suitable for questions in a driving context as they are in a general scenario and this is a research gap that could be exploited. Additionally, the authors observed that there has not been a thorough performance analysis conducted on this topic.

## 2 Background Study and Related Work

VQA models incorporate multimodal architectures that utilize transformers to handle the fusion of visual and textual modalities. Transformers enable contextual understanding and information exchange between the visual and textual components of the input, facilitating more accurate and comprehensive question answering [Vaswani et al., 2017]. Therefore, multimodal models employ transformers to process and fuse information from different modalities. Within the domain of VQA, multimodal models leverage transformers to handle the integration of visual and textual information, allowing for enhanced understanding and improved performance in answering questions based on visual inputs.

Transformers in multimodal models with vision and NLP refer to the application of transformer-based architectures in tasks that involve both visual and textual information. Transformers have demonstrated great success in natural language processing (NLP) tasks, thanks to their ability to capture long-range dependencies and model sequential data effectively. However, the integration of visual information poses unique challenges, and incorporating it into transformers allows for more powerful multimodal models.

Traditionally, multimodal models combined visual and textual information using separate pathways, such as using convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for language processing. Transformers offer an alternative approach that enables joint processing of both modalities within a unified architecture.

Transformers are utilized in multimodal models for early fusion, late fusion, and cross-modal attention according to [Nagrani et al., 2021]. Early fusion involves the simultaneous processing of modalities to learn joint representations. Late fusion includes separate processing of modalities, followed by fusion to capture interdependencies. Cross-modal attention enables information exchange and alignment between modalities, enhancing multimodal understanding and integration. More details on how transformers are utilized in each of these types of models can be read in [Boulahia et al., 2021] and [Nagrani et al., 2021].

### 3 Methodology

In this study, we collected a comprehensive corpus of 78 research papers on Visual Question Answering (VQA)<sup>1</sup>. From this collection, we carefully selected nine models based on specific criteria for our analysis. These models were evaluated for user interface quality, code replication ease, and compatibility with our pretrained models. The initial experiment aimed to enhance the models' performance using the German Traffic Sign Recognition Benchmark (GTSRB) dataset, explicitly focusing on signboard interpretation [Stallkamp et al., 2011]. However, the results revealed limited comprehension of driving-related matters by the models. This led us to conduct an additional experiment with computer vision experts, presenting them with contextually minimal images from our dataset, mirroring the approach used with the pre-trained models.

For our experiment comparing human responses to multimodal models, we selected three models solely based on their utilization of transformers in their architectures from the previously mentioned nine models. A brief introduction about the three models chosen for the analysis:

- Vision and Language BERT (ViLBERT)- Early Fusion: extends BERT with a co-attention mechanism, integrating vision-attended language features into visual representations. It enables joint reasoning about text and images for visual grounding [Lu et al., 2019].
- Vision-and-Language Transformer (ViLT)- Cross-Modal Attention: aligns visual and textual features and generates joint representations through a visual encoder and a language encoder [Kim et al., 2021].
- Learning Cross-Modality Encoder Representations from Transformers (LXMERT)- Late Fusion: incorporates multi-level interactions between vision and language by employing cross-attention mechanisms. It captures the interplay between different modalities and generates accurate answers to the posed questions [Tan and Bansal, 2019].

The authors conducted a survey consisting of two specific questions, namely "What are the contents of the image?" and "What should the driver do?", targeting a carefully chosen set of images all pertaining to driving scenarios. These images were selected from the MS COCO dataset. The survey was distributed among a cohort of ten Computer Vision Experts who provided responses to the questions based on the available options and the accompanying images. The answer that received the most votes was selected as the ground truth. The comprehensive outcomes of this survey are presented in Figure 1.

The rationale behind asking both subjective and objective questions, namely "What are the contents of the image?" and "What should the driver do?", is to assess the model's ability to comprehend and respond to different types of questions in the context of visual information. Subjective questions, like "What are the contents of the image?", require the model to understand and interpret the visual content and provide a descriptive answer. These questions evaluate the model's capability to recognize objects, scenes, and other relevant visual elements depicted in the image. Objective questions, like "What should the driver do?", require the model to provide a specific action or response based on the given visual information. These questions assess the model's understanding of driving scenarios and ability to reason about the appropriate course of action.

<sup>1</sup>Full list of papers available at our Github page.

By including both subjective and objective questions, the experiment aims to evaluate different aspects of the model's performance. Subjective questions focus on the model's visual comprehension and scene understanding abilities, while objective questions assess its ability to provide contextually appropriate and practical responses in a driving context. This comprehensive evaluation helps to gauge the model's overall proficiency in visual question answering and its potential utility in real-world applications such as self-driving cars.

The ground truth for the respective questions was evaluated against the answers generated by the pre-trained models. Figure 1 provides a visual representation of the model's performance in addressing the posed questions, allowing for an assessment of their effectiveness based on the ground truth. The rationale behind comparing the answers of three Visual Question Answering (VQA) models with human answers and using colour coding (green for correct, orange for wrong, yellow for partially correct) is to visually highlight the performance and discrepancies between the models and human responses as done in [Dzelzkaleja, 2020]. This visual representation allows for a quick and intuitive understanding of the accuracy and effectiveness of the models in comparison to human performance.

## 4 Results and Analysis

Figure 1 concisely summarises the results from the experiment conducted on the selected models.

Category	Image	Questions	CV Experts	VILBERT	ViLT	LXMERT
Dark setting		what are the contents of the image?	An Accident	Trucks	Cars	Cars
		what should the driver do?	Slow down and go left	Run	Stop	Go
Light Setting		what are the contents of the image?	Traffic Lights	Clouds	Traffic Lights	Power Lines
		what should the driver do?	Stop	Sleep	Stop	Stop
Parking		what are the contents of the image?	Parking Lot	Clouds	Cars	Cars
		Can the driver park here?	Yes	No	No	Yes
Signboard		what are the contents of the image?	Road	Paint	Traffic Lights	Cars
		what should the driver do?	Stop	Sleep	Stop	Go
Pedestrian Crossing		what are the contents of the image?	People crossing the road	Clouds	Buildings	People
		what should the driver do?	Stop	Run	Stop	Stop
Traffic		what are the contents of the image?	Traffic	Trucks	Buses	Buses
		what should the driver do?	Go	Run	Stop	Go
Accident		what are the contents of the image?	An Accident	Clouds	Cars	Cars
		what should the driver do?	Stop	Stop	Stop	Stop

Figure 1: Comparison of Responses: Computer Vision Experts vs. Selected Models

### 4.1 Analysis

In summary, the analysis of the evaluated models to date yields the following observations: ViLBERT demonstrates a lack of comprehension regarding the question "What are the contents of the image?", as it consistently provided the answer "nothing". However, when posed with the question "What are the objects of the image?", ViLBERT manages to produce answers, albeit quite often incorrect or of limited utility for the application at hand. Consequently, ViLBERT is not an optimal choice for fine-tuning within the context of self-driving scenarios. ViLT exhibits a certain level of capability in generating answers based on the provided images. Notably, when addressing the question "What should the driver do?", ViLT frequently responds with "Stop," as depicted in Figure 1. However, upon further investigation, it becomes apparent that the model does perform well in terms of question comprehension, and its object identification performance surpasses that of ViLBERT. This finding suggests that ViLT holds promise for fine-tuning with the GTSRB dataset, enabling it to learn how to effectively answer questions within a driving context. The LXMERT model demonstrates better performance in answering questions within a driving context. Although the model exhibits excellent object identification capabilities, the accuracy of its answers requires refinement. The authors noted that LXMERT's object identification algorithm effectively recognizes objects in various scenes, and can offer accurate scene descriptions with the important exception of accident-related images. This observation implies there exists potential to enhancing LXMERT's performance in driving scenarios through fine-tuning with the GTSRB dataset, thereby improving its performance in driving-specific use cases.

## 5 Conclusions and Future Work

This paper has reviewed the performance of three VQA models, namely ViLBERT, ViLT, and LXMERT, from a driver assistance perspective, focusing on model efficacy in terms of similarity to expert responses for posed questions. Based on the analysis presented in this paper, it is inferred that both ViLT and LXMERT exhibit promising performance in this application space. However, despite the advancements observed in these models, further research and development are required to address the specific challenges associated with driver assistance. The ability to accurately comprehend and respond to user queries in real-time scenarios remains a crucial aspect of enhancing the interaction between drivers and vehicles. Achieving a VQA model that can effectively interpret diverse driver inquiries, provide accurate answers, and adapt to dynamic driving conditions is essential for optimizing user-car interaction.

Moving forward, the work will expand its scope by conducting a more comprehensive performance analysis that considers six additional selected models, including basic and fine-tuned pretrained models using the GTSRB dataset. The ultimate objective is to identify a preferred model that can be extensively trained with an expanded dataset that encompasses driving scenarios with subjective reference responses. Future research will focus on providing better codified contextual information to both experts and models, including camera location, velocity, acceleration, handbrake, and steering inputs, to enable more informed assessment of performance and to enable decisions on the next highest priority action to be taken with greater confidence.

## Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Boulahia et al., 2021] Boulahia, S. Y., Amamra, A., Madi, M. R., and Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121.
- [Dzelzkaleja, 2020] Dzelzkaleja, L. (2020). Color code method design evaluation and data analysis. *International Journal of Engineering & Technology*, 7(2.28):106–109.
- [Kim et al., 2021] Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- [Lu et al., 2019] Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- [Nagrani et al., 2021] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.
- [Stallkamp et al., 2011] Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2011). The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE.
- [Tan and Bansal, 2019] Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [Xiong et al., 2022] Xiong, W., Fan, H., Ma, L., and Wang, C. (2022). Challenges of human—machine collaboration in risky decision-making. *Frontiers of Engineering Management*, 9(1):89–103.
- [Zhong et al., 2022] Zhong, Y., Ji, W., Xiao, J., Li, Y., Deng, W., and Chua, T.-S. (2022). Video question answering: datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.

# Evaluate Fine-tuning Strategies for Fetal Head Ultrasound Image Segmentation with U-Net

Fangyijie Wang<sup>1</sup>, Guénolé Silvestre<sup>2</sup>, Kathleen M. Curran<sup>1</sup>

<sup>1</sup>*School of Medicine, University College Dublin, Dublin, Ireland*

<sup>2</sup>*School of Computer Science, University College Dublin, Dublin, Ireland*

## Abstract

Fetal head segmentation is a crucial step in measuring the fetal head circumference (HC) during gestation, an important biometric in obstetrics for monitoring fetal growth. However, manual biometry generation is time-consuming and results in inconsistent accuracy. To address this issue, convolutional neural network (CNN) models have been utilized to improve the efficiency of medical biometry. But training a CNN network from scratch is a challenging task, we proposed a Transfer Learning (TL) method. Our approach involves fine-tuning (FT) a U-Net network with a lightweight MobileNet as the encoder to perform segmentation on a set of fetal head ultrasound (US) images with limited effort. This method addresses the challenges associated with training a CNN network from scratch. It suggests that our proposed FT strategy yields segmentation performance that is comparable when trained with a reduced number of parameters by 85.8%. And our proposed FT strategy outperforms other strategies with smaller trainable parameter sizes below 4.4 million. Thus, we contend that it can serve as a dependable FT approach for reducing the size of models in medical image analysis. Our key findings highlight the importance of the balance between model performance and size in developing Artificial Intelligence (AI) applications by TL methods. Code is available at [https://github.com/13204942/FT\\_Methods\\_for\\_Fetal\\_Head\\_Segmentation](https://github.com/13204942/FT_Methods_for_Fetal_Head_Segmentation).

**Keywords:** Medical Imaging, Transfer Learning, Ultrasound, Biometry, Convolutional Neural Network.

## 1 Introduction

Training a deep CNN from scratch can prove to be a formidable undertaking, particularly in medical applications that are often constrained by limited annotated data and require a substantial time investment. However, Transfer Learning (TL) can help alleviate these challenges. TL is a technique in which a network learns from a large dataset and then applies that knowledge to another application, typically a smaller dataset. This approach can be especially advantageous in medical applications where annotated data is scarce, as it permits the utilization of pre-trained models to enhance performance on smaller datasets. TL approaches entail the adoption of pre-trained models and fine tuning (FT).

In this study, we conducted a segmentation task on fetal head US images using deep neural networks with various FT strategies. The dataset HC18 comprises of 1334 ultrasound images obtained from 551 pregnant women and is publicly available [van den Heuvel et al., 2018]. To perform semantic segmentation on the HC18 fetal head US images, we performed the FT of the U-Net [Ronneberger et al., 2015] network, with a pre-trained MobileNet [Howard et al., 2018] as its backbone. In order to develop a lightweight model using FT techniques, this research work considered a comparison of model sizes for various pre-trained CNN models. Furthermore, we investigated the impact of FT on different decoder layers for fetal head segmentation. In terms of segmentation outcomes on tests, the results were competitive in comparison to the state-of-the-art (SOTA) results, 97% ( $\pm 0.3\%$ ) achieved by [Amiri et al., 2020] with FT the encoder. Our research is of significance when analyzing the trade-off between performance and model size in the development of mobile AI applications.

The main contributions of this paper are as follows: (1) We analyzed eight different fine-tuning strategies on a U-Net network that used a MobileNet V2 encoder to predict segmentation masks from a fetal head ultrasound dataset. (2) We achieved SOTA accuracy on the HC18 Grand Challenge by providing a pre-trained U-Net model that had only 4.4 million trainable parameters. (3) Our experiments showed that

unfreezing the decoder of a pre-trained U-Net network was the most effective fine-tuning strategy compared to the others we tested.

## 2 Related Work

In recent years, DL techniques have been developed to achieve high precision outcomes in semantic segmentation tasks. Ronneberger et al. [Ronneberger et al., 2015] proposed the U-Net architecture to perform biomedical image segmentation tasks with annotated samples more efficiently. In 2019, Howard et al. [Howard et al., 2018] constructed MobileNet V2 for semantic segmentation by making use of lightweight depth-wise separable convolutions to filter features. Therefore, it has a lower computational cost, less memory, and consumes less power. As a result, MobileNet V2 is a low-cost, efficient deep neural network suitable for mobile and embedded vision applications.

In terms of US image segmentation tasks, [Amiri et al., 2020] employs TL techniques to overcome limited and costly data issues in DL for medical applications. The authors investigate the impact of FT various layers of a pre-trained U-Net and assess their performance in fetal US image segmentation tasks on the HC18 US dataset. Their FT strategies consist of three schemes, FT shallow, deep layers, and the entire network. Across all US datasets analyzed in their work, FT the entire pre-trained U-Net yielded better results than training from scratch. [Cheng and Lam, 2021] utilizes cross-domain TL with U-Net architecture for precise and fast image segmentation. The cross-domain TL techniques are utilized in [Monkam et al., 2023] for the purpose of fetal head segmentation on HC18. The researchers have proposed a speedy and efficient method to produce a considerable number of annotated US images, based on a limited number of manually annotated biometrics. Besides cross-domain TL techniques, Alzubaidi et al. [Alzubaidi et al., 2022] demonstrated an ensemble TL technique with a segmentation model that includes eight CNN models. This technique is evaluated on the US dataset HC18 by achieving 98.53% mIoU. However, the ensemble TL model has 28.12 million trainable parameters, which is 7 times more than the best model we proposed with 4.4 million trainable parameters. [Kim et al., 2022] provides an overview study of TL methods on medical image classification. They demonstrated the efficacy of TL. The authors suggest that utilizing CNN models as feature extractors can save computational costs. Inspired by the investigation from [Kim et al., 2022], we think similar FT methods can be utilized in medical image segmentation.

