Irish Machine Vision and Image Processing Conference

Dublin City University, September 2-3, 2021

Conference Proceedings

IPRCS

Irish Pattern Recognition and Classification Society

Published by the Irish Pattern Recognition & Classification Society

iprcs.org

ISBN 978-0-9934207-6-4

Welcome

We are happy to welcome the speakers and delegates of the 23rd Irish Machine Vision and Image Processing Conference, IMVIP 2021! This year the conference is hosted at Dublin City University over two days, September 2-3.

The IMVIP Conference is Ireland's primary meeting for those researching in the fields of machine vision and image processing. The conference has been running since 1997 and provides a forum for the exchange of ideas and the presentation of research conducted both in Ireland and worldwide. IMVIP is a single-track conference consisting of high quality previously unpublished contributed papers focusing on both theoretical research and practical experiences in all areas. IMVIP is run in association with the Irish Pattern Recognition and Classification Society (iprcs.org), a member organisation of the International Association for Pattern Recognition (IAPR) and the International Federation of Classification Societies (IFCS).

This is the fourth time that DCU hosts IMVIP conference after 1999, 2006 and 2011 editions which were held at DCU in the past. We are delighted to be able to return to on-site format, after the fully online version of 2020. More specifically, this year we are offering a mixed format where the majority of the participants present on-site and some participate remotely. The technical program this year consists of 12 full and 4 short papers, all of which are delivered as oral presentations. We are delighted to have two keynote speakers this year: Prof. Xavier Giro-i-Nieto from Universitat Politecnica de Catalunya (UPC), Barcelona, and Prof. Noel O'Connor from Dublin City University.

We thank sincerely the members of the Programme Committee for generously giving their time, effort and expertise in reviewing the submissions. We wish all the attendees a pleasant and engaging experience with the mixed delivery format at IMVIP 2021.

Vladimir A. Krylov, Kevin McGuinness
Dublin City University
September 2021

Programme Chairs

Vladimir A. Krylov, Dublin City University
Kevin McGuinness, Dublin City University

Programme Committee

Martin Alain, Trinity College Dublin
Vincent Andrearczyk, HES-SO The University of Applied Sciences and Arts of Western Switzerland
Donald Bailey, Massey University, New Zealand
Francesco Bianconi, Università degli Studi di Perugia, Italy
Kathy Clawson, University of Sunderland, UK
Sonya Coleman, Ulster University
Joan Condell, Ulster University
David Corrigan, Huawei Technologies
Jane Courtney, Technological University, Dublin
Kathleen Curran, University College Dublin
Rozenn Dahyot, National University of Ireland, Maynooth
Kenneth Dawson-Howe, Trinity College Dublin
Catherine Deegan, Technological University, Dublin
Soumyabrata Dev, University College Dublin
Cem Direkoglu, Middle East Technical University, Cyprus
Antonio Fernández, Universidad de Vigo, Spain
Bob Fisher, University of Edinburgh, UK
Guillaume Gales, Foundry
Bryan Gardiner, Ulster University
Jonathan Horgan, Valeo Vision Systems
Ciaran Eising, Valeo Vision Systems, UL
Dermot Kerr, Ulster University
Yasuyo Kita, National Institute of Advanced Industrial Science and Technology (AIST), Japan
Vladimir A. Krylov, Dublin City University
Suzanne Little, Dublin City University
Charles Markham, National University of Ireland, Maynooth
Sally Mcclean, Ulster University
John McDonald, National University of Ireland, Maynooth
Kevin McGuinness, Dublin City University
Paul McKevitt, Ulster University
Derek Molloy, Dublin City University
Sean Mullery, Institute of Technology Sligo
Omar Nibouche, Ulster University
Robert Sadleir, Dublin City University
Bryan Scotney, Ulster University
Shane Gilroy, IT Sligo
Matej Ulicny, Trinity College Dublin
David Vernon, Carnegie Mellon University Africa, Rwanda
Rudi Villing, National University of Ireland, Maynooth
Paul Whelan, VSG DCU
Reyer Zwiggelaar, Aberystwyth University, Wales

Keynote Speakers

Prof. Xavier Giro-i-Nieto is an associate professor at the Universitat Politecnica de Catalunya (UPC) in Barcelona, member of the Image Processing Group (GPI), Intelligent Data Science and Artificial Intelligence Research Center (IDEAI-UPC), Institute of Industrial Robotics (IRI), and also a visiting researcher at Barcelona Supercomputing Center (BSC). He graduated in Telecommunications Engineering at ETSETB (UPC) in 2000, after completing his master thesis on image compression at the Vrije Universiteit in Brussels (VUB) with Prof. Peter Schelkens. After working one year in Sony Brussels, he returned to UPC to obtain a PhD on computer vision, supervised by Prof. Ferran Marqués and Prof. Shih-Fu Chang from the Digital Video and MultiMedia laboratory at Columbia University, that he repeatedly visited between 2008-2014. He regularly collaborates with the Insight Center of Data Analytics at Dublin City University, and is a member of the Governance Committee of the Science Foundation Ireland Centre for Research Training in Machine Learning. He serves as associate editor at IEEE Transactions on Multimedia, and reviews for top tier conferences in machine learning (NeurIPS, ICML), computer vision (CVPR, ECCV, ICCV) and multimedia (ACMMM, ICMR).