Our proposed FT strategy achieved competitive head segmentation results on HC18 with fewer trainable parameters and training epochs compared to existing SOTA methods, see Figure 1b. The U-Net is a strong CNN architecture widely applied in medical image analysis. The most notable segmentation outcomes on the present HC18 leaderboard were obtained by leveraging U-Net and its expansion networks. Hence, we utilize U-Net architecture to construct a CNN model and evaluate our FT strategies.

## 3 Methodology

**Data Preparation:** The HC18 dataset comprises a two-dimensional (2D) US image collection that has been split into 999 images for training purposes and 335 images for testing. All HC18 images are standard planes that are suitable for measuring fetal HC. Each of these images is of dimensions 800 by 540 pixels. Because these 999 images were annotated with biometrics by experienced medical experts, they were selected as the experimental dataset, whereby 799 images and 200 images were assigned for training and testing, respectively. 999 images were resized to  $512 \times 512$  pixels. In this study, we used the standard data-augmentation techniques: rotation by an angle from  $[-25^\circ, 25^\circ]$ , horizontal flipping, vertical flipping, and pixel normalization.

**Model Design:** In our work, based on Ronneberger's work [Ronneberger et al., 2015], we built a U-Net baseline model with 4 encoder layers, 4 decoder layers and 1 bottleneck. The model has input features [64, 128, 256, 512]. We apply a MobileNet V2 model to the U-Net's encoder part. The MobileNet V2 model was pre-trained on dataset ImageNet.

**Fine-tuning Strategies:** Our FT methods include a collection of seven distinct schemes, see Figure 1a. The baseline U-Net model has no pre-trained encoder and all layers unfrozen. The FT methods include training the entire decoder, the entire encoder, 0 layer within decoder, 0,1 layers within decoder, 0,1,2 layers within decoder, 2,3,4 layer within decoder, and 4 layer within decoder. In the baseline U-Net model, the encoder is not pre-trained and all layers remain unfrozen. The FT methods are comprised of a range of

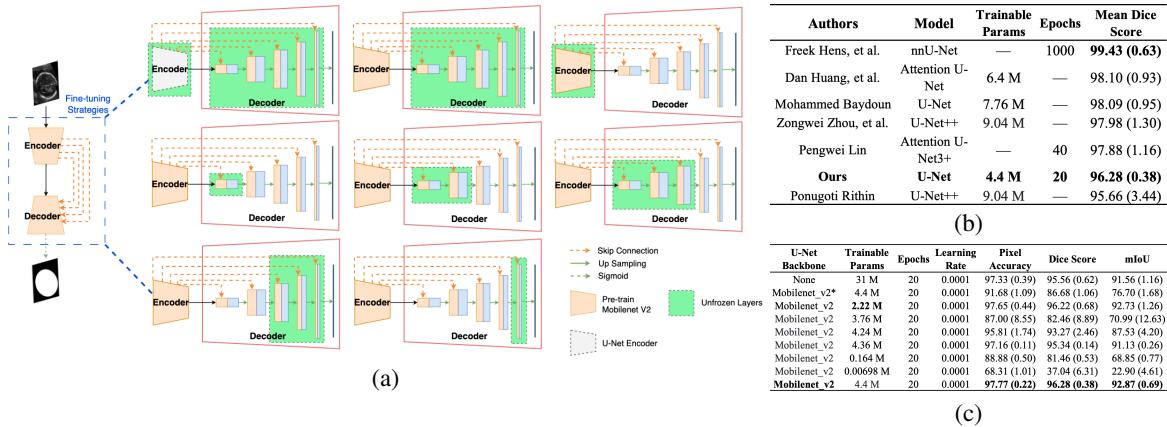


Figure 1: (a) The first row shows three fine-tuning strategies: U-Net baseline, 0 to 4 layers remain unfrozen within the decoder, and the encoder remains unfrozen. The second row shows three fine-tuning strategies: 0 layer remains unfrozen within the decoder, 0 to 1 layers remain unfrozen within the decoder, and 0 to 2 layers remain unfrozen within the decoder. The last row shows two fine-tuning strategies: 2 to 4 layers remain unfrozen within the decoder, 4 layer remains unfrozen within the decoder. (b) Comparison of our methods with the SOTA methods. (c) Comparison of Pixel Accuracy, Dice Score, and mIoU on Test data set. Mobilenet\_v2\* is the encoder with random weights.

techniques, including training the entire decoder, the entire encoder, the layer 0 within the decoder, layers 0 to 1 within the decoder, layers 0 to 2 within the decoder, layers 2 to 4 within the decoder, and the layer 4 specifically within the decoder. In all experiments, the training and testing operations are executed four times repeatedly.

**Training and Evaluation:** We implemented all of our experiments using Pytorch. After comparing performance between different CNN architectures, we train a U-Net model on HC18 from scratch by using Segmentation Models [Iakubovskii, 2019]. We trained the U-Net model with 20 epochs from scratch. Each epoch took around 75 seconds. Also, we fine-tuned the pre-trained U-Net with MobileNet V2 encoder with 20 epochs. Each epoch took around 25 seconds. The training dataset and test dataset both have a batch size of 10. The Adam optimiser was used in training processes with a decaying learning rate of  $1e - 4$ . All training processes were performed on an NVIDIA Tesla T4 graphics card. The typical metrics applied to evaluate the performance of segmentation models are Pixel Accuracy (PA), Dice coefficient, and Mean Intersection over Union (IoU). Mean IoU is defined as the average IoU over all classes  $K$ .

## 4 Experimental Results

Figure 1c summarises the segmentation metrics achieved through the implementation of various FT strategies on the HC18 test set, 200 fetal US images. The act of unfreezing the entire decoder within the pre-trained U-Net model has contributed to the generation of more accurate predictions on segmentation masks when compared to both the U-Net baseline model and other FT strategies. Our proposed FT strategy improved PA, Dice score, and mIoU by 0.45%, 0.75%, and 1.4% respectively when compared to training our U-Net baseline from scratch. Furthermore, the size of trainable parameters has been reduced by 85.8%. Despite the fact that the size of trainable parameters for other FT strategies is smaller than 4.4 million, our proposed FT strategy outperformed their evaluation results. In comparison to Amiri's methods, our proposed FT strategy has also yielded a 1.24% increase in their results (95.1%) [Amiri et al., 2020] in terms of Dice score. Another FT strategy involving training U-Net pre-trained encoder only has also shown competitive results with a 96.22% Dice score.

## 5 Conclusion

We presented a FT strategy for a pre-trained U-Net that enables accurate fetal head segmentation in US images while utilizing only 4.4 million parameters. To evaluate the effectiveness of various fine-tuning approaches, we conducted experiments on the HC18 Grand Challenge dataset. Our findings suggest that

utilizing a pre-existing network enhances segmentation precision, whereas augmenting the amount of trainable parameters does not significantly impact accuracy. To reduce model size and the number of trainable parameters, we used the MobileNet V2 model as the encoder in our U-Net. Our fine-tuned model has significantly reduced 85.8% trainable parameters in comparison to training an initialized U-Net. Our research suggests that the ideal approach for FT is to adjust the decoder's 0, 1, 2, 3, 4 layers of the pre-trained U-Net based on our experiments. This methodology yielded a PA of 97.77%, a Dice coefficient of 96.28%, and a mIoU of 92.87% on the HC18 test dataset. Alternatively, FT the U-Net pre-trained encoder only is another TL method producing competitive results potentially. Our findings propose that adjusting the decoder of the U-Net might serve as an efficient approach for FT small models in US image analysis.

## 6 Future Work

Future research may be conducted in order to reduce noise on US images by introducing image processing methods. And we will further investigate the resilience of the model that has been trained by TL techniques. Furthermore, we intend to investigate alternative pre-trained models in order to achieve an optimized model that is smaller in size.

## Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [Alzubaidi et al., 2022] Alzubaidi, M., Agus, M., Shah, U., Makhlof, M., Alyafei, K., and Househ, M. (2022). Ensemble transfer learning for fetal head analysis: From segmentation to gestational age and weight prediction. *Diagnostics*, 12(9).
- [Amiri et al., 2020] Amiri, M., Brooks, R., and Rivaz, H. (2020). Fine-tuning U-Net for ultrasound image segmentation: Different layers, different outcomes. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(12):2510–2518.
- [Cheng and Lam, 2021] Cheng, D. and Lam, E. Y. (2021). Transfer learning U-Net deep learning for lung ultrasound segmentation. *arXiv*, 2110.02196.
- [Howard et al., 2018] Howard, A., Zhmoginov, A., Chen, L.-C., Sandler, M., and Zhu, M. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In *CVPR*.
- [Iakubovskii, 2019] Iakubovskii, P. (2019). Segmentation models pytorch. [https://github.com/qubvel/segmentation\\_models.pytorch](https://github.com/qubvel/segmentation_models.pytorch).
- [Kim et al., 2022] Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):69.
- [Monkam et al., 2023] Monkam, P., Jin, S., and Lu, W. (2023). Annotation cost minimization for ultrasound image segmentation using Cross-Domain transfer learning. *IEEE Journal of Biomedical and Health Informatics*, 27(4):2015–2025.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *LNCS: Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241. Springer.
- [van den Heuvel et al., 2018] van den Heuvel, T. L. A., de Bruijn, D., de Korte, C. L., and van Ginneken, B. (2018). Automated measurement of fetal head circumference using 2D ultrasound images. *PLoS One*, 13(8):e0200412.

# Hardware Accelerators in Autonomous Driving

Ken Power<sup>1,2</sup> , Shailendra Deva<sup>1</sup>, Ting Wang<sup>1</sup>, Julius Li<sup>1</sup>, Ciarán Eising<sup>2</sup>

<sup>1</sup>*Motional AD Inc, USA.* <sup>2</sup>*University of Limerick, Ireland.*

## Abstract

Computing platforms in autonomous vehicles record large amounts of data from many sensors, process the data through machine learning models, and make decisions to ensure the vehicle's safe operation. Fast, accurate, and reliable decision-making is critical. Traditional computer processors lack the power and flexibility needed for the perception and machine vision demands of advanced autonomous driving tasks. Hardware accelerators are special-purpose coprocessors that help autonomous vehicles meet performance requirements for higher levels of autonomy. This paper provides an overview of ML accelerators with examples of their use for machine vision in autonomous vehicles. We offer recommendations for researchers and practitioners and highlight a trajectory for ongoing and future research in this emerging field.

**Keywords:** Machine Vision, Perception, Autonomous Vehicles, Machine Learning, Hardware Accelerators

## 1 Introduction

Autonomous vehicles are part of the field of mobile autonomous robots as much as they are part of the automotive domain [Correll *et al.* 2022]. Autonomous taxis, also called autonomous ride-hailing, or robotaxis, are one of the significant emerging markets and opportunities for autonomous vehicles [Li *et al.* 2022]. Advances in machine learning and deep learning directly contribute to advances in vehicle autonomy. ML accelerators for autonomous driving are enabling new levels of autonomy and performance. Section 2 of this paper provides an overview of machine learning accelerators in the context of autonomous driving. This overview includes architecture styles for ML accelerators. Section 3 provides examples of how ML accelerators improve the performance of ML workloads in AVs, with a focus on machine vision use cases. Finally, this paper summarises the implications for machine vision and AV development, offers some recommendations for researchers and practitioners, and outlines a trajectory for future research in ML accelerators.

## 2 Background

Robots are autonomous when they make decisions in response to their environment rather than following pre-programmed instructions [Correll *et al.* 2022]. They achieve autonomy using multiple techniques, including signal processing, control theory, and artificial intelligence [Correll *et al.* 2022]. An autonomous vehicle is a safety-critical system with constraints on timing, scheduling, performance, and safety that are not present in other forms of robots [Jo *et al.* 2014]. The Society of Automotive Engineers (SAE) defines six levels of driving autonomy (Level 0 through Level 5) for vehicles that perform part or all of the *dynamic driving task* (DDT) on a sustained basis [SAE International 2021]. A human driver must be present in Levels 0 through 3. Level 4 (L4) and Level 5 (L5) autonomous driving systems, where there is no human driver, must operate autonomously and safely in complex and diverse conditions [SAE International 2021]. Autonomous robots need to sense and perceive their environment,

which they achieve using various sensors that measures some aspects of the environment [Ben-Ari and Mondada 2018]. The most common sensors for sensing and perception are cameras, radar, and lidar.

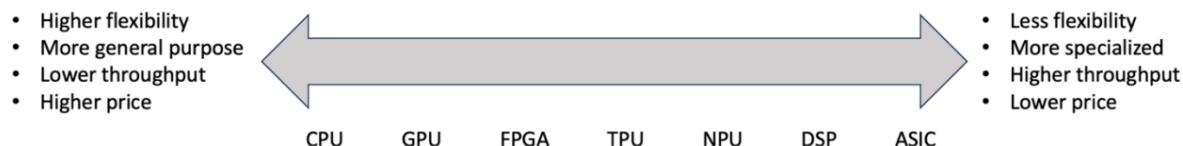
Machine learning is increasingly adopted to enable autonomous vehicle functionality [Gharib *et al.* 2018]. In the context of autonomous driving, machine learning is a crucial tool for enabling vehicles to recognize and respond to the complex and dynamic environments in which they operate [Bachute and Subhedar 2021]. By leveraging large datasets and powerful algorithms, machine learning can help autonomous vehicles identify objects, navigate roads, and make decisions in real-time [Correll *et al.* 2022]. New machine learning models are continuously emerging and finding application in the domain of mobile robots and autonomous driving [Grigorescu *et al.* 2020]. For example, researchers have used deep learning to train autonomous vehicles to recognize and avoid obstacles, pedestrians, and other vehicles on the road [Bojarski *et al.* 2016]. The machine learning models for autonomous driving are typically trained in a public cloud or data centre. The trained models for autonomous driving deploy to and execute on computing platforms onboard the autonomous vehicles.