Prof. Noel E. O'Connor is a Full Professor in the School of Electronic Engineering at Dublin City University (DCU) Ireland. He is CEO of the Insight SFI Research Centre for Data Analytics, Ireland's largest SFI-funded research centre. He was previously Academic Director of DCU's Research and Enterprise Hub on Information Technology and the Digital Society, with the responsibility of coordinating multi-disciplinary ICT-related research across the university. The focus of his research is in multimedia content analysis, computer vision, machine learning, information fusion and multi-modal analysis for applications in security/safety, autonomous vehicles, medical imaging, IoT and smart cities, multimedia content-based retrieval, and environmental monitoring. Since 1999 he has published over 400 peer-reviewed publications, made 11 standards submissions, and filed 7 patents. He is an Area Editor for Signal Processing: Image Communication (Elsevier) and an Associate Editor for the Journal of Image and Video Processing (Springer). He is a member of the ACM and IEEE.

Table of Contents

Multi-Head Self-Attention via Vision Transformer for Zero-Shot Learning

Faisal Alamri and Anjan Dutta

Department of Computer Science, University of Exeter, United Kingdom

Abstract

Zero-Shot Learning (ZSL) aims to recognise unseen object classes, which are not observed during the training phase. The existing body of works on ZSL mostly relies on pretrained visual features and lacks the explicit attribute localisation mechanism on images. In this work, we propose an attention-based model in the problem settings of ZSL to learn attributes useful for unseen class recognition. Our method uses an attention mechanism adapted from Vision Transformer to capture and learn discriminative attributes by splitting images into small patches. We conduct experiments on three popular ZSL benchmarks (i.e., AWA2, CUB and SUN) and set new state-of-the-art harmonic mean results on all the three datasets, which illustrate the effectiveness of our proposed method.

Keywords: Generalised zero-shot learning, Inductive learning, Attention, Semantic embedding, Vision Transformer.

1 Introduction

Relying on massive annotated datasets, significant progress has been made on many visual recognition tasks, which is mainly due to the widespread use of different deep learning architectures [Ren et al., 2015, Dosovitskiy et al., 2021, Khan et al., 2021]. Despite these advancements, recognising any arbitrary real-world object still remains a daunting challenge as it is unrealistic to label all the existing object classes on the earth. Zero-Shot Learning (ZSL) addresses this problem, requiring images from the *seen* classes during the training, but has the capability of recognising *unseen* classes during the inference [Xian et al., 2019a, Xie et al., 2019, Xu et al., 2020, Federici et al., 2020]. Here the central insight is that all the existing categories share a common semantic space and the task of ZSL is to learn a mapping from the imagery space to the semantic space with the help of side information (attributes, word embeddings) [Xian et al., 2017, Mikolov et al., 2013, Pennington et al., 2014] available with the seen classes during the training phase so that it can be used to predict the class information for the unseen classes during the inference time.

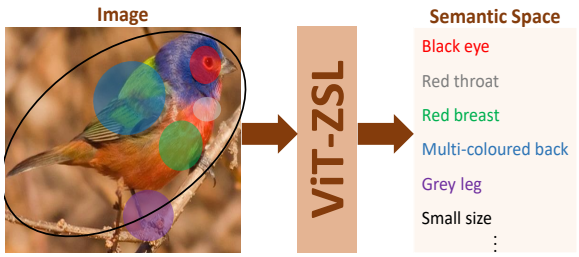


Figure 1: Our method embeds each attribute-based feature with the semantic space. It learns the visual discriminative features through multi-head attention. *Best to view in colour*: colours in the image correspond to the same-colour attribute in the semantic space.

Most of the existing ZSL methods [Xian et al., 2018, Schönfeld et al., 2019] depends on pretrained visual features and necessarily focus on learning a compatibility function between the visual features and semantic attributes. Although modern neural network models encode local visual information and object parts [Xie et al., 2019], they are not sufficient to solve the localisation issue in ZSL models. Some attempts have also

been made by learning visual attention that focuses on some object parts [Zhu et al., 2019]. However, designing a model that can exploit a stronger attention mechanism is relatively unexplored.

Therefore, to alleviate the above shortcomings of visual representations in ZSL models, in this paper, we propose a Vision Transformer (ViT) [Dosovitskiy et al., 2021] based multi-head self-attention model for solving the ZSL task. Our main contribution is to introduce ViT for enhancing the visual feature localisation to solve the zero-shot learning task. Without any object part-level annotation or detection, this is the first attempt to introduce ViT into ZSL. As illustrated in Figure 1, our method maps the visual features of images to the semantic space with the help of scaled dot-product of multi-head attention employed in ViT. We have also performed detailed experimentation on three public datasets (i.e., AWA2, CUB and SUN) following Generalised Zero-Shot Learning (GZSL) setting and achieved very encouraging results on all of them, including the new state-of-the-art harmonic mean on all the datasets.

2 Related Work

Zero-Shot Learning: ZSL is employed to bridge the gap between seen and unseen classes using semantic information, which is done by computing similarity function between visual features and previously learned knowledge [Romera-Paredes and Torr, 2015]. Various approaches address the ZSL problem by learning probabilistic attribute classifiers to predict class labels [Lampert et al., 2009, Norouzi et al., 2014] and by learning linear [Frome et al., 2013, Akata et al., 2015, Akata et al., 2016], and non-linear [Xian et al., 2016] compatibility function associating image features and semantic information. Recently proposed generative models synthesise visual features for the unseen classes [Xian et al., 2018, Schönfeld et al., 2019]. Although those models achieve better performances compared to classical models, they rely on features of trained CNNs. Recently, attention mechanism is adapted in ZSL to integrate discriminative local and global visual features. Among them, S^2GA [Yu et al., 2018] and AREN [Xie et al., 2019] use an attention-based network with two branches to guide the visual features to generate discriminative regions of objects. SGMA [Zhu et al., 2019] also applies attention to jointly learn global and local features from the whole image and multiple discovered object parts. Very recently, APN [Xu et al., 2020] proposes to divide an object into eight groups and learns a set of attribute prototypes, which further help the model to decorrelate the visual features. Partly inspired by the success of attention-based models, in this paper, we propose to learn local and global features using multi-scaled-dot-product self-attention via the Vision Transformer model, which to the best of our knowledge, is the first work on ZSL involving Vision Transformer. In this model, we employ multi-head attention after splitting the image into fixed-size patches so that it can attend to each patch to capture discriminative features among them and generate a compact representation of the entire image.