The capabilities of the onboard computing platform in the autonomous vehicle are a crucial factor in determining the successful performance of the machine learning models and, in turn, the vehicle's successful autonomous operation. Traditional CPUs are not sufficiently powerful to process machine learning workloads to meet the necessary performance criteria for safe and effective autonomous driving. While useful for many AI-based workloads, GPUs are just one tool in an increasingly rich ecosystem of coprocessors known as machine learning accelerators. A challenge is designing compute systems that are high performance enough to meet the demands of running increasingly sophisticated machine learning models. The challenges for L4+ systems are orders of magnitude more complex and more demanding than for L2 and L3 systems. The safety demands are more complex, with higher stakes. There are redundancy scenarios that must be factored in. Speed, performance, and latency demands are all much higher. The machine learning models are more sophisticated in L4+. There are more sensors, resulting in higher data volumes and data processing demands. 3D object detection, object tracking, and trajectory planning are autonomous driving functions that require increased compute power. ML accelerators need to deal with these challenges and offer design options that traditional compute architectures do not.

Hardware accelerators are not a new concept. Floating-point coprocessors, graphics processing units (GPUs), and video codecs are familiar and widely used examples [Patel and Hwu 2008]. What is new is the development and growth of special purpose hardware accelerators, known collectively as ML accelerators, dedicated to processing machine learning workloads [Park and Kim 2021]. [Reuther *et al.* 2022] summarize recent trends in commercial ML accelerators. Figure 1 highlights that ML accelerators exist on a continuum and can be compared along several dimensions in terms of flexibility, application, throughput, and cost. For example, at one end CPUs have general flexibility to a wide number of applications. While they can execute ML workloads, they have no specific support for them. At the other end of the continuum, neural processing units (NPU) are designed to provide increased performance for specialized ML workloads, but at the cost of being less flexible for general-purpose applications. ASICs have high parallel processing power and relatively lower cost but lack flexibility, only suitable for specific use cases. System designers therefore need to balance consider a blend of processing resources tailored to their application needs. Most of the innovations are focusing on optimising the compositions from the compute architecture spectrum for their specific applications. These accelerators are usually using silicon CMOS logic technology where the hardware manufacturing and software infrastructures are most mature and advanced to obtain the highest performance with the least costs.

There is no definitive set of standards that defines whether a processor or other hardware element is automotive grade. However, there are certain standards and operating conditions that must be met by processors, including ML accelerators, used in autonomous driving (as distinct from, say, chips used in a vehicle's entertainment system). They typically require an automotive safety integrity level (ASIL) rating in compliance with the ISO 26262 functional safety standard [ISO 2018]. Other common characteristics of automotive-qualified

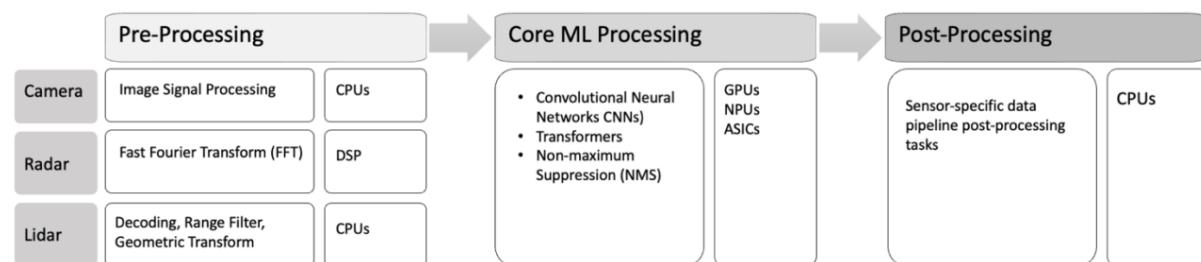
processors include operating temperature range from  $-40^{\circ}\text{C}$  to  $125^{\circ}\text{C}$  or  $150^{\circ}\text{C}$  [Automotive Electronics Council 2019], and supporting automotive-specific communications protocols and signaling interfaces [Staron 2021]. ML silicon should also be qualified according to AEC-Q100 standard for ICs and AEC-Q104 for multi-package IC reliability standards [Automotive Electronics Council 1994]. Other system considerations include protection against electro-static discharge and support for failure modes effects and diagnostic analysis (FMEDA).



**Figure 1: Continuum of processor architectures from general purpose CPUs to special-purpose ML accelerators**

### 3 Matching Machine Vision Workloads to Processor Architectures

These engines are used mostly for core ML processing in autonomous driving machine vision tasks. Figure 2 shows the common stages of processing for the three primary sensor modalities used in AVs. Figure 2 shows some common examples of tasks processed at each stage. While state-of-the-art machine vision algorithms make use of ML, when it comes to the full machine vision pipeline a lot of work is also performed in CPUs or DSP. Hardware-based ML accelerators are most useful in performing core ML processing on camera, radar, and lidar data.



**Figure 2: Examples of matching machine vision tasks to processor types for different sensor modalities**

Support for functional safety and cybersecurity are table stakes. Design trade-offs to consider include performance, power, and cost. Factors influencing system performance include on-chip memory (e.g., HBM, SRAM), off-chip memory (e.g., DDR), and core clock rate. Moving data around is expensive and one of the main contributors to system latency. We need to consider relative strengths of interconnect options for connecting multiple chips in a system. Options include PCIe, MIPI, Ethernet, and GMSL. Special bus structures, non-uniform memory access, and custom cache coherence mechanisms can improve performance by treating memory from multiple sources as a single addressable block, rather than moving data around between multiple blocks. Software and toolkit support is a differentiator, with associated development effort, maintenance cost, and impact on overall system architecture.

### 4 Conclusions

Machine vision applications in autonomous vehicles use data from a variety of sensors including cameras, radar, and lidar to enable the vehicle to perceive its environment. A variety of processor architectures are available for processing data in these applications. Some tasks are better suited to more general-purpose architectures such as CPUs and GPUs. Others can benefit from specialized processing architectures such as neural processing units,

DSPs, or ASICs. Choosing the right architecture for a given task helps meet safety and performance demands of machine vision applications in autonomous driving. The authors are researching ML accelerator architectures for autonomous driving systems that support L4+ autonomy. They are designing next-generation autonomy compute systems that leverage the power of ML accelerators. Ongoing and future research will provide deeper analysis of specific machine vision use cases using data from a variety of sensors including camera, lidar, and radar.

## References

- [Automotive Electronics Council 1994] Automotive Electronics Council (1994) *Automotive Electronics Council*, available: <http://www.aecouncil.com/> [accessed May 1, 2023].
- [Automotive Electronics Council 2019] Automotive Electronics Council (2019) 'AEC - Q103-002 Rev - (Initial Release): Failure Mechanism Based Stress Test Qualification for Micro Electro-Mechanical System (MEMS) Pressure Sensor Devices', available: [http://www.aecouncil.com/Documents/AEC\\_Q103-002\\_Rev-.pdf](http://www.aecouncil.com/Documents/AEC_Q103-002_Rev-.pdf).
- [Bachute, M.R. and Subhedar, J.M. 2021] Bachute, M.R. and Subhedar, J.M., (2021) 'Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms', *Machine Learning with Applications*, 6, 100164, available: <http://dx.doi.org/https://doi.org/10.1016/j.mlwa.2021.100164>.
- [Ben-Ari, M. and Mondada, F. 2018] Ben-Ari, M. and Mondada, F., (2018) *Elements of robotics*, Springer International Publishing.
- [Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U. and Zhang, J. 2016] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U. and Zhang, J., (2016) 'End to end learning for self-driving cars', *arXiv preprint arXiv:1604.07316*.
- [Correll, N., Hayes, B., Heckman, C. and Roncone, A. 2022] Correll, N., Hayes, B., Heckman, C. and Roncone, A., (2022) *Introduction to Autonomous Robots: Mechanisms, Sensors, Actuators, and Algorithms*, Cambridge, MA: The MIT Press.
- [Gharib, M., Lollini, P., Botta, M., Amparore, E., Donatelli, S. and Bondavalli, A. Year] Gharib, M., Lollini, P., Botta, M., Amparore, E., Donatelli, S. and Bondavalli, A. (2018) 'On the safety of automotive systems incorporating machine learning based components: a position paper', in *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, IEEE, 271-274.
- [Grigorescu, S., Trasnea, B., Cocias, T. and Macesanu, G. 2020] Grigorescu, S., Trasnea, B., Cocias, T. and Macesanu, G., (2020) 'A survey of deep learning techniques for autonomous driving', *Journal of Field Robotics*, 37(3), 362-386.
- [ISO 2018] ISO (2018) *ISO 26262-11: Road vehicles - Functional safety - Part 11: Guidelines on application of ISO 26262 to semiconductors*, Geneva, Switzerland: ISO.
- [Jo, K., Kim, J., Kim, D., Jang, C. and Sunwoo, M. 2014] Jo, K., Kim, J., Kim, D., Jang, C. and Sunwoo, M., (2014) 'Development of Autonomous Car—Part I: Distributed System Architecture and Development Process', *IEEE Transactions on Industrial Electronics*, 61(12), 7131-7140, available: <http://dx.doi.org/10.1109/TIE.2014.2321342>.
- [Li, D., Huang, Y. and Qian, L. 2022] Li, D., Huang, Y. and Qian, L., (2022) 'Potential adoption of robotaxi service: The roles of perceived benefits to multiple stakeholders and environmental awareness', *Transport policy*, 126, 120-135, available: <http://dx.doi.org/10.1016/j.tranpol.2022.07.004>.
- [Park, H. and Kim, S. 2021] Park, H. and Kim, S. (2021) 'Chapter Three - Hardware accelerator systems for artificial intelligence and machine learning' in Kim, S. and Deka, G. C., eds., *Advances in Computers* Elsevier, 51-95.
- [Patel, S. and Hwu, W.m.W. 2008] Patel, S. and Hwu, W.m.W., (2008) 'Accelerator Architectures', *IEEE Micro*, 28(4), 4-12, available: <http://dx.doi.org/10.1109/MM.2008.50>.
- [Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S. and Kepner, J. Year] Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S. and Kepner, J. (2022) 'AI and ML accelerator survey and trends', in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, 1-10.
- [SAE International 2021] SAE International (2021) *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016\_202104*.
- [Staron, M. 2021] Staron, M., (2021) *Automotive Software Architectures*, 2nd Ed, Springer International Publishing.

# Self-Supervised Online Camera Calibration for Autonomous Driving Applications

Ciarán Hogan<sup>1,2</sup>, Ganesh Sistu<sup>1</sup>, Ciarán Eising<sup>2</sup>

<sup>1</sup>Valeo Vision Systems Galway, Ireland

<sup>2</sup>Department of Electronic and Computer Engineering University of Limerick Ireland

## Abstract

Camera calibration is a critical step for accurate perception in autonomous vehicles. However, traditional calibration methods are often laborious and time-consuming, and require specialised data collection and careful tuning. This process must be repeated whenever the parameters of the camera change, which can be a frequent occurrence due to vibration, impacts or extreme temperatures. This paper proposes a novel deep learning framework for self-supervised camera calibration. The framework learns the intrinsic and extrinsic camera parameters from unlabeled driving data. This allows the framework to calibrate the camera in real time, without the need for any physical targets or special planar surfaces. This could cut the cost of re-calibration for both the manufacturer and consumer as the vehicle can re-calibrate itself over the lifetime of the vehicle, which also improves the accuracy and safety of the ADAS systems on the car.

**Keywords:** Camera Calibration, Self-Supervised Depth & Pose Estimation, Machine Vision

## 1 Introduction

There are two stages of camera calibration: intrinsic and extrinsic. Intrinsic calibration refers to the calibration of the camera's internal parameters, such as the focal length and principal point. Extrinsic calibration refers to the calibration of the camera's external parameters, such as its location and orientation in space. Traditionally, camera calibration is performed once in a laboratory setting using special calibration targets. This can be a time-consuming and expensive process, and the calibration results may not be accurate over time due to factors such as temperature changes, vibration, and wear and tear. In recent years, deep learning has opened new possibilities for camera calibration [Bogdan et al., 2018]. The advantage of Deep learning methods is that they can continually learn calibration over the lifetime of the vehicle, without the need for special calibration targets. This makes deep learning-based calibration methods more efficient as they stay up to date with the constantly changing camera parameters due to external factors

Hence, we propose a novel deep learning-based camera calibration method for autonomous vehicles. Our method is able to learn the intrinsic and extrinsic parameters of the camera from data collected while the vehicle is in motion. This allows for provide accurate calibration over the lifetime of the vehicle, without the need for manual intervention.

Traditional camera calibration methods require specialized lab settings and target patterns. Zhang's method [Zhang, 2000], is a popular traditional CV method that only requires a checkerboard pattern and can be done in any environment. The method works by detecting feature points in the checkerboard pattern under different orientations, which can be achieved by moving either the camera or the plane.

The proposed approach uses properties of Self supervised depth and pose estimation networks like [Guizilini et al., 2020] and [Godard et al., 2019] as proxy to also learn calibration. [Garg et al., 2016] first introduced the idea of the joint learning of depth and ego-motion. The proposed method provided a significant contribution to the subject of depth estimation as it allows for the learning of depth estimation from monocular images totally unsupervised.

## 2 Methodology

The proposed calibration framework uses self-supervised monocular depth and pose estimation as a proxy for learning camera calibration with the addition of a third network to learn intrinsics. Self-supervised depth and ego-motion architectures consist of a depth network that produces a depth map and a pose network that predicts the transformation between the current frame and the context frame. With this known transformation and depth map, one can warp/project the current frame into a target image and train the networks jointly by minimising the photometric loss between the actual image and the synthesised image from the projection. Monodepth2 [Godard et al., 2019] was chosen as a base framework for the project. Monodepth2 is a popular and well documented Pytorch-based self-supervised monocular depth estimation network. At the core of the project is a depth network with a multi-input ResNet [He et al., 2016] encoder and UNet [Ronneberger et al., 2015] decoder alongside a pose CNN, also with a multi-input ResNet encoder.

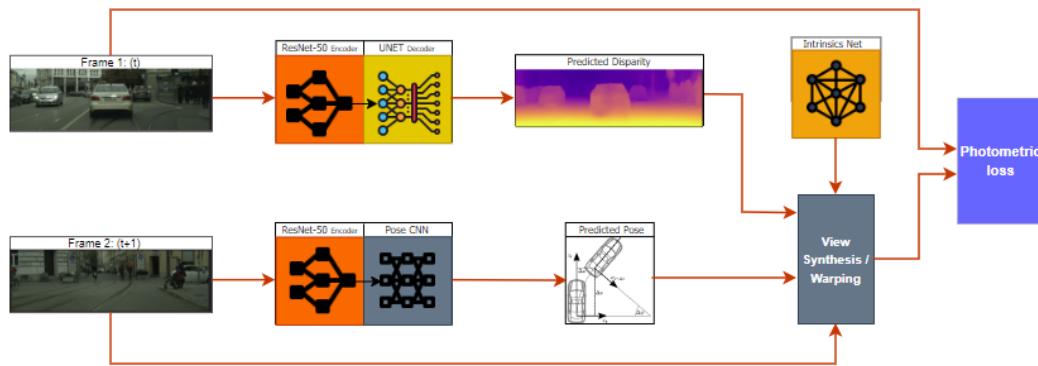
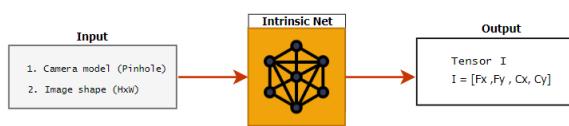


Figure 1: Calibration Framework Architecture

### 2.1 Architecture

The modified monodepth2 framework consists of:

- Depth network: ResNet-50 encoder pre-weighted on ImageNet & UNet decoder.
- Pose network: Multi-input ResNet-50 encoder pre-weighted on ImageNet & Pose CNN.
- Intrinsic network consists of 4 or more trainable parameters (Depending on the camera model) representing intrinsic camera parameters which feed into the view synthesis function for image warping.



(a) Intrinsic Network pseudo diagram

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}$$

(b) Traditional Intrinsic matrix

### 2.1.1 Intrinsic Network

The intrinsic network consists of 4 or more trainable parameters (Depending on camera model) which represent the intrinsic parameters of the relevant camera model. The network takes two inputs, the camera model and the image shape. The network outputs a tensor of size 1x4 (depending on camera model). This tensor is then manipulated into a traditional 3x3 intrinsic matrix which is fed into the image warping function and the network is trained in simultaneously with the depth and pose network by minimising the photometric loss.

### 2.1.2 Data

The framework was trained on the KITTI dataset. The KITTI dataset is a benchmark dataset commonly used in computer vision for depth estimation. The Eigen zhou split was used which is a subset of the KITTI dataset which consists of 39K images for training and 4k images for validation. A further 679 unseen images are used for evaluation.

## 3 Experiments and Results

The framework is trained in a self-supervised manner by minimising photometric reconstruction error. The framework was trained for 20 Epochs with a learning rate of 0.0001 and a batch size of 4. Training took 36 hours on an NVIDIA GeForce RTX 2080ti graphics card.

Model	Abs Rel $\downarrow$	Sq Rel $\downarrow$	RMSE $\downarrow$	RMSE log $\downarrow$
Base monodepth2	0.114	0.931	4.810	0.192
monodepth2 x IntrinsicNet	0.112	0.851	4.706	0.188

Table 1: Depth metric results vs baseline

Comparing the depth metrics of the baseline vs learned intrinsics is one way to evaluate the learned intrinsics in the absence of synthetic data as we are using depth estimation as a proxy to learn camera calibration. Comparing the two we see that the modified framework has a slight improvement in depth evaluation metrics, possibly due to the learned intrinsics being more accurate than the static KITTI parameters.

Model	Intrinsics	Abs Rel $\downarrow$	Sq Rel $\downarrow$	RMSE $\downarrow$
Gordon et al	K (P)	0.129	0.982	5.23
	L (P)	0.128	0.959	5.23
J. Fang et al	L (P)	0.129	0.8393	4.96
Monodepth2	K (P)	0.114	0.931	4.810
	L (P)	0.112	0.851	4.70

Table 2: Depth metric results vs baseline

Table 4 shows a comparison of depth metrics from other camera based learning methods [Fang et al., 2022, Gordon et al., 2019] trained on the Eigen split of the KITTI dataset. K denotes known intrinsics and L denotes learned intrinsics. Note the proposed implementation is trained on the Eigen Zhou split which contains the same files but has extra data added from more challenging scenes and lower lighting. This means the Eigen Zhou split is more challenging and provides a better estimation of the model's generalisation capability. The table provides a quick comparison of improved metrics between baseline and learned intrinsics of other implementations.

## 4 Conclusion & Future work

This paper proposes a novel self-supervised deep learning framework that learns camera calibration from unlabelled video input. As seen the framework learns calibration, which improved depth evaluation metrics by providing more accurate calibration parameters than the static calibration parameters provided in KITTI. As with any real world dataset the calibration provided does not represent the true parameters of the camera, only a very good estimate of the parameters at time of calibration.

Future work would be to train the framework on synthetic data. The advantage of using synthetic data is that we know the exact true calibration parameters of the camera so we can accurately evaluate the networks outputs. Other future work would be to adapt the framework to work with wider field of view camera models like the fisheye camera model.

In summary the proposed framework could potentially be very effective as a recalibration tool for vehicle perception systems and could cut the cost of recalibration for both manufacturer and consumer.

## References

- [Bogdan et al., 2018] Bogdan, O., Eckstein, V., Rameau, F., and Bazin, J.-C. (2018). Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10.
- [Fang et al., 2022] Fang, J., Vasiljevic, I., Guizilini, V., Ambrus, R., Shakhnarovich, G., Gaidon, A., and Walther, M. R. (2022). Self-supervised camera self-calibration from video.
- [Garg et al., 2016] Garg, R., Bg, V. K., Carneiro, G., and Reid, I. (2016). Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer.
- [Godard et al., 2019] Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838.
- [Gordon et al., 2019] Gordon, A., Li, H., Jonschkowski, R., and Angelova, A. (2019). Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras.
- [Guizilini et al., 2020] Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., and Gaidon, A. (2020). 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.
- [Zhang, 2000] Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334.

# An Ensemble Deep Learning Approach for COVID-19 Severity Prediction Using Chest CT Scans

Sidra Aleem<sup>\*1</sup>, Mayug Manipambil<sup>†1</sup>, Suzanne Little<sup>2</sup>, Noel O'Connor<sup>2</sup> and Kevin McGuinness<sup>2</sup>

<sup>1,2</sup>*ML Labs, SFI Centre for Research Training, Dublin City University*

*sidra.aleem2@mail.dcu.ie*

*mayugmanipambil@gmail.com*

*{suzanne.little, noel.oconnor, kevin.mcguinness}@dcu.ie*

## Abstract

Chest X-rays have been widely used for COVID-19 screening; however, 3D computed tomography (CT) is a more effective modality. We present our findings on COVID-19 severity prediction from chest CT scans using the STOIC dataset. We developed an ensemble deep learning based model that incorporates multiple neural networks to improve COVID-19 severity prediction. To address data imbalance, we used slicing functions and data augmentation. We further improved performance using test time data augmentation. Our approach, which employs a simple yet effective ensemble of deep learning-based models with strong test time augmentations, achieved results comparable to more complex methods and secured the fourth position in the STOIC2021 COVID-19 AI Challenge. Code is available online at: STOIC AI Challenge.

**Keywords:** COVID-19 severity, Automated medical diagnosis, Radiology, CT scans.

## 1 Introduction

COVID-19 has created a global health crisis with millions of cases and deaths reported worldwide. Timely and accurate COVID severity prediction is essential for effective clinical management and treatment. Deep learning has shown tremendous potential in the medical domain. Automating the severity prediction of COVID-19 based on deep learning can lead to improved clinical workflow, resulting in faster diagnosis and better prognosis for severe COVID-19 cases. While a large number of studies have utilized deep learning methods for COVID-19 prediction based on chest X-ray data, CT scans have been found to be more effective in detecting COVID-19 positivity and severity [Aswathy et al., 2021]. Previous studies have primarily focused on COVID-19 positivity prediction or improving feature extraction methods. However, limited work has been done on COVID-19 severity prediction using CT scans.

## 2 Material and Method

We employed an ensemble deep learning approach to predict COVID-19 severity in high-resolution 3D CT scans provided by the STOIC2021 COVID-19 AI Challenge [Boulogne, 2022]. The scans had a spatial resolution of  $512 \times 512$  and a depth of 128 to 600 slices. To standardize the number of slices, we used a uniform sampling function, which samples 32 uniformly spaced slices from the slice dimension. The radiodensity is measured using Hounsfield Units (HU), ranging from -1024 HU to 3071 HU, and stored as 12-bit numbers [Glide-Hurst et al., 2013]. However, directly scaling these values to  $[0, 1]$  results in low contrast images that make identification of COVID-19-related features difficult [J et al., 1995]. To improve contrast, we utilize

<sup>\*</sup>Contributed equally

<sup>†</sup>Contributed equally

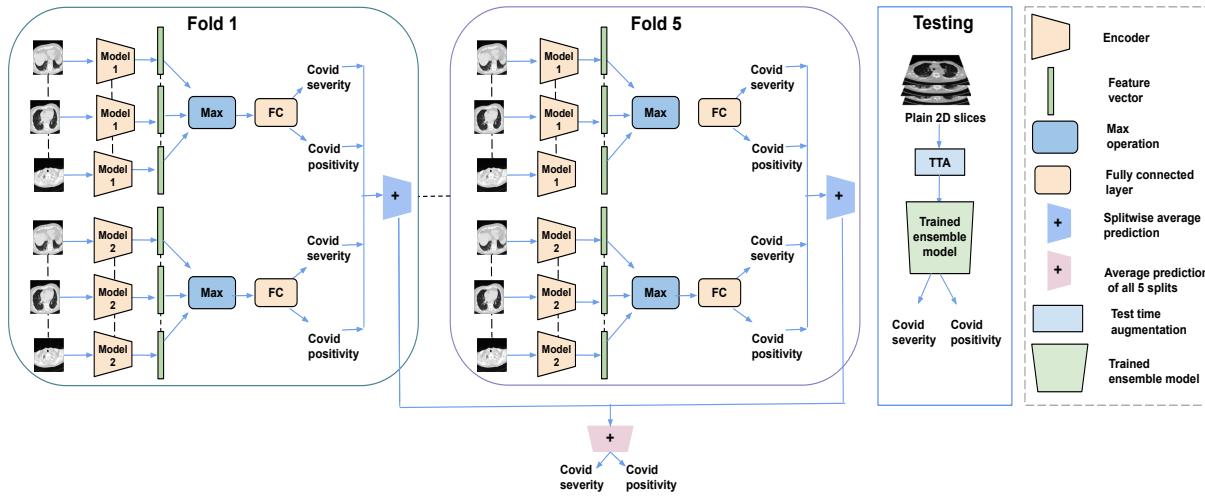


Figure 1: The schematic overview of our method.

the lung window with a window width of 1500 HU and a window level of -600 HU [J et al., 1995]. To deal with data scarcity, we partitioned the training set into five random splits. ResNet18 and MobileNetV3 are used for prediction. The overview of our method is shown in Figure 1. For every split, each pre-processed slice  $S_i$  is passed through both the ResNet18 and MobileNetV3 encoder  $E$  to obtain feature vectors  $z_i = f(S_i)$ . The maximally activated features are selected along the slice dimension on the feature vectors of the slices as  $z_j^{\max} = \max\{z_i\}_{i=1}^{132}$  where  $z_i \in R^D$ . Only  $z_j^{\max}$  feature vector is then used by the fully connected layer to predict COVID severity  $\hat{x}_s$  and COVID positivity  $\hat{x}_p$ . The two models are combined through an ensemble technique, where the probabilities from both models are averaged to obtain the final prediction. Finally, the probabilities obtained from all five splits are averaged together to get final prediction  $\hat{y}_s$  and  $\hat{y}_p$ .

A number of challenges were encountered while experimenting with STOIC dataset [Boulogne, 2022]. a) The publicly available subset of the STOIC dataset is composed of only 2000 subjects, training the model with such a limited data is highly challenging. To avoid overfitting, various image level augmentations: random rotations, random crops, random gamma adjustment, color jitter, median blur, multiplicative noise, and *mixup* were used. These augmentations led to improvements in our results (Table. 2). Furthermore, performance was improved by test time data augmentation (TTA) with center crop, crops around four corners, safe rotate. b) The publicly available sample of STOIC dataset is heavily imbalanced, with only 301 subjects identified as severe COVID cases. To ensure training stability, a balanced sampler is used to assign weights to samples in proportion to the inverse occurrence of their respective classes in the training set. It helped to prevent the model's bias to the majority class. c) In addition to implementation challenges, the submission limit was restricted to one per week.

### 3 Results and Discussion

All the models were trained for 100 epochs using the Adam optimizer with a weight decay of 0.0005. The binary cross-entropy with logits criterion was used for evaluation. The training hyperparameters included a batch size of 16, a learning rate of  $5 \times 10^{-4}$  with a decay of 0.5 every 40 epochs using StepLR. The STOIC dataset [Boulogne, 2022] comprises of CT scans from 10,735 subjects. For this challenge, the dataset has been divided randomly into a publicly available training set (2,000 subjects) released under the CC BY-NC 4.0 license, a test set ( $\sim 1,000$  subjects) and a private training set (7,000+ subjects). The scans are in .MHA format and have a resolution of  $512 \times 512$  with varying number of slices ranging from 128 to 600. The meta-data contains information about the subject's age and gender, with age ranging from 35 to 85 years and a gender distribution of 57.4% male and 42.6% female. The ground truth consists of two labels: COVID-19 positivity and COVID-19 severity. The primary evaluation criterion is COVID-19 severity Area Under the Curve (AUC). We experimented

with multiple algorithms on the public set as shown in Table 1, and chose the one that performed best on the public test set. We then used this algorithm on the qualification leaderboard (LB) test set. Based on the results obtained from the qualification LB, we were able to evaluate the generalization performance of the model and submitted the improved version to final LB. ConvNext Tiny [Liu et al., 2022] performed well on the public data set; however, when tested on the qualification LB, AUC severity dropped significantly to 0.748, indicating poor generalization performance. Alternative models were also investigated: 3D CNN, MobileNetV3 small, and the use of encoders as fixed feature extractors along with age and sex metadata. These did not improve the predictive performance. Building on our initial results from Table 1, we conducted further experiments with ResNet18 and MobileNetV3. To evaluate the impact of augmentations on the overall prediction, we experimented with four set

Table 1: Performance evaluation based on AUC Severity and AUC COVID on public STOIC subset [Boulogne, 2022].

Model	STOIC Public Data		Qualification LB	
	AUC Severity	AUC COVID	AUC Severity	AUC COVID
CNN	0.687	0.780	-	-
ConvNext	0.845	0.748	0.748	0.800
ResNet18	0.775	0.784	<b>0.752</b>	0.784
MobileNet	0.817	0.780	<b>0.779</b>	0.735

of augmentations. 1) Default: consisting of horizontal flip, random crop to  $224 \times 224$ , random gamma and color jitter with brightness 0.5, contrast 0.5 and saturation 0.4. 2) Default + Strong: consisting of safe rotate with limit 30, median blur, and the default augmentations. 3) Default + Strong + Mixup [Zhang et al., 2017]: consisting of set 1, set 2 and mixup augmentation. The mixup function is applied to the batch of images and labels i.e. COVID severity and COVID positivity with  $\alpha = 0.8$ . It returns augmented images and mixed-up labels for both the classes respectively. 4) Default + Strong + Mixup + TTA: consisting of TTA, center crop, crops around four corners, safe rotate along -5, 10, +5. As shown in Table 2, the integration of augmentation further improved the performance. TTA proved to be effective and greatly complemented the other augmentation techniques, further enhancing the overall performance of the model. To select the best model for final LB submission, for each split,

Table 2: Impact of augmentation on AUC Severity.