Vision Transformer: Self-attention-based architectures, especially Transformer [Vaswani et al., 2017] has shown major success for various Natural Language Processing (NLP) [Brown et al., 2020] as well as for Computer Vision tasks [Alamri et al., 2021, Dosovitskiy et al., 2021]; the reader is referred to [Khan et al., 2021] for further reading on Vision Transformer based literature. Specifically, CaiT [Touvron et al., 2021] introduces deeper transformer networks, and Swin Transformer [Liu et al., 2021] proposes a hierarchical Transformer, where the representation is computed using self-attention via shifted windows. In addition, TNT [Han et al., 2021] proposes transformer-backbone method modelling not only the patch-level features but also the pixel-level representations. CrossViT [Chen et al., 2021] shows how dual-branch Transformer combining different sized image patches produce stronger image features. Since the applicability of transformer-based models is growing, we aim to expand and judge its capability for GZSL tasks; to the best of our knowledge, this is still unexplored. Therefore, different from the existing works, we employ ViT to map the visual information to the semantic space, benefiting from the great performance of multi-head self-attention to learn class-level attributes.

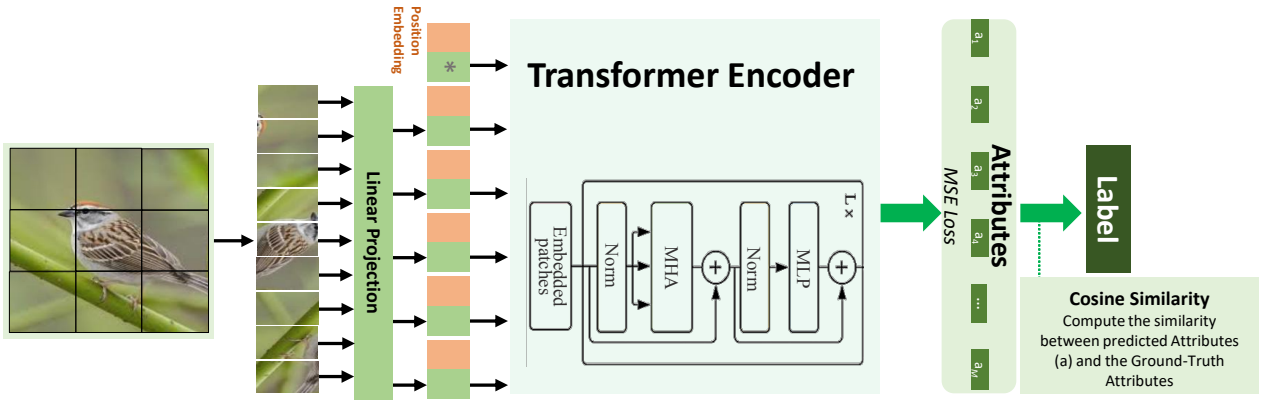


Figure 2: ViT-ZSL Architecture. An image is split into small patches fed into the Transformer encoder after attaching positional embeddings. During the training the output of the encoder is compared with the semantic information of the corresponding image via MSE loss. At inference the encoder output is used to search for the nearest class label.

3 Vision Transformer for Zero-shot Learning (ViT-ZSL)

We follow the inductive approach for training our model, i.e. during training, the model only has access to the images and corresponding image/object attributes from the *seen* classes $\mathcal{S} = \{\mathbf{x}, \mathbf{y} | \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}^s\}$, where $\mathbf{x}$ is an RGB image and $\mathbf{y}$ is the class-level attribute vector annotated with M different attributes, as provided with the dataset. As depicted in Figure 2, a 224×224 image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ with resolution $H \times W$ and C channels is fed into the model. The model follows ViT [Dosovitskiy et al., 2021] as closely as possible; hence the image is divided into a sequence of N patches denoted as $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2.C)}$, where $N = \frac{H.W}{P^2}$. Each patch with a resolution of $P \times P$ is encoded into a patch embedding by a trainable 2D convolution layer (i.e., Conv2d with kernel size=(16, 16) and stride=(16, 16)). Position embeddings are then attached to the patch embeddings to preserve the relative positional information of the order of the sequence due to the lack of recurrence in the Transformer. An extra learnable classification token ($\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$) is appended at the beginning of the sequence to encode the global image representation. Patch embeddings ($\mathbf{z}$) are then projected thought a linear projection $\mathbf{E}$ to D dimension (i.e., $D = 1024$) as in Eq. 1. Embeddings are then passed to the Transformer encoder, which consists of Multi-Head Attention (MHA) (Eq. 2) and MLP blocks (Eq. 3). Before every block, a layer normalisation (Norm) is employed, and residual connections are also applied after every block. Image representation ($\hat{\mathbf{y}}$) is produced as in Eq. 4.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \mathbf{x}_p^3\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2.C) \times D}, \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MHA}(\text{Norm}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \ (L = 24) \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{Norm}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\hat{\mathbf{y}} = \text{Norm}(\mathbf{z}_L^0) \tag{4}$$