Augmentation	Public data		Qualification LB	
	ResNet18	MobileNetV3	ResNet18	MobileNetV3
Default	0.775	0.817	0.752	0.779
Default + Strong	0.795	0.831	0.781	0.793
Default + Strong + Mixup	0.842	0.829	0.790	0.795
Default + Strong + Mixup + TTA	<b>0.863</b>	<b>0.841</b>	<b>0.815</b>	<b>0.821</b>

we created an ensemble of ResNet18 and MobileNetV3 with most effective augmentation as shown in Table. 2 and finally ensembled predictions from all five splits as described in Section 2.

#### 4 Comparison with top methods on final LB and conclusion [Boulogne, 2022]

The first ranked team, pre-trained ConvNext [Liu et al., 2022] with MosMed [Morozov et al., 2020] for severity classification and UperNet [Xiao et al., 2018] with TCIA [Aerts et al., 2015] for lesion segmentation. It was followed by training on STOIC dataset using metadata and the output of both backbones as vectors. They used a 5-fold cross-validation and ensemble model for testing. Team 2, used two vision encoders pre-trained on iBot [Zhou et al., 2021] via self-supervised learning on plain slices and segmented lung regions. They concatenated the features with age and sex features and used logistic regression for predictions. Team 3, used a lung segmentation model and autodidactic pre-training on segmented images. The network's output was combined with age and passed to a fully connected layer and finally ensemble of five models and TTA was

Table 3: Final leader board results: comparison with top ranked methods [Boulogne, 2022].

<b>Model</b>	<b>STOIC Public Data</b>	
	<b>AUC Severity</b>	<b>AUC COVID</b>
First	0.815	0.616
Second	0.811	0.845
Third	0.794	0.837
<b>Fourth (our method)</b>	0.787	0.829

used. In contrast to other methods, our approach did not use highly complex models with additional data sets. Despite its simplicity, our method is highly effective and competitive with the more complex techniques, as can be seen from Table 3. Unlike other methods, we did not use metadata for the final LB submission as it did not improve performance on the public data set. Despite observing the same phenomenon, other teams opted to include metadata in their final submission and it proved to be effective. If our approach had included metadata, it might have been ranked among the top three teams in the competition.

## Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183.

## References

- [Aerts et al., 2015] Aerts, H., Rios Velazquez, E., Leijenaar, R., Parmar, C., Grossmann, P., Carvalho, S., and Lambin, P. (2015). Data from nsclc-radiomics. the cancer imaging archive.
- [Aswathy et al., 2021] Aswathy, A., Hareendran, A., and SS, V. C. (2021). Covid-19 diagnosis and severity detection from ct-images using transfer learning and back propagation neural network. *Journal of Infection and Public Health*, 14(10):1435–1445.
- [Boulogne, 2022] Boulogne, L. (2022). STOIC2021- COVID-19 AI Challenge. <https://stoic2021.grand-challenge.org/stoic2021/>. [Online; accessed 22-Feb-2022].
- [Glide-Hurst et al., 2013] Glide-Hurst, C., Chen, D., Zhong, H., and Chetty, I. (2013). Changes realized from extended bit-depth and metal artifact reduction in ct. *Medical physics*, 40(6Part1):061711.
- [J et al., 1995] J, S. E., S, F. M., and Godwin, J. D. (1995). Chest computed tomography display preferences. survey of thoracic radiologists. *Invest Radiol*, 99:106906.
- [Liu et al., 2022] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.
- [Morozov et al., 2020] Morozov, S. P., Andreychenko, A., Pavlov, N., Vladzimyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I. A., Gelezhe, P., Gonchar, A., and Chernina, V. Y. (2020). Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.
- [Xiao et al., 2018] Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434.
- [Zhang et al., 2017] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.
- [Zhou et al., 2021] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.

# Explaining decisions of a light weight deep network model for Coronary Artery Disease classification in Magnetic Resonance Imaging

Aaleen Khalid<sup>1</sup>, Talha Iqbal<sup>2</sup>, and Ihsan Ullah<sup>1,2</sup>

<sup>1</sup>*School of Computer Science, University of Galway, Galway, H91 TK33, Ireland.*

<sup>2</sup>*Insight SFI Research Centre for Data Analytics, University of Galway, Galway, H91 TK33, Ireland.*

## Abstract

In certain healthcare settings such as emergency or critical care units, where quick and accurate real-time analysis and decision-making are required, the healthcare system can leverage the power of artificial intelligence (AI) models to support decision-making and prevent complications. This paper investigates the optimization of healthcare AI models based on time complexity, hyper-parameter tuning, and XAI for a classification task. The paper highlights the significance of a light weight convolutional neural network (CNN) alone or in combination with other classifiers, e.g. CNN-RandomForest (CNN-RF) for analysing and classifying Magnetic Resonance Imaging (MRI). The role of hyper-parameter is also examined in finding optimal configurations that enhance the model's performance while efficiently utilizing the limited computational resources. Finally, the benefits of incorporating the XAI technique (GradCAM) in providing transparency and interpretable explanations of AI model predictions, fostering trust, and error/bias detection are explored. Using the proposed model, clinicians/cardologists can achieve accurate and reliable results while ensuring patient's safety and answering questions imposed by GDPR. The proposed investigative study will advance the understanding and acceptance of AI systems in healthcare settings.

**Keywords:** Healthcare Models, Time Complexity, Hyper-parameter Tuning, Explainable AI, Classification.

## 1 Introduction

According to the World Health Organization <sup>1</sup>, in 2019 an estimated 17.9 million people died from cardiovascular diseases, representing 32% of all global deaths. Statistics published by the American Heart Association in 2023 state that from 2017-2020, an estimated 20.5 million Americans had coronary heart disease (CHD) [Tsao et al., 2023]. Specifically, Coronary artery disease (CAD) accounts for approximately 610,000 deaths annually in the United States and is the third leading cause of death worldwide, with 17.8 million deaths annually [Brown et al., 2020].

The patient's symptoms of CAD are neither sensitive nor specific thus making it difficult for clinicians or cardiologists to rely only on them. The reference standard for CAD detection is through coronary angiography which is an invasive diagnostic imaging procedure done using cardiac catheterisation. This method is expensive and carries potential risks. Other methods include cardiac imaging techniques, which are safe, non-invasive, cheaper and can help doctors in early detection and provide timely interventions to treat CAD patients. These techniques include X-rays, Computer Tomography (CT), Echo-cardiogram and Magnetic Resonance Imaging (MRI) or Cardiac Magnetic Resonance (CMR) Imaging.

Among these techniques, CMR imaging uses magnetic waves and is considered a viable alternative for the non-invasive assessment of CAD. MRI/CMR images provide precise measurements of heart structure and functions, as well as myocardial perfusion and parametric quantification. Manual interpretation of these scans is time-consuming and needs expertise. Thus, artificial intelligence methods are exploited to automate the CAD diagnosis to reduce the analysis time with potentially improved accuracy.

<sup>1</sup><https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29>

## 1.1 Related Work

Recently, Khozeimeh *et al.* [Khozeimeh et al., 2022] adopted a hybrid approach (CNN-RF model) combining random forest (RF) and convolutional neural networks (CNN) to develop a novel classification method. They trained their model with three different optimizers (Adagrad, RMSprop and Adam) and reported the results.

The authors used 2D CMR images. Their pre-processing steps included resizing the images to 100x100 pixels and normalization between 0 and 1. The next step was feature generation using 10 CNNs, where each CNN was two-layered. The CNNs convert the input images into vectors of real value. For classification, Random Forest is fed with the feature vector generated by the CNN model. This approach preserves the inter-pixel spatial relationship and avoids the issue of dimensionality associated with pixel-by-pixel feeding.

A 5-fold cross-validation approach was implemented for training 10 CNN models. The trained CNNs were then used to generate a decision tree in the random forest, and based on the decision trees, the random forest determines whether the test image has CAD or not. For the final prediction, majority voting between decision trees was implemented. The authors also provide an open-source dataset of CAD Cardiac MRI repository<sup>2</sup> and reported to achieve a classification accuracy of 99.18% using Adam optimizer. Our objective in this work is to show that a single properly designed lightweight deep neural network can achieve better performance than an ensemble approach of multiple CNN models.

## 2 Proposed Work

In the proposed work, we implemented a lightweight deep convolutional neural network similar to the LeNET-5 model trained on the above-mentioned dataset for a comprehensive comparison. The motivation behind the proposed work is to propose a lightweight deep network model to reduce training and inference time and complexity while maintaining accuracy, enabling real-time or near-real-time CAD diagnosis. We optimized the deep model by exploring different hyper-parameter settings to identify configurations that maximize the model's performance. Furthermore, to understand why certain decisions are made, we used GradCam [Selvaraju et al., ] an XAI technique to provide interpretable insights into the decision-making process of the model, generating heat maps that highlight the regions of MRI / CMR images that contribute to the classification of CAD in respective categories.

**Dataset description** The dataset consists of 63,151 multiparametric CMR images that include 37,290 healthy images and 25,861 CAD patients. CAD diagnosis was confirmed by invasive coronary angiography. During the pre-processing stage, a manual inspection was conducted on images from both subsets, and any images with poor MRI/CMR quality were excluded from further analysis. Following the pre-processing stage, the dataset consisted of 34,216 images from healthy patients and 17,438 images from patients with CAD.

**Performance Assessment Matrices** The performance of the classifier is assessed using Positive Predictive Value (PPV), Recall (Sensitivity or True Positive Rate), Specificity (True Negative Rate), F1-Score, Area Under the Curve (AUC), Accuracy and Balanced Accuracy.

## 3 Results

Figure 1 illustrates the proposed model inspired from LeNET-5 architecture<sup>3</sup>. All the experiments were implemented in Python using Keras library<sup>4</sup>. The models were trained on Apple M2 Pro with 16 GB RAM. The following subsections discuss the time complexity calculations, the effect of hyper-parameter tuning, and feature explanations using XAI results.

### 3.1 Time Complexity Calculation and Comparison

<sup>2</sup><https://www.kaggle.com/danialsharifrazi/cad-cardiac-mri-dataset>

<sup>3</sup>[https://d2l.ai/chapter\\_convolutional-neural-networks/lenet.html](https://d2l.ai/chapter_convolutional-neural-networks/lenet.html)

<sup>4</sup><https://keras.io/>

The time taken by a model is determined by the number of layers and the operations performed in each layer. The proposed model comprises of seven layers as shown in figure 1, excluding the input layer. The input image shape was (100,100,1), in model 1 the number of filters in each convolutional layer i.e. C1 and C3 is 6, whereas in model 2 it is 12 in C1 and 6 in C3, the kernel size in convolutional layers is 5x5 (with stride (2,2))

and 2x2 in pooling layers, neurons in fully connected layer 1 and 2 are 128 and 84, respectively. Whereas, the output layer has only 1 unit, as the model is performing binary classification.

Considering the previously mentioned size, our model is much smaller compared to the CNN-RF ensemble model in [Khozeimeh et al., 2022]. Our inference time on a MacBook laptop for 323 test images of size 100x100 is only 2.6 sec, which is only 8 milliseconds per image. Additionally, it provides better or equal classification accuracy. Our model's lower computational complexity enables faster image analysis and diagnosis, improving efficiency, and facilitating deployment on resource-constrained systems.

### 3.2 Hyper-parameter tuning and classification results

We explored various hyperparameter configurations to achieve optimal model performance for CAD image classification. The PReLU activation function combined with the RMSprop optimizer resulted in the highest classification accuracy, achieving 99.35% overall and a 99.13% balanced accuracy. This surpasses the previously achieved highest accuracy of 99.18% obtained by the reference CNN-RF model, see table 1.

Table 1: Model performances achieved with different settings: Comparison

Model	Activation Function	Optimizer	PPV	Recall	Specificity	F1-Score	AUC	Accuracy
Our Model 1	PReLU	Adam	99.19	98.03	99.59	98.60	99.88	99.06
Our Model 2	PReLU	RMSprop	99.62	98.45	99.81	99.04	99.92	99.35
CNN-RF	ReLU	Adam	100	98.88	99.66	99.70	99.00	99.18

### 3.3 Explaining the decision with GradCAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [Selvaraju et al., ] is an explainable AI technique used to generate a heatmap of the important regions in an image that significantly contributes to the deep learning model's prediction. Figure 2 illustrates some examples of generated GradCAM heatmaps highlighting the focused regions (region of interest) for the prediction of CAD in test images. In the GradCAM visualization, the intensity of the heatmap represents the importance of each pixel in the input image. Higher intensity (e.g., brighter colours) and high contrast colour with respect to the background are indicative of a more significant region that contributed to the model's prediction. Figure 2(a) shows an image with CAD, therefore, the XAI technique highlighted the region of interest with a pattern of bluish in the centre and yellowish from sides that are predicting the image with a disease. Whereas, 2(b) is an abnormal (CAD) image but misclassified. Its heatmap is yellowish as a whole unlike the images correctly classified as having patterns of blue and yellow.

## 4 Conclusion

Considering that time and hyper-parameter tuning play a vital role in the development of healthcare AI models, optimizing them to be resource efficient, suitable for real-time applications, and ensuring the reliability of the

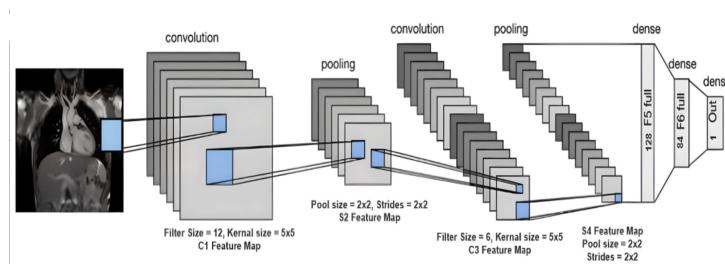


Figure 1: Proposed Model Architecture



(a) Normal case - Region of interest is shown with bluish color in the center

(b) Region of interest: The bright region around the center (misclassified)

Figure 2: Heatmaps generated by GradCAM on test images

model is important. In addition, XAI provides interpretability and fosters an understanding of the model's decision-making. XAI helps in ensuring AI recommendations are aligned with clinical experts' interpretations, making them safer and GDPR compliant.

The CNN-RF (baseline) and proposed models were presented for the diagnosis of CAD patients using open-access 2D CMR images. The models perform binary classification and distinguish normal images from abnormal/sick patients' images. The classification model performance was measured using PPV, Recall, Specificity, F1-Score, AUC and Accuracy matrices. As the dataset has a small class imbalance, an additional performance metric i.e., Balance Accuracy was calculated. The combination of different hyperparameters revealed different classification accuracy, tabulated in table 1. Among all the settings, our model (model 2) achieved the highest test accuracy of 99.35% (with balanced accuracy = 99.13%) with activation function 'PReLU', RMSprop optimizer, batch size of 32 and binary-crossentropy as loss-function.

The proposed investigative work aimed to provide insights into the optimization of healthcare AI models, ensuring accurate and reliable results with a small lightweight convolutional neural network while prioritizing patient safety, and advancing the acceptance and understanding of AI in healthcare settings. While the results on the 2D CMR images are promising, in future investigated CNN-based models will be tested on other healthcare images such as Computer Tomography (CT), X-rays and/or Echocardiogram (Echos) images to determine the model's comprehensive diagnostic capabilities, cross-domain scalability, and performance on multi-model data.

## Acknowledgments

This publication has emanated from research supported by a grant from Science Foundation Ireland under Grant number SFI/12/RC/2289\_P2. We acknowledge the School of Computer Science summer EDI scholarship for providing funding in part for the completion of this work. For correspondence: ihsan.ullah@universityofgalway.ie

## References

- [Brown et al., 2020] Brown, J. C., Gerhardt, T. E., and Kwon, E. (2020). Risk factors for coronary artery disease.
- [Khozeimeh et al., 2022] Khozeimeh, F., Sharifrazi, D., Izadi, N. H., Joloudari, J. H., Shoeibi, A., Alizadehsani, R., Tartibi, M., Hussain, S., Sani, Z. A., Khodatars, M., et al. (2022). Rf-cnn-f: random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance. *Scientific Reports*, 12(1):11178.
- [Selvaraju et al., ] Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Gradcam: Visual explanations from deep networks via gradient-based localization. arxiv 2016. *arXiv preprint arXiv:1610.02391*.
- [Tsao et al., 2023] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E., et al. (2023). Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621.

# Non-Contact Breathing Rate Detection Using Optical Flow

Robyn Maxwell, Timothy Hanley, Dara Golden, Adara Andonie, Joseph Lemley, and Ashkan Parsi  
*OCTO Sensing Team, Xperi Inc., Galway, Ireland*

## Abstract

Breathing rate is a vital health metric that is invaluable indicator of the overall health of a person. In recent years, the non-contact measurement of health signals such as breathing rate has been a huge area of development, with a wide range of applications from telemedicine to driver monitoring systems. This paper presents an investigation into a method of non-contact breathing rate detection using a motion detection algorithm, optical flow. Optical flow is used to successfully measure breathing rate by tracking the motion of specific points on the body. In this study, the success of optical flow when using different sets of points is evaluated. Testing shows that both chest and facial movement can be used to determine breathing rate, but to different degrees of success. The chest generates very accurate signals, with a RMSE of 0.63 on the tested videos. Facial points can also generate reliable signals when there is minimal head movement but are much more vulnerable to noise caused by head/body movements. These findings highlight the potential of optical flow as a non-invasive method for breathing rate detection and emphasize the importance of selecting appropriate points to optimize accuracy.

**Keywords:** Breathing Rate, Optical Flow, Open Pose, Face Detection, Driver Monitoring

## 1 Introduction

Breathing rate (BR) is a vital physiological signal that can indicate the overall health of a person by providing an insight into their respiratory function and physical fitness, as well as stress and anxiety levels. In part due to the devastation of the 2021 global pandemic, there has been a call for a more comprehensive telemedicine service in recent years. This demand has spiked a huge push for research into the possibility of non-invasive methods of measuring human vital signs such as BR. Contact based measurement of this metric is typically carried out using a wearable sensor like a spirometer or breathing belt however, in many cases, where these methods are not possible, it must be detected through manual counting. The downsides to these methods, although accurate, is that they are intrusive, can lead to the spread of infection and are often not feasible in situations where the person is acutely unwell.

In addition to this, the need for accurate non-invasive BR detection is increasingly present due to recent legislation requiring all EU cars to implement a form of Driver Monitoring System (DMS) from 2026 [M. Bassani, 2023]. The purpose of DMS is to alert drivers to deteriorating health levels, such as drowsiness, stress and distraction and will be instrumental in reducing the number of collisions on the road and improving overall driver safety. For this reason, this study is being carried out by Xperi Inc. to further research in this area.

The majority of existing solutions for this application in terms of optical systems use thermal imaging to detect the BR of a person by analysing the variation in the temperature of the skin caused by respiration. However, the drawbacks of this method are the high cost, lack of spatial resolution and depth perception and the sensitivity to environmental factors. For this reason, this study explores the use of optical flow as a method of detecting BR using both near infrared (NIR) and RGB videos. Optical flow is a computer vision technique that tracks the movement of specific pixels between consecutive frames in a video. The method outlined in this paper compares the accuracy of the detected BR when using specified points on the face compared with points on the chest.

## 2 Breathing Rate Extraction Method

### 2.1 Optical Flow

Optical flow is a computer vision technique that is commonly used to track the motion of objects in a video [Tianqi Guo1 and Allebach1, 2021]. It works under the ‘brightness constancy assumption’ which is that pixel intensity does not change between consecutive frames. This assumption allows for the displacement of each pixel between frames to be calculated and processed to generate a meaningful signal. In this case, sparse optical flow is used, meaning that motion vectors are calculated only from the specific set of points passed in from the face detector or open pose. The optical flow algorithm used is the Lucas-Kanade feature tracker with a window size of 20 x 20. Testing determined this to be the most effective window size for this study as it could most effectively track pixels of homogeneous texture or colour.

The points to track are detected in the first frame of the video and passed into optical flow. These points are then tracked through each consecutive frame of the video. The difference in position of the y co-ordinate of the points between frames is used to generate the resulting raw signal which is then filtered and used to calculate breathing rate.

This study assesses whether BR can be accurately detected from the face and chest and compares the relative merits of the three methods of detection outlined in further detail below. The first and second method use three points on the face and chest respectively, and the third method uses a triangular grid as points for optical flow to track.

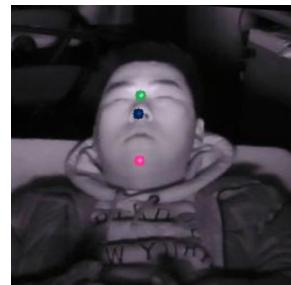
### 2.2 Face and Chest Point Detection

In order to determine the facial points to track, the KAPAO (Keypoints and Poses as Objects) human pose estimation method was used [McNally et al., 2022]. Face detection is carried out on the first frame of the input video to ascertain the location of the three points of interest. The chosen points are the midpoint between the eyes, the nose, and the chin. These points were identified as those least affected by facial deformation or noise. The points on the lips and eyes were too susceptible to noise and therefore discarded.

The chest points were identified using Open Pose, an open-source computer vision library that uses deep learning to track key points on the human body [Zhe Cao, 2021]. The chest movement was tracked using two similar approaches. The first approach uses the left shoulder, right shoulder, and neck points as arguments for optical flow to track while the second approach uses a triangular grid of points with the shoulder points used as the base.

### 2.3 Post Processing

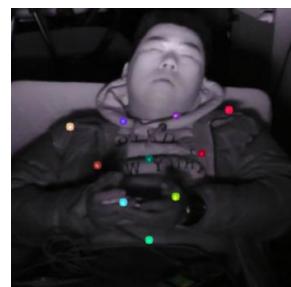
To extract the BR signal from the raw output signal of the optical flow, the signal is band passed in the range of 0.1 Hz to 0.5 Hz which corresponds to a BR of 6-30 bpm. The BR of a healthy adult is normally between 10 and 20 bpm. The BR itself is calculated by dividing the number of peaks detected by the peak detector on the filtered signal by the length of the video in seconds and multiplying by 60.



**Figure 1 Points tracked on face**



**Figure 2 Points tracked on chest**



**Figure 3 Triangular chest grid**

## 2.4 Dataset

This method was tested on two different sets of data in order to capture different poses and rates of respiration. The first set of data consisted of 6 videos, captured internally at Xperi Inc. Each video was 30 seconds in length and consisted of a participant sitting upright in a chair against a neutral background breathing at different rates (14-26 bpm). The subject changes clothes after the first three videos from a patterned shirt to a single colour top to test the effectiveness of optical flow when tracking pixels of the same colour. To compliment this dataset, videos from a publicly available sleep database were incorporated in the testing [Menghan Hu et al., 2018]. This database consists of videos of 12 sleeping participants captured using near infrared (NIR) cameras. From the sleep database, only 6 videos were selected for analysis due to excessive movement by the subject which distorted the results. The purpose of the additional data was to test the success of the method when participants were lying down and to assess whether the method was effective using NIR videos.

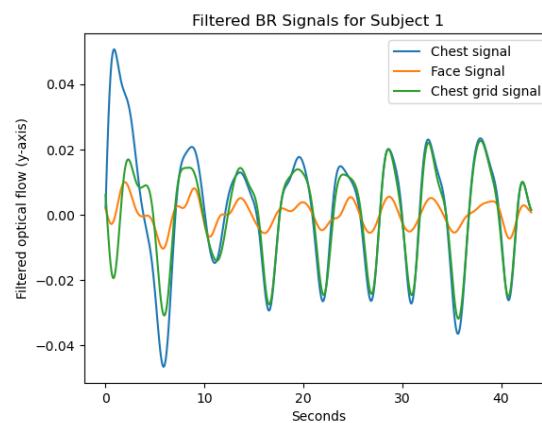
## 3 Results

Each method was tested on six videos, three NIR and three from the webcam footage. The videos selected all had range of slight participant movement. This aided in determining the effects of noise on the signal. The resulting graphs were firstly compared against the ground truth signal for accuracy and then against each other to determine signal strength. The chest points achieved an accurate BR signal in each test.

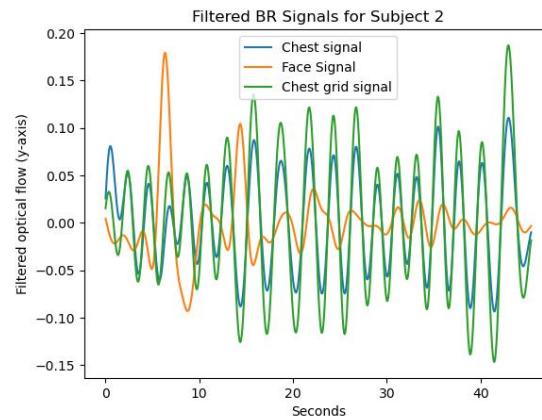
The resulting signals showed that BR was accurately detectable using each of the three methods, if there was no external noise such as head movement/arm movement. The BR signal from the face provided the weakest signal as overall it was the signal most consistently effected by noise. However it did provide an accurate signal in the videos with the low participant motion and particularly the sleeping videos with the face at a slight angle, as shown in Figure 4.

The signal from the three points on the chest resulted in a strong signal and had the lowest RMSE of the three methods as shown in Table 1 below in all the tests as it could function at a low optical flow window size and was not overly affected by noise. The chest grid provided the strongest signal overall but did not perform as well in the webcam videos where the participant was wearing a solid green jumper. Optical flow had difficulty tracking the points and the window size had to be increased to 40 x 40. This was not an issue with the NIR footage.

The graph in Figure 5 is an example of a video with a relatively large amount of facial movement. The signal from the face is distorted while the chest signals are both accurate.



**Figure 4** BR signal for subject 1 (NIR video)



**Figure 5** BR signal for subject 5 (webcam video)

	Face Points	Chest Points	Chest Grid
Average RMSE	7.0288	0.6305	0.76511

**Table 1** Average RMSE's for each method

Table 1 shows that this method produces a lower RMSE than many other BR detection methods which report RMSE of 1.7 and 2.03 [Jafar Pourbemany and Zhu, 2021] [Tianqi Guo1 and Allebach1, 2021]. The two chest point variations give very low RMSEs with the three-point input being slightly more accurate in detecting the actual BR value. Despite the strength of the results, further development is required to improve the model so that it can function in situations when there is increased movement from the participant. This method performs well on footage with low levels of noise but would not work on a moving participant.

## 4 Conclusion and Further Work

This method of breathing rate detection has produced accurate results and could be further developed to perform accurate real-time non-invasive breathing rate detection. Future work on this will involve improving the robustness of the system against noise caused by participant motion. This could be done by analysing the driver pose and rejecting signals where there is a large amount of noise in the frame is evident. It is also possible that a mixture of all three variations could be used in conjunction with each other to create a system that adapts to the motion present in the system and selects the most probable signal with the least noise. In terms of the triangular chest grid, the strength and reliability of the signal could be further improved by preprocessing the image in order to enhance the gradients in the image and provide more distinct points for optical flow to track.

## References

- [Jafar Pourbemany and Zhu, 2021] Jafar Pourbemany, A. E. and Zhu, Y. (2021). Real time video based heart and respiration rate monitorings. *NAECON 2021-IEEE National Aerospace and Electronics Conference*.
- [M. Bassani, 2023] M. Bassani, L. Catani, A. H. A. H. A. L. A. P. L. T. (2023). Do driver monitoring technologies improve the driving behaviour of distracted drivers? a simulation study to assess the impact of an auditory driver distraction warning device on driving performance. *Transportation Research Part F: Traffic Psychology and Behaviour*, 95:239–250
- [Menghan Hu et al., 2018] Menghan Hu, G. Z., Duo Li, Y. F., Huiyu Duan, W. Z., and Yang, X. (2018). Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation. *PLoS ONE*, 13(1):e0190466.
- [McNally et al., 2022] McNally, W., Vats, K., Wong, A., and McPhee, J. (2022). Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. *In European Conference on Computer Vision*, pages 37–54. Springer.
- [Chengxu Yang and Duan, 2021] Chengxu Yang, Xinxin Huang, Y. Z. Y. X. and Duan, X. (2021). Non-contact breathing rate detection based on time of flight sensor. *Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 43:7284-7287
- [Tianqi Guo1 and Allebach1, 2021] Tianqi Guo1, Q. L. and Allebach1, J. (2021). Remote estimation of respiration rate by optical flow using convolutional neural networks. *Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 43:7284 – 7287.
- [Zhe Cao, 2021] Zhe Cao, Gines Hidalgo, T. S. S.-E. W. Y. S. (2021). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186

# Non-Contact NIR PPG Sensing through Large Sequence Signal Regression

Timothy Hanley, Dara Golden, Robyn Maxwell, Joseph Lemley, and Ashkan Parsi

*OCTO Sensing Team, Xperi Inc., Galway, Ireland*

## Abstract

Non-Contact sensing is an emerging technology with applications across many industries from driver monitoring in vehicles to patient monitoring in healthcare. Current state-of-the-art implementations focus on RGB video, but this struggles in varying/noisy light conditions and is almost completely unfeasible in the dark. Near Infra-Red (NIR) video, however, does not suffer from these constraints. This paper aims to demonstrate the effectiveness of an alternative Convolution Attention Network (CAN) architecture, to regress photoplethysmography (PPG) signal from a sequence of NIR frames. A combination of two publicly available datasets, which is split into train and test sets, is used for training the CAN. This combined dataset is augmented to reduce overfitting to the ‘normal’ 60 – 80 bpm heart rate range by providing the full range of heart rates along with corresponding videos for each subject. This CAN, when implemented over video cropped to the subject’s head, achieved a Mean Average Error (MAE) of just 0.99 bpm, proving its effectiveness on NIR video and the architecture’s feasibility to regress an accurate signal output.