In terms of MHA, self-attention is performed for every patch in the sequence of the patch embeddings independently; thus, attention works simultaneously for all the patches, leading to multi-head self-attention. Three vectors, namely Query (Q), Key (K) and Value (V), are created by multiplying the encoder's input (i.e., patch embeddings) by three weight matrices (i.e., W^Q, W^K and W^V) trained during the training process to compute the self-attention. The Q and K vectors undergo a dot-product to output a scoring matrix representing how much a patch embedding has to attend to every other embedding; the higher the score is, the more attention is considered. The score matrix is then scaled down and passed into a softmax to convert the scores into probabilities, which are then multiplied by the V vectors, as in Eq. 5, where d_k is the dimension of the K vectors. Since the multi-attention mechanism is employed, self-attention matrices are then concatenated and

fed into a linear layer and passed to the regression head.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

We argue that self-attention allows our model to attend to image regions that can be semantically relevant for classification and learns the visual features across the entire image. Since the standard ViT has one classification head implemented by an MLP, it has been edited to meet our model objective: to predict M number of attributes (i.e., depending on the datasets used). The motivation behind this is that the network is assumed to learn the notion of classes to predict attributes. For the objective function, we employed the Mean Squared Error (MSE) loss, as the continuous attributes are used as in Eq. 6, where $\mathbf{y}_i$ is the observed attributes, and $\hat{\mathbf{y}}_i$ is the predicted ones.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \tag{6}$$

During testing, instead of applying the extensively used dot product as in [Xu et al., 2020], we consider the cosine similarity as in [Gidaris and Komodakis, 2018] to predict class labels. The cosine similarity between the predicted attributes and every class embedding is measured. The output of the similarity measure is then used to determine the class label of the test images.

4 Experiments

Implementation Details: All images used in training and testing are adapted from the ZSL datasets mentioned below and sized 224×224 without any data augmentation. We employ the Large variant of ViT (ViT-L) [Dosovitskiy et al., 2021], with input patch size 16×16, 1024 hidden dimension, 24 layers, 16 heads on each layer, and 24 series encoder. There are 307M parameters in total in this architecture. ViT-L is then fine-tuned using Adam optimiser with a fixed learning rate of 10^{-4} and a batch size of 64. All methods are implemented in PyTorch[1] on an NVIDIA RTX 3090 GPU, Xeon processor, and a memory sized 32GB.

Datasets: We have conducted our experiments on three popular ZSL datasets: AWA2, CUB, and SUN, whose details are presented in Table 1. The main aim of this experimentation is to validate our proposed method, ViT-ZSL, demonstrate its effectiveness and compare it with the existing state-of-the-arts. Among these datasets, AWA2 [Xian et al., 2017] consists of $37,322$ images of 50 categories (40 seen + 10 unseen). Each category contains 85 binary as well as continuous class attributes. CUB [Wah et al., 2011] contains $11,788$ images forming 200 different types of birds, among them 150 classes are considered as seen, and the other 50 as unseen, which is split by [Akata et al., 2016]. Together with images CUB dataset also contains 312 attributes describing birds. Finally, SUN [Patterson and Hays, 2012] has the largest number of classes among others. It consists of 717 types of scene, divided into 645 seen and 72 unseen classes. The SUN dataset contains $14,340$ images with 102 annotated attributes.

Table 1: Dataset statistics in terms of granularity, number of classes (seen + unseen classes) as shown within parenthesis, number of attributes and number of images.

Datasets	Granularity	# Classes (S + U)	# Attributes	# Images
AWA2 [Xian et al., 2017]	coarse	50 (40 + 10)	85	37,322
CUB [Wah et al., 2011]	fine	200 (150 + 50)	102	11,788
SUN [Patterson and Hays, 2012]	fine	717 (645 + 72)	312	14,340

Evaluation: In this work, we train our ViT-ZSL model following the inductive approach [Wang et al., 2019]. Following [Xian et al., 2019a], we measure the top-1 accuracy for both seen as well as unseen classes. To

[1]Our code is available at: `https://github.com/FaisalAlamri0/ViT-ZSL`

capture the trade-off between both sets of classes performance, we use the harmonic mean, which is the primary evaluation criterion for our model. Following the recent papers (e.g., [Xu et al., 2020], [Chao et al., 2016]), we apply Calibrated Stacking [Chao et al., 2016] to evaluate the considered methods under GZSL setting, where the calibration factor γ is dataset dependant and decided based on a validation set.

Quantitative Results: We consider the AWA2, CUB and SUN datasets to show the performance of our proposed model and compare the performance with related arts. Table 2 shows the quantitative comparison between the proposed model and various other GZSL models. The performance of each model is shown in terms of Seen (S) and Unseen (U) classes and their harmonic mean (H).

Table 2: Generalised zero-shot classification performance on AWA2, CUB and SUN

Models	AWA2			CUB			SUN		
	S	U	H	S	U	H	S	U	H
DAP [Lampert et al., 2009]	84.7	0.0	0.0	67.9	1.7	3.3	25.1	4.2	7.2
IAP [Lampert et al., 2009]	87.6	0.9	1.8	72.8	0.2	0.4	37.8	1.0	1.8
DeViSE [Frome et al., 2013]	74.7	17.1	27.8	53.0	23.8	32.8	30.5	14.7	19.8
ConSE [Norouzi et al., 2014]	90.6	0.5	1.0	72.2	1.6	3.1	39.9	6.8	11.6
SSE [Zhang and Saligrama, 2015]	82.5	8.1	14.8	46.9	8.5	14.4	36.4	2.1	4.0
SJE [Akata et al., 2015]	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
ESZSL [Romera-Paredes and Torr, 2015]	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8
LATEM [Xian et al., 2016]	77.3	11.5	20.0	57.3	15.2	24.0	28.8	14.7	19.5
ALE [Akata et al., 2016]	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3
SAE [Kodirov et al., 2017]	82.2	1.1	2.2	54.0	7.8	13.6	18.0	8.8	11.8
AREN [Xie et al., 2019]	92.9	15.6	26.7	78.7	38.9	52.1	38.8	19.0	25.5
SGMA [Zhu et al., 2019]	87.1	37.6	52.5	71.3	36.7	48.5	-	-	-
APN [Xu et al., 2020]	78.0	56.5	65.5	69.3	65.3	67.2	34.0	41.1	37.6
*GAZSL [Zhu et al., 2018]	86.5	19.2	31.4	60.6	23.9	34.3	34.5	21.7	26.7
*f-CLSWGAN [Xian et al., 2018]	64.4	57.9	59.6	57.7	43.7	49.7	36.6	42.6	39.4
Our model (ViT-ZSL)	**90.0**	**51.9**	**65.8**	**75.2**	**67.3**	**71.0**	**55.3**	**44.5**	**49.3**

S, U, H denote Seen classes ($\mathcal{Y}^s$), Unseen classes ($\mathcal{Y}^u$), and the Harmonic mean, respectively. For each scenario, the best is in red and the second-best is in blue. * indicates generative representation learning methods.

DAP and IAP [Lampert et al., 2009] are some of the earliest works in ZSL, which perform poorly compared to other models. This is due to the assumptions claimed in these approaches regarding attributes dependency. In real-world animals with attributes 'terrestrial' and 'farm' are dependent but are assumed independent by such models, which are noted as incorrect by [Akata et al., 2016]. Our model ViT-ZSL does not assume this, but rather it considers the correlation between attributes, which self-attention helps to achieve by considering both positional and contextual information of the entire sequence of patches. DeViSE [Frome et al., 2013] and ConSE [Norouzi et al., 2014] learn a linear mapping between images and their semantic embedding space. They both make use of the same text model trained on 5.4B words from Wikipedia to construct 500-dimensional word embedding vectors. Both use the same baseline model, but DeViSE replaces the last layer (i.e., softmax layer) with a linear transformation layer. In contrast, ConSE keeps it and computes the predictions via a convex combination of the class label embedding vectors. ConSE, as presented in Table 2 outperforms DeViSE, but DeViSE is generally performing better on the unseen classes. Similarly, SJE [Akata et al., 2015] learns a bilinear compatibility function using the structural SVM objective function to maximise the compatibility between image and class embeddings. ESZSL [Romera-Paredes and Torr, 2015] uses the square loss to learn bilinear compatibility. Although ESZSL is claimed to be easy to implement, its performance, in particular for GZSL, is poor. ALE [Akata et al., 2016], which belongs to the bilinear compatibility approach group, performs better than most of its group member. LATEM [Xian et al., 2016], instead of learning a single bilinear map, extends the bilinear compatibility of SJE [Akata et al., 2015] as to be an image-class pairwise linear by learning multiple linear mappings. It performs better than SJE on unseen classes but with a lower harmonic

Figure 3: Representative examples of attention. First row: Original images, Middle: Attention maps, and last: Attention fusions. From left to right, ViT-ZSL is able to focus on object-level attributes and learn objects discriminative features when objects are partly captured (first three columns images), occluded (fourth column images) or fully presented (last two columns images).

mean due to its poor performance on seen classes. Generative ZSL models such as GAZSL [Zhu et al., 2018], and f-CLSWGAN [Xian et al., 2018] are seen to reduce the effect of the bias problem due to the inclusion of synthesised features for the unseen classes. However, this does not apply to our method, as no synthesised features are used in our case; instead, solely the features extracted from seen classes are used during training. AREN [Xie et al., 2019], SGMA [Zhu et al., 2019] and APN [Xu et al., 2020] are non-generative ZSL models focusing on object region localisation using image attention. They are the most relevant works to ours as attention mechanism is included in these models architecture. However, they consist of two branches in their models, where the first learns local discriminative visual features and the second captures the image's global context. In contrast, our model uses only one compact network, where the input is the image patches so that the global and local discriminative features can be learned using the multi-head self-attention mechanism.

Our model ViT-ZSL, as shown in Table 2, achieves the best harmonic mean on AWA2. It also performs as the third best on both seen and unseen classes. Compared with the other models, it scores 90.02%, where the highest is the highest is AREN with 92.9% accuracy. As the comparison illustrated follows the GZSL setting using the harmonic mean as the primary evaluation metric for GZSL models, ViT-ZSL outperforms all state-of-the-art models. In terms of the CUB dataset, our method achieves the second-highest accuracy for seen classes, but the highest for unseen. In addition, our ViT-ZSL obtains the best harmonic mean score among all the reported approaches. On SUN datasets, which has the most significant number of object classes among other datasets, our model performs as the best for both seen and unseen classes, achieving a harmonic mean of 47.9%, the highest compared to all other models.

Attention Maps: In Figure 3, we show how our model attends to image regions semantically relevant to the object class. For example, in the images of the first three columns, the entire objects' shapes are absent (i.e., only the top part is captured), and in the image in the fourth column, the groove-billed ani bird is impeded by a human hand. Although these images suffer from occlusion, our model accurately attends to the objects in the image. Thus, we believe that ViT-ZSL definitely benefits from the attention mechanism, which is also involved in the human recognition system. Clearly, we can say that our method has learned to map the relevance of local regions to representations in the semantic space, where it makes predictions on the visible attribute-based regions. Similarly, in the last two columns images of Figure 3, it can be noticed how the model pays more attention to some object-level attributes (i.e., *Deer*: forest, agility, furry etc., and *Vermilion Flycatcher*: solid and red breast, perching-like shape, notched tail). It can also be noticed that the model focuses on the context of the object, as in the second column images. This can be due to the guidance of some attributes (i.e., forest, jungle, ground and tree) which are associated with *leopard* class. However, as shown in the first column, the model did not pay much attention to the bird's beak compared to the head and the rest of the body, which

needs to be investigated further and building an explainable model as in [Xian et al., 2019b] could be a way to accomplish this.

5 Conclusion

In this paper, we proposed a Vision Transformer-based Zero-Shot Learning (ViT-ZSL) model that specifically exploits the multi-head self-attention mechanism for relating visual and semantic attributes. Our qualitative results showed that the attention mechanism involved in our model focuses on the most relevant image regions related to the object class to predict the semantic information, which is used to find out the class label during inference. Our results on the GZSL task, including the highest harmonic mean scores on the AWA2, CUB and SUN datasets, illustrate the effectiveness of our proposed method.

Although our method achieves very encouraging results for the GZSL task on three publicly available benchmarks, the bias problem towards seen classes remains a challenge, which will be addressed in future work. Training the model in a transductive setting, where visual information for unseen classes could be included, is a direction to be examined.

Acknowledgement

This work was supported by the Defence Science and Technology Laboratory and the Alan Turing Institute. The TITAN Xp and TITAN V used for this research were donated by the NVIDIA Corporation.

References

[Akata et al., 2016] Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2016). Label-embedding for image classification. *IEEE TPAMI*.

[Akata et al., 2015] Akata, Z., Reed, S. E., Walter, D., Lee, H., and Schiele, B. (2015). Evaluation of output embeddings for fine-grained image classification. In *CVPR*.

[Alamri et al., 2021] Alamri, F., Kalkan, S., and Pugeault, N. (2021). Transformer-encoder detector module: Using context to improve robustness to adversarial attacks on object detection. In *ICPR*.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *NeurIPS*.

[Chao et al., 2016] Chao, W.-L., Changpinyo, S., Gong, B., and Sha, F. (2016). An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*.

[Chen et al., 2021] Chen, C.-F., Fan, Q., and Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv*.

[Dosovitskiy et al., 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[Federici et al., 2020] Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. (2020). Learning Robust Representations via Multi-View Information Bottleneck. In *ICLR*.

[Frome et al., 2013] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *NIPS*.

[Gidaris and Komodakis, 2018] Gidaris, S. and Komodakis, N. (2018). Dynamic few-shot visual learning without forgetting. In *CVPR*.

[Han et al., 2021] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. *arXiv*.

[Khan et al., 2021] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F., and Shah, M. (2021). Transformers in vision: A survey. *arXiv*.

[Kodirov et al., 2017] Kodirov, E., Xiang, T., and Gong, S. (2017). Semantic autoencoder for zero-shot learning. In *CVPR*.

[Lampert et al., 2009] Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv*.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*.

[Norouzi et al., 2014] Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G., and Dean, J. (2014). Zero-shot learning by convex combination of semantic embeddings. In *ICLR*.

[Patterson and Hays, 2012] Patterson, G. and Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*.

[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 TPAMI*.

[Romera-Paredes and Torr, 2015] Romera-Paredes, B. and Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *ICML*.

[Schönfeld et al., 2019] Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., and Akata, Z. (2019). Generalized zero- and few-shot learning via aligned variational autoencoders. *CVPR*.

[Touvron et al., 2021] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. (2021). Going deeper with image transformers. *arXiv*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.

[Wah et al., 2011] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology.

[Wang et al., 2019] Wang, W., Zheng, V., Yu, H., and Miao, C. (2019). A survey of zero-shot learning. *ACM-TIST*.

[Xian et al., 2016] Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., and Schiele, B. (2016). Latent embeddings for zero-shot classification. In *CVPR*.

[Xian et al., 2019a] Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2019a). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*.

[Xian et al., 2018] Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. (2018). Feature generating networks for zero-shot learning. In *CVPR*.

[Xian et al., 2017] Xian, Y., Schiele, B., and Akata, Z. (2017). Zero-shot learning - the good, the bad and the ugly. In *CVPR*.

[Xian et al., 2019b] Xian, Y., Sharma, S., Schiele, B., and Akata, Z. (2019b). F-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*.

[Xie et al., 2019] Xie, G.-S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., and Shao, L. (2019). Attentive region embedding network for zero-shot learning. In *CVPR*.

[Xu et al., 2020] Xu, W., Xian, Y., Wang, J., Schiele, B., and Akata, Z. (2020). Attribute prototype network for zero-shot learning. In *NIPS*.

[Yu et al., 2018] Yu, y., Ji, Z., Fu, Y., Guo, J., Pang, Y., and Zhang, Z. M. (2018). Stacked semantics-guided attention model for fine-grained zero-shot learning. In *NeurIPS*.

[Zhang and Saligrama, 2015] Zhang, Z. and Saligrama, V. (2015). Zero-shot learning via semantic similarity embedding. In *ICCV*.

[Zhu et al., 2018] Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., and Elgammal, A. (2018). Imagine it for me: Generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*.

[Zhu et al., 2019] Zhu, Y., Xie, J., Tang, Z., Peng, X., and Elgammal, A. (2019). Semantic-guided multi-attention localization for zero-shot learning. In *NIPS*.

A Study of Image and Video Reconstruction Applications of a Novel Frequency-Domain Loss Function

Ojasvi Yadav*, Sebastian Lutz+, Koustav Ghosal+, Aljosa Smolic+

*Dukaan, India. Previously with V-SENSE, Trinity College Dublin, Ireland.
+V-SENSE, Trinity College Dublin, Ireland.

Abstract

Image and video reconstruction are well-researched problems in computer vision. Several problems involving reconstruction are actively researched areas in computer vision such as image denoising and deblurring, super resolution, inpainting etc. In recent years, the state-of-the-art (SoA) in these areas has come from deep learning methods, which train neural networks for the specific task, often in a supervised manner. Naturally, the choice of the loss function in these algorithms is important. In this paper, we study a general-purpose loss function that can be used for several SoA image or video reconstruction method for performance enhancements. We show that the addition of the novel loss function during training improves the performance of SoA algorithms in five different image and video reconstruction tasks.

Keywords: Image and Video Reconstruction, Loss Function, Frequency Domain, Fast Fourier Transform.

1 Introduction

The non-linear nature of image and video reconstruction makes it a difficult task. In other words, the desired output may not linearly map to the input due to the presence of distortions such as noise, incorrect exposure, inconsistent colours, poor resolution, etc. Recently, Convolutional Neural Networks (CNNs) have proven quite effective in several reconstruction tasks such as image deblurring [Kupyn et al., 2019], super-resolution [Dong et al., 2015], image inpainting [Xie et al., 2019], exposure correction [Yadav et al., 2021], denoising [Lehtinen et al., 2018] etc. CNNs by design are capable of learning complex, non-linear mappings between different data domains. But despite the superior performance, they suffer from most, if not all of the distortions mentioned above. In a recent work, [Yadav et al., 2021], proposed a generic loss function by exploiting the frequency domain and showed improvements in image exposure correction in several of these aspects. However, the capacity of their approach in other reconstruction tasks is still not explored and therefore remains an open problem. In this paper, we extend their work by studying a wide variety of image and video reconstruction tasks namely deblurring, super-resolution, inpainting, and denoising. In the rest of this section, we briefly highlight some of the task specific challenges and the intuition behind addressing these challenges in the frequency domain.

In image deblurring, the aim is to remove blur from images which can be caused by many factors such as unsuitable focal length or aperture size, low light, low frame rate and fast motion, camera shake, or lack of focus. Blur can also be of different types such as motion blur, gaussian blur, bokeh blur, zoom blur, etc. These different blur types make this problem ill-posed and challenging, which has been addressed by many deep learning and classical solutions. In the fast Fourier transform (FFT), high and low frequency components present away from and close to the center, respectively. Image blur generally results in a reduction of high-frequency components of the image. Thus, the frequency coefficients of a blurred image are dominantly skewed towards low-frequency coefficients as shown in figure 1.

In super-resolution, the goal is to upscale the image to a higher resolution while preserving the high-frequency components. It has several applications such as displaying low-resolution pictures on giant screens,

person re-identification from low-resolution surveillance cameras, image classification [Hao et al., 2018], edge detection [Dai et al., 2016] etc. Image interpolation or upsampling is a basic form of super-resolution. It increases the resolution of the image by averaging the nearby pixels to approximate the newly added pixels. The averaging operation introduces blur to the image and the loss of sharp edges and other high-frequency components, as shown in figure 1. Recent deep learning-based methods for super-resolution can learn to accurately predict high-frequency information by performing an inverse non-linear mapping.

In image-inpainting, the goal is to fill in the pixels of an image that are missing or occluded such as occlusions in the form of subtitles in videos or by removing unwanted objects from the image. It is a difficult problem to solve due to the ill-posed nature of filling in the missing/occluded pixels: For an 8-bit image, there are 2^8 possible values for a missing pixel. However, prediction of the missing pixels can be assisted by CNN based networks which take advantage of the spatial correlations that are characteristic in most natural images. To fill in an occluded region in an image, the frequency domain can help fill in pixels such that the filled-in region matches the frequency characteristics of the surrounding pixels. Small occlusions in an image can lead to dramatic changes in its frequency domain representation, as shown in figure 1.

In image and video denoising, the aim is to separate noise from the original data. There are a variety of causes for noise. It can be caused by the camera heat generated during the camera's operation or it can be external due to low light i.e. transmission noise. Each type of noise has a particular characteristic: it can be random, or it can be related to the underlying image. For example, salt and pepper noise is independent of the underlying image, while multiplicative noise is dependent on the underlying image. Such diversity makes denoising a cumbersome problem to solve. Often, noise is better represented in the frequency domain because its frequency components can be quite distinct from the frequency components in the original image, as shown in figure 1. The FFT of the noisy image in figure 1 is more crowded than the FFT of the original image. Deep learning-based approaches work well for image denoising due to their ability to understand spatial relationships in an image [LeCun et al., 1995]. This ability allows deep learning-based models to accurately estimate the correct pixel values in noisy images.

The goal of this work is to use a loss function that can improve existing SoA image and video reconstruction approaches in deep learning. Loss functions quantify how different the model's output is from the ground truth. This value is used during the back-propagation of the model training phase. Back-propagation is then used for re-tuning the weights of the model so that the loss function gets minimized. To minimize a loss function at a faster rate, its rate of change should be higher than that of other loss functions at the beginning of the training phase. This can be achieved by using a loss function that measures the distance in both the image and the frequency domain. This increases the room for improvement by presenting an error from diverse representations.

In general, the frequency domain is often used in the computer vision community for image processing and digital signal processing. But, to the best of our knowledge, barring [Yadav et al., 2021], it has not been used for general-purpose loss functions for image and video reconstruction tasks. Frequency domain transforms have characteristics that make them suitable for use in image and video reconstruction. We show this in figure 1. The common deformations in the images lead to dramatic changes in the frequency domain. We theorize that compensating for these dramatic changes in the frequency domain may improve the existing image and video reconstruction tasks. The comparison in figure 1 shows that the frequency domain provides value in locating frequencies that would be hard to locate in the image domain.

There are many frequency domain transforms available to build such a loss function, Namely, DCT, FFT, DFT, etc. DCT is used for image compression because it can condense an image's frequency components into a very small space. This is unfortunately the opposite of what a loss function should ideally do. When comparing DFT to DCT, frequency domain components are spread out over a larger area for DFT. This makes it more suitable to be used in a loss function. The time complexity required for DFT is of the order of N^2 where N is the number of pixels in an image. Next, FFT is another frequency domain transform that can be used to create a frequency domain loss function. It is a faster implementation of the DFT which also works for continuous signals, making it more general-purpose than DFT. Time complexity to compute the FFT of an image is equal to $N \times \log N$. This makes it an ideal choice to be used as a loss function for image reconstruction tasks. It offers

similar boosts in accuracy as [Jiang et al., 2020], and is more general-purpose than [Jiang et al., 2020] due to its compatibility with continuous signals, and is computationally more efficient.

Our contribution in this study is that we thoroughly investigate the capacity of the frequency loss function proposed by [Yadav et al., 2021] in each of the aforementioned domains — image deblurring, image super-resolution, image inpainting, and image and video denoising and subsequently, notice significant improvements. For each application, we choose the SoA method, retrain it using the frequency loss and compare the performance both quantitatively and qualitatively.

2 Related Works

Loss Functions: In the literature, several loss functions are available for model training. We divide these loss functions into two categories: general purpose loss functions and specific purpose loss functions. General purpose loss functions such as L1 loss and L2 loss are commonly used in image and video reconstruction tasks. Specific loss functions are developed by researchers to address a particular problem. They are usually not general-purpose or are not tested for a wide variety of problems. L1 loss function is the summed absolute difference between the output and the ground truth. L2 loss function is similar to the L1 loss function. However, the difference between the output and the ground truth gets squared in L2 loss. Other general-purpose loss functions have been shown to give favourable results but are not adopted as an industry standard yet.

[Mechrez et al., 2018] present a loss function that does not require paired datasets. Instead of having to compare the output with its aligned ground truth, their loss function can deal with unaligned image pairs. To do this, they find contextual similarity between features, which can be present in different areas of the image. They apply this loss function in four SoA models for style transfer [Gatys et al., 2016], single-image animation [Johnson et al., 2016], puppet control [Isola et al., 2017], and domain transfer [Zhu et al., 2017]. Since these problems deal with unaligned datasets, they were unable to provide quantitative comparisons with the SoA models. However, they did present comparable or improved qualitative results. In this paper, we try to show how the novel loss function introduced by [Yadav et al., 2021] can improve existing SoA approaches quantitatively.

Image and Video Reconstruction: Due to the plethora of work that exists in this area, an exhaustive discussion is beyond our scope. We mention the baselines compared with and other papers directly related to this work. A prominent SoA approach for deblurring is DeblurGAN-v2 [Kupyn et al., 2019] which improves the efficiency of the entire working of its predecessor, DeblurGAN [Kupyn et al., 2018]. For super-resolution, we build on SRCNN [Dong et al., 2015] that uses a L2 loss. Newer approaches such as [Ledig et al., 2017, Wang et al., 2018] use GAN based loss functions to achieve impressive results. In image inpainting, the goal is to predict the obscured or missing areas of an image. [Xie et al., 2019] present an approach by combining style, perceptual and L1 loss. Other recent important approaches include [Yu et al., 2019, Yu et al., 2018]. For image denoising approach, the SoA is Noise2Noise [Lehtinen et al., 2018], which performs denoising using a U-net architecture [Ronneberger et al., 2015] and L2 loss. A recent comprehensive survey of recent approaches in image denoising using deep learning can be found in [Tian et al., 2020]. A recent SoA approach in video denoising is VideNN [Claus and van Gemert, 2019]. Their model is based on two types of networks: spatial denoising and temporal denoising. Spatial denoising is built upon the architecture provided by [Zhang et al., 2017] which is a SoA CNN architecture for gaussian denoising.

3 Proposed Approach

To devise a loss function based on a mathematical function such as FFT, we need to ensure that the loss function stays differentiable. FFT outputs a complex signal along with a real signal. This output denotes the sines and cosine waves of all the frequency present in the image. Unfortunately, this output is not reliably differentiable due to the presence of the complex domain signal. To make it differentiable, we compute the magnitude of

the FFT and convert it into a purely real signal by computing its magnitude. After creating a differentiable frequency domain transform, a loss function should also compute the distance between the model output and the ground truth. To do this, we use either L1 or L2 distance between the two as shown in 1, where y_{GT} and y_O represent ground-truth and prediction, respectively.

$$L_1 = \sum_{i=1}^{n} |y_{GT} - y_O|, \quad L_2 = \sum_{i=1}^{n} (y_{GT} - y_O)^2 \tag{1}$$

For an output image I_1, whose ground truth is I_2, the F-loss is defined in equation 2. Here, K is a scaling factor, it is the scale at which the loss is being computed. For example, to calculate the loss at half the resolution, K is set to 2, for calculating loss at quarter the resolution, K is set to 4. $FFT(I)$ refers to the fast Fourier transform of image I. M and N are the number of pixels horizontally and vertically of the input images. We compute this loss at different scales (K) of the images and obtain the loss function (F-loss) shown in equation 3.

$$L_{FFT}^{\frac{M}{K} \times \frac{N}{K}} = \frac{K^2}{M \times N} \