**Keywords:** Remote Photoplethysmography, NIR Sensing, Heart Rate, Driver Monitoring

## 1 Introduction

Non-contact sensing is the act of obtaining an individual’s health signals, without any hardware, etc. being physically in contact with them. This is usually achieved using cameras to detect changes or motions often imperceptible to the human eye, that can be regressed to obtain the desired health metric. In the case of PPG, there are slight changes in colour to the skin, caused by blood rushing to and from the heart [Wu et al., 2012]. These colour changes are detectable in both NIR and RGB video, however, they are more pronounced in RGB.

This has enormous potential across multiple sectors from applications in the health industry to in-cabin driver monitoring systems (DMS). There are health situations where it may be uncomfortable for the subject to ‘wear’ the sensors, or it may be the case that it is simply unfeasible to deploy a contact-based sensor, such as in a DMS. In a health setting, NIR can operate in the dark, to allow for monitoring of the patient’s heart rate throughout the night or when they are sleeping, with no discomfort. In a DMS, NIR, especially in the range of 940nm, provides substantial reductions in noise in comparison to RGB, reducing the noise produced by external and uncontrollable factors [Magdalena Nowara et al., 2018]. The use of NIR cameras, along with suitable NIR illuminators, can offset some of the problems encountered in these scenarios.

Xperi’s research group proposes a method to accurately calculate heart rate by means of regressing a PPG signal from NIR video. This method consists of a CAN architecture that predicts a large sequence of PPG signals, given a large sequence [Liu et al., 2020] of NIR frames as inputs.

## 2 Methodology

### 2.1 Model

The model utilises a CAN architecture, heavily influenced by DeepPhys [Chen and McDuff, 2018], with the final layer adjusted to predict N signal samples. This adjusted layer also employs the Snake activation function [Ziyin et al., 2020], to improve the model's capability to learn the semi-periodic signal. When regressing a signal where the length of the PPG sample is greater than N, the inference is run for every N sample sequence, with the common outputs averaged to produce the signal waveform.

### 2.2 Dataset

The dataset used for training and testing purposes is a combination of both publicly available MR-NIRP datasets produced by the Rice Computational Imaging Lab [Magdalena Nowara et al., 2018] [Nowara et al., 2020].

#### 2.2.1 Dataset Corrections

Due to discrepancies in the dataset, such as an inconsistent PPG sampling frequency and dropped frames, much of the initial work was focused on correcting these. The varying sampling frequency of the PPG ground truth signal was rectified by considering the dropping of samples at the buffer and interpolating those missing samples. Further, any videos/portions of videos where the frames/signals could not be verified were removed for the purpose of this experiment.

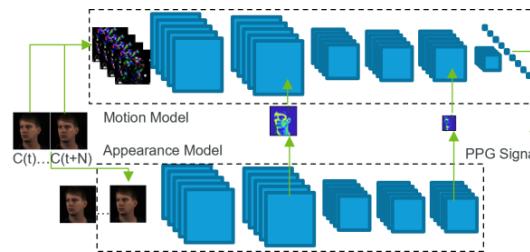
#### 2.2.2 PPG Normalisation

To prevent the model from encountering issues learning the DC component of the PPG signal, these signals were normalised between 1 and 0, in such a way that each peak is a 1, and each trough is a 0. The theory behind this is that the model may, given a large sequence (64 samples,  $\approx 2$  seconds), find it easier to locate the peaks in the sequence rather than detect and quantify the increase/decrease in the signal.

#### 2.2.3 Heart Rate Augmentation

Initial work showed that the model was liable to overfit to the average heart rate range of the dataset, which in this case was discovered to be 60 – 80 bpm. Therefore, to ensure a broader range of heart rates are successfully detected, the dataset was augmented to provide an equal distribution of heart rates in the 40 – 140 bpm range.

This augmentation is achieved by effectively ‘stretching’ or ‘squeezing’ the signals and videos with samples interpolated to create an effective 30 fps video with a corresponding 30 Hz PPG signal. Heart rates were chosen at random in bins of 10 bpm, and this is used as the target heart rate for augmentation. The data provided by each



**Figure 1: CAN architecture [Chen and McDuff, 2018]**

	MR-NIRP (Indoor)	MR-NIRP (Driving)
No. of Subjects	8	19
No. of Videos	15	190
NIR Wavelengths	940 nm	940 nm, 975 nm
Image Dimension	640 x 640	640 x 640
Scenarios	Indoor	Garage (Indoor), Driving
Motion Levels	Still, Small	Still, Small, Large

**Table 1: MR-NIRP datasets outline**

	MR-NIRP Augmented
No. of Subjects	26
No. of Videos	2079 (Augmented & Original)
NIR Wavelengths	940 nm, 975 nm
Image Dimension	64 x 64
Scenarios	Indoor, Garage (Indoor), Driving
Motion Levels	Still, Small, Large
Heart Rate Ranges	40 – 140 bpm

**Table 2: Augmented and combined MR-NIRP dataset**

original video has effectively been multiplied by 10. All videos with augmented heart rates are trimmed to the same length, to prevent overfitting to the slower heart rates.

#### 2.2.4 Face Detection, Cropping, and Resizing

One further augmentation of the data is to remove some unnecessary data in the background. In general, the subjects in MR-NIRP (Indoor) were closer to the camera than those in MR-NIRP (Driving). This is rectified by cropping with 25% padding to the face, detected by a non-public industry face detector. This should allow for more detail from the face to transfer into the resized images.

### 2.3 Training & Evaluation

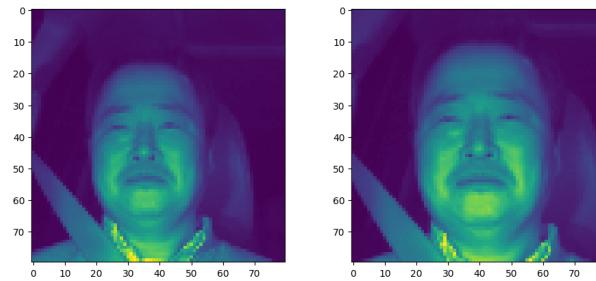
The model is trained on an 19/7 subject train/test split, training on the augmented data but testing on the original data. For continuity, only 940 nm video is used, and the videos with motion are excluded from this experiment.

MSE is used as the loss function, however, potential improvements could be seen from frequency-aware loss functions, as MSE can severely punish phase-shifted signals. For validation, the MAE of the HR (calculated from R-R Intervals) across a whole video is used, to provide a fair comparison between this and different solutions.

## 3 Results

As expected, the model performs slightly better when trained on the cropped frames, as seen in Table 4, however, the difference is more marginal than anticipated. This architecture performs better than [Nowara et al., 2021], which was trained on RGB video and tested only on the MR-NIRP (Indoor). DeepPhys [Chen and McDuff, 2018] performed better on the NIR video, however, that video focused on the neck/underside of the head, which is subject to substantially less noise than the face, while also using a much smaller dataset.

A visual inspection, as shown in Figure 3, shows that the model can consistently correctly predict peak locations, along with the respective waveforms. It also shows promise of detecting the dicrotic notch, however, it is likely that cleaner signals would be required for the model to learn these successfully.



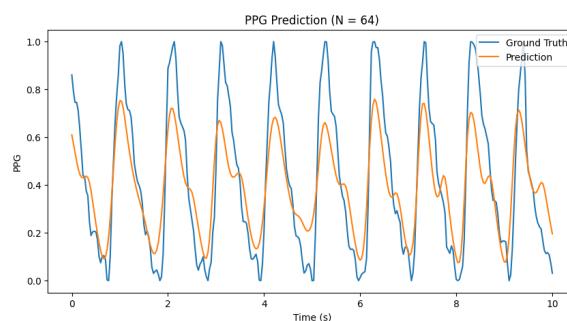
**Figure 2: Uncropped vs cropped comparison**

	Train	Test
No. Subjects	19	7
No. Subjects (Indoor)	5	3
No. Subjects (Driving)	14	4
NIR Wavelengths	940 nm	940 nm
Motion Levels	Still	Still
Normalised PPG	Yes	Yes
Augmented HR	Yes	No

**Table 3: Train and test sets**

	MAE (bpm)
Uncropped	1.07
Cropped	0.99

**Table 4: MAE results**



**Figure 3: Example prediction**

Overall, the model shows promising signs of being able to produce accurate heart rate results while also regressing an accurate PPG signal, which may be required for further analysis.

## 4 Conclusion

Future work on this concept will include different image dimensions and different sequence lengths. The model should also be retrained on subsets of the dataset that contain the videos with motion, to improve robustness to subject movement. Further performance gains may be made by cleaning the PPG signal through band passing or other filtering methods to remove unnecessary noise. Further testing should include validation on Xperi's in-house datasets to ensure extensibility to other data.

## References

- [Chen and McDuff, 2018] Chen, W. and McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365.
- [Liu et al., 2020] Liu, X., Fromm, J., Patel, S., and McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411.
- [Magdalena Nowara et al., 2018] Magdalena Nowara, E., Marks, T. K., Mansour, H., and Veeraraghavan, A. (2018). Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1272–1281.
- [Nowara et al., 2020] Nowara, E. M., Marks, T. K., Mansour, H., and Veeraraghavan, A. (2020). Near-infrared imaging photoplethysmography during driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3589–3600.
- [Nowara et al., 2021] Nowara, E. M., McDuff, D., and Veeraraghavan, A. (2021). The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4955–4964.
- [Wu et al., 2012] Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., and Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8.
- [Ziyin et al., 2020] Ziyin, L., Hartwig, T., and Ueda, M. (2020). Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594.

# Spurious Correlation Mitigation in CXR Images via Reinforcement learning and Self-Supervision

Weichen Huang<sup>1</sup>, Kathleen M. Curran<sup>2</sup>

<sup>1</sup>*St Andrew's College Dublin, Dublin, Ireland*

<sup>2</sup>*School of Computer Science, University College Dublin, Dublin, Ireland*

## Abstract

In the medical domain, accurate interpretation of chest X-ray (CXR) images is critical for diagnosis and treatment decisions. However, deep learning models trained on large datasets can be susceptible to spurious correlations, leading to erroneous interpretations and potentially harmful decisions. This study aims to address this issue in the CXR domain by proposing the use of reinforcement learning techniques and semi-supervised training. These methods actively select relevant CXR data samples while mitigating the influence of spurious correlations. The results demonstrate the effectiveness of these approaches in improving prediction accuracy and decision-making performance compared to traditional data selection methods. This research contributes to the advancement of both technical state-of-the-art and clinical applications of deep learning in healthcare.

**Keywords:** Medical Imaging, Machine Vision, Reinforcement Learning, Self-Supervised Learning, Spurious Correlations

## 1 Introduction

The article tackles the challenge of spurious correlations in deep learning, specifically in medical domains such as CXR image classification [Nguyen et al., 2021]. Spurious correlations can lead to inaccurate diagnoses and treatment decisions. To address this issue, the article proposes a reinforcement learning-based data selection framework that integrates self-supervised training [Calude and Longo, 2017]. By actively selecting relevant data samples, the framework enhances the accuracy of the classification model and improves diagnostic reliability. Empirical evaluations demonstrate the effectiveness of this approach in mitigating spurious correlations and enhancing prediction accuracy. In the medical domain, this framework shows promise in improving the trustworthiness and effectiveness of diagnostic models, resulting in accurate predictions and improved patient outcomes [Nguyen et al., 2021]. This work offers several contributions in tackling spurious correlations in the medical domain. Firstly, it emphasizes the importance of developing solutions to mitigate these correlations, particularly in diagnostic tasks. Secondly, the work proposes a unified reinforcement learning (RL) based framework that incorporates an adaptable data relevance assessment system. This framework considers the inter-dependence between task-specific data relevance assessment and the target task, aiming to reduce the impact of spurious correlations. Lastly, the framework is evaluated on a public dataset of chest X-ray images, with a specific focus on the diagnostic task of pneumonia detection.

## 2 Related Work

Addressing spurious correlations in data is a critical challenge in AI and machine learning. Previous methods, such as feature selection and engineering [Deng et al., 2023], regularization techniques [Kirichenko et al., 2023], and causal inference [Cui and Athey, 2022], have limitations in complex scenarios. Our proposed approach

combines reinforcement learning (RL) and self-supervised learning (SSL) techniques to overcome these limitations. Feature selection and engineering, while effective, are manual processes that may not capture all relevant features in complex scenarios [Deng et al., 2023]. In contrast, our RL-based approach automatically learns the most relevant features and discards spurious correlations, reducing the need for manual intervention. Causal inference techniques rely on observational data and human intervention, limiting their applicability in complex scenarios [Cui and Athey, 2022]. Our RL-based method actively selects informative data points, reducing bias and spurious correlations without solely relying on observational data. By combining RL and SSL, our approach effectively addresses spurious correlations [Cui and Athey, 2022]. RL enables dynamic data selection based on rewards and penalties, while SSL reduces the influence of spurious correlations and improves generalization capabilities. Our approach overcomes the shortcomings of previous methods, enhancing the model's feature selection, interpretability, and robustness [Deng et al., 2023, Kirichenko et al., 2023, Cui and Athey, 2022]. It provides an effective solution for mitigating spurious correlations in data, improving the reliability of AI and machine learning models.

### 3 Methodology

#### 3.1 Problem formulation

Spurious correlations refer to statistical relationships that appear to exist between variables but are not causally related. These correlations are often coincidental or arise due to confounding factors.

Let's assume we have two variables,  $X$  and  $Y$ , and we denote their sample sets as  $S_X$  and  $S_Y$  respectively. A correlation between  $X$  and  $Y$  can be quantified using the Pearson correlation coefficient, denoted by  $r(X, Y)$ . The Pearson correlation coefficient measures the linear relationship between two variables and ranges from -1 to 1.

The presence of a spurious correlation means that  $X$  and  $Y$  exhibit a non-causal, coincidental association. In other words, their correlation arises due to the influence of an unaccounted third variable,  $Z$ , which affects both  $X$  and  $Y$  independently.

Mathematically, we can describe spurious correlations as follows:

Given  $X$ ,  $Y$ , and  $Z$ , the observed correlation between  $X$  and  $Y$ , denoted as  $r_{XY}$ , can be expressed as:

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Where:  $X_i$  and  $Y_i$  represent individual data points from the sample sets  $S_X$  and  $S_Y$  respectively.  $\bar{X}$  and  $\bar{Y}$  are the mean values of the sample sets  $S_X$  and  $S_Y$  respectively.

However, if we consider the correlation between  $X$  and  $Y$  while controlling for the influence of  $Z$ , denoted as  $r_{XY.Z}$ , we would perform partial correlation. The partial correlation coefficient is used to measure the relationship between  $X$  and  $Y$  after accounting for the effect of  $Z$ .

The spurious correlation between  $X$  and  $Y$  is present if  $r_{XY.Z}$  is substantially different from  $r_{XY}$  (the observed correlation without considering  $Z$ ). In this case, the correlation between  $X$  and  $Y$  in the presence of  $Z$  disappears or becomes significantly weaker, indicating that the original correlation was spurious.

Our approach combines self-supervised learning and spurious feature correction to simultaneously train a task predictor  $f(x; w)$  and a data selection controller  $h(x; \theta)$ . Reinforcement learning is used to modify a parameter associated with the target task through the controller, aiming to maximize task performance. We utilize a recurrent neural network (RNN) for adaptability. The integrated approach incorporates self-supervised learning, emphasizes spurious feature correction, and operates within a meta-learning framework. In this section, we address spurious correlations by assessing data relevance using the task predictor and controller functions. The controller assigns data relevance scores to improve task performance over time, guided by task performance feedback.

### 3.2 Justification of Reinforcement Learning Approach

The proposed data relevance framework can be formulated as the following bi-level minimization problem:

$$\begin{aligned} & \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{P}_{xy}} [L_h(f(x; w^*), y) h(x; \theta)], \\ \text{s.t. } & w^* = \arg \min_w \mathbb{E}_{(x,y) \sim \mathcal{P}_{xy}} [L_f(f(x; w), y) h(x; \theta)], \\ & \mathbb{E}_{x \sim \mathcal{P}_x} [h(x; \theta)] \geq c > 0. \end{aligned}$$

This problem can be restructured to allow sampling or selection based on controller outputs by considering the data  $x$  and  $(x, y)$  to be sampled from the controller-selected or -sampled distributions  $P^h(X)$  and  $P^h(XY)$ , with probability density functions  $p_h(x) \propto p(x)h(x; \theta)$  and  $p_h(x, y) \propto p(x, y)h(x; \theta)$ , respectively. Thus, reformulating to facilitate sampling or selection, we can rewrite the bi-level minimization problem as follows:

$$\begin{aligned} & \min_{\theta} \mathbb{E}_{(x,y) \sim p_{xy}^\dagger} [L_h(f(x; w^*), y)], \\ \text{s.t. } & w^* = \arg \min_w \mathbb{E}_{(x,y) \sim p_{xy}'} [L_f(f(x; w), y)], \\ & \mathbb{E}_{x \sim p_x^\dagger} [1] \geq c > 0. \end{aligned}$$

The formulated data relevance assessment problem can be learned in a RL-based meta-learning framework. In this work, we outline a general RL-based meta-learning framework to learn adaptable data relevance assessment.

The proposed formulation can be modeled as a finite-horizon Markov decision process (MDP) [Puterman, 1990], with the controller interacting with, and influencing, an 'environment,' which contains the task predictor and the data used to train such a function. The MDP environment for this data relevance problem consists of the data from  $\mathcal{P}_X$  and the target task predictor  $f(\cdot; w)$ . At time-step  $t$ , the observed state of the environment  $s_t = (f(\cdot; w_t), \mathcal{B}_t)$  is composed of the target task predictor  $f(\cdot; w)$  and a batch of samples  $\mathcal{B}_t = \{x_i\}_{i=1}^B$  from a train set  $\mathcal{D}_{\text{train}} = \{x_i\}_{i=1}^N$  from the distribution  $\mathcal{P}_X$ . If each MDP environment is defined as  $M_k$ , the distribution and task predictor within the environment can be defined as  $\mathcal{P}_{X,k}$  and  $f_k(\cdot; w_k)$ , respectively. However, in further analysis, we omit  $k$  from these expressions for notational convenience.

Reinforcement learning allows training of a controller to maximize a reward obtained based on controller-environment interactions, considered as an MDP. In RL, the MDP is represented as a 5-tuple  $(S, \mathcal{A}, p, r, \pi)$ .  $S$  is the state space and  $\mathcal{A}$  is the continuous action space.  $p: S \times S \times \mathcal{A} \rightarrow [0, 1]$  is the state transition distribution conditioned on state-actions, where  $p(s_{t+1} | s_t, a_t)$  represents the probability of the next state  $s_{t+1} \in S$  given the current state  $s_t \in S$  and action  $a_t \in \mathcal{A}$ .

The reward function is denoted by  $r: S \times \mathcal{A} \rightarrow \mathbb{R}$ , and  $R_t = r(s_t, a_t)$  denotes the reward given the current state  $s_t$  and action  $a_t$ . The policy,  $\pi(a_t | s_t): S \times \mathcal{A} \in [0, 1]$ , represents the probability of performing action  $a_t$  given the state  $s_t$ . The controller interacting with an environment creates a trajectory of states, actions, and rewards,  $(s_1, a_1, R_1, s_2, a_2, R_2, \dots, s_T, a_T, R_T)$ , where the subscript indicates the time-step.

The goal of the agent is to maximize the cumulative reward over a trajectory. The cumulative reward is the discounted sum of accumulated rewards starting from time-step  $t$ :  $Q^\pi(s_t, a_t) = \sum_{k=0}^T \gamma^k R_{t+k}$ , where the discount factor  $\gamma \in [0, 1]$  is used to discount future rewards. The objective of the controller is to learn a parameterized policy  $\pi_\theta$  that maximizes the expected return  $J(\theta) = \mathbb{E}_{\pi_\theta}[Q^\pi(s_t, a_t)]$ . The central optimization problem in RL can be expressed as:

$$\theta^* = \operatorname{argmax}_\theta J(\theta),$$

where  $\theta^*$  denotes optimal policy parameters.

We propose to train the controller using RL, where the controller outputs sampling probabilities  $\{h(x_{i,t}, \theta)\}_{i=1}^B$  based on the input images. The action  $a_t = \{a_{i,t}\}_{i=1}^B \in \{0, 1\}^B$  leads to a sample selection decision for target task

predictor training if  $a_{i,t} = 1$ . The selection is done based on  $a_{i,t} \sim \text{Bernoulli}(h(x_{i,t}; \theta))$ . The policy  $\pi_\theta(a_t | s_t)$  is defined as:

$$\log \pi_\theta(a_t | s_t) = \sum_{i=1}^B h(x_{i,t}; \theta) a_{i,t} + (1 - h(x_{i,t}; \theta))(1 - a_{i,t}).$$

In this formulation, the reward  $R_t$  is formulated based on the metric function that measures the performance of the target task,  $L_h$ .

The proposed reinforcement learning (RL) approach addresses the issue of data relevance by training a controller to select or sample data points based on their relevance to the target task. The RL agent learns an optimal policy by maximizing the expected return, effectively identifying and emphasizing the most relevant data during training. This iterative process improves the performance of the target task predictor by efficiently utilizing informative examples while minimizing the negative impact of noisy or irrelevant data. Ultimately, the RL-based data relevance assessment optimizes the data selection strategy, resulting in improved training efficiency and better task performance.

### 3.3 Optimizing the task predictor

We propose SimCLR [Chen et al., 2020], a self-supervised learning technique, to optimize the task predictor  $f(x; w)$ . SimCLR maximizes similarity between augmented views of the same image while minimizing similarity between views of different images, enabling robust and transferable image representations. After pre-training, we perform supervised fine-tuning on a small labeled dataset selected based on the controller's scores. In fine-tuning, we utilize cross-entropy loss to adapt the pre-trained CNN to the labeled dataset. Additionally, we experiment with a purely supervised learning method, leveraging all labels from the training set.

### 3.4 Optimizing the controller model

The controller is optimized by minimizing the weighted metric function  $L_h : Y \times Y \rightarrow R_{\geq 0}$  on the validation set. The controller predicts lower data relevance scores for samples with higher metric function values, indicating lower task performance. A constraint prevents the trivial solution. The data relevance assessment problem is learned using a RL-based meta-learning framework modeled as a finite-horizon Markov decision process (MDP) [Puterman, 1990]. Reinforcement learning trains the controller to maximize the reward obtained through controller-environment interactions within the MDP framework. The MDP is defined by its state space  $S$ , continuous action space  $\mathcal{A}$ , state transition distribution  $p$ , reward function  $r$ , and policy  $\pi$ . The goal is to learn a parameterized policy  $\pi_\theta$  that maximizes the expected return  $J(\theta)$ .

### 3.5 Data Environment

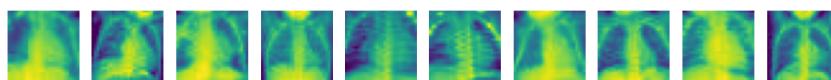


Figure 1: These images are examples of normal chest X-ray images from the PneumoniaMNIST dataset.

The proposed framework's ability to detect spurious features was evaluated using the publicly available PneumoniaMNIST dataset [Yang et al., 2023], consisting of a total of 5856 chest X-ray images with binary labels for pneumonia diagnosis, split into train, validation and holdout sets. This dataset allows for evaluating the framework's performance in pneumonia detection while demonstrating its applicability beyond a single modality, dataset, or task. Gaussian noise with random intensities and random obstructions were added to simulate real-world variations and occlusions in imaging data. These corruptions allowed assessing the framework's ability to detect and mitigate spurious features in pneumonia detection using chest X-ray images, ensuring relevance and comparability to existing studies. The evaluation focused on the framework's capability to detect

and exclude irrelevant data. A separate holdout set was used for evaluation, where samples were sorted based on controller predictions. The holdout set rejection ratio varied from 0% to 100% in 10% increments to assess the controller's ability to detect spurious features. The remaining training data was used to train the controller.

### 3.6 Training

The controller is trained using the DDPG algorithm [Li et al., 2019] with empirically configured hyperparameters. An Alex-Net-style architecture [Yan et al., 2015] serves as the target task predictor, trained with cross-entropy loss and classification accuracy-based rewards. The controller underwent 100 episodes of training with a batch size of 64, while the task predictor underwent 100 epochs of training with the same batch size.

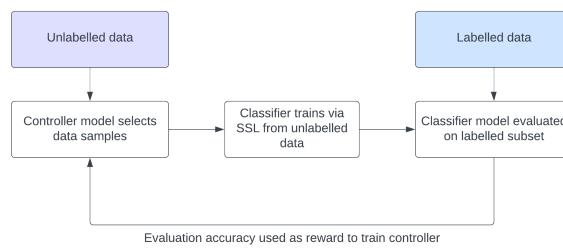


Figure 2: Diagram of the proposed framework. The controller takes in the data and outputs a ranking of data relevance. The task predictor takes in the data and outputs a prediction of the target task. The controller is trained to maximise the reward based on the task predictor output.

## 4 Evaluation

### 4.1 Evaluation Procedure

The goal is to evaluate the task-predictor accuracy based on reinforcement learning data selection and random data selection. We also evaluate the efficacy of self-supervised learning vs supervised learning of the task-predictor.

These methods are evaluated on the holdout dataset. The procedure for evaluation is the following:

- The trained controller model takes in the holdout set and generates the data-relevance rankings.
- A proportion (with value  $k$ ,  $0 < k < 1$ ) of the data samples are removed based on the ranking ( $nk$  data samples are removed for a holdout set of size  $n$ )
- The remaining data is used to evaluate the task predictor model which outputs evaluation metrics.

### 4.2 Results

Reject ratio	0.0	0.2	0.4
RL (Supervised)	<b>0.835 ± 0.0251</b>	<b>0.836 ± 0.0325</b>	<b>0.821 ± 0.031</b>
RL (Unsupervised)	0.8109 ± 0.0317	<b>0.824 ± 0.0354</b>	0.8027 ± 0.0346
No selection (Supervised)	<b>0.815 ± 0.032</b>	0.815 ± 0.032	<b>0.815 ± 0.032</b>
Random selection (Supervised)	0.810 ± 0.0272	0.800 ± 0.03	0.792 ± 0.034

Table 1: This table shows the accuracy of each technique on the holdout set. The standard deviation of each value is also shown.

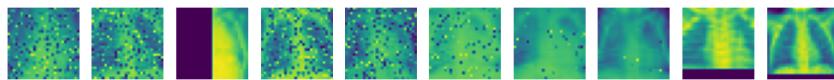


Figure 3: This is a spectrum of images from lowest score to highest score based on the controller model’s predictions. This shows that the model can highlight images that have little to no distortions (right side), and can classify when there are spurious correlations in the images (left side).

Comparing results from Table 1 in the classification task, both proposed RL-based algorithms show significant improvements over non-selective baselines. However, the unsupervised method has lower accuracy due to the lack of labeled data. While it still provides insights, its accuracy is generally lower than supervised methods. The peak accuracy for unsupervised learning was 82.4% at 0.2 rejection ratio, while supervised learning achieved 83.6% at the same ratio. Figure 3 provides a visual interpretation of the controller model’s predictions. Images on the left have low selection scores, while images on the right have high scores. The model effectively highlights images with minimal distortions (right) and detects spurious correlations (left). This empirical evidence supports the model’s ability to identify images with and without spurious correlations.

## 5 Conclusions and Future Work

The study concludes that RL techniques and self-supervised training effectively select relevant data samples from large unlabeled datasets given a small labeled dataset. The proposed RL-based framework addresses spurious correlations in deep learning by considering task-specific data relevance assessment and the target task. Unsupervised methods provide valuable insights but have slightly lower accuracy compared to supervised methods due to the lack of labeled data. The observation of peak performance before decreasing raises questions about the reasons behind this behavior, such as prediction variance, RL algorithm overfitting, and dataset limitations. Future work can explore different RL algorithms or self-supervised learning techniques for data selection, evaluate on larger datasets and complex models for scalability, and incorporate explanatory supervision for enhanced efficacy and interpretability. In the medical domain, the RL-based framework has potential to improve diagnostic models, reducing misdiagnosis risks and improving patient outcomes. Future work in medicine could focus on fine-tuning the framework for medical imaging tasks, incorporating domain-specific knowledge and evaluating on larger and diverse medical datasets.

## 6 Acknowledgements

We would like to express our gratitude to the MedMNIST team for providing their data, which was instrumental in conducting this research. Additionally, we acknowledge the contributions of previous works in this field that have paved the way for our study. Furthermore, we extend our thanks to our colleagues and collaborators for their valuable feedback and support throughout this project, as their insights have greatly contributed to the success of this research.

## References

- [Calude and Longo, 2017] Calude, C. S. and Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of science*, 22:595–612.
- [Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

- [Cui and Athey, 2022] Cui, P. and Athey, S. (2022). Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115.
- [Deng et al., 2023] Deng, Y., Yang, Y., Mirzasoleiman, B., and Gu, Q. (2023). Robust learning with progressive data expansion against spurious correlation.
- [Kirichenko et al., 2023] Kirichenko, P., Izmailov, P., and Wilson, A. G. (2023). Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*.
- [Li et al., 2019] Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., and Russell, S. (2019). Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4213–4220.
- [Nguyen et al., 2021] Nguyen, T., Nagarajan, V., Sedghi, H., and Neyshabur, B. (2021). Avoiding spurious correlations: Bridging theory and practice. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.
- [Puterman, 1990] Puterman, M. L. (1990). Markov decision processes. *Handbooks in operations research and management science*, 2:331–434.
- [Yan et al., 2015] Yan, L. C., Yoshua, B., and Geoffrey, H. (2015). Deep learning. *nature*, 521(7553):436–444.
- [Yang et al., 2023] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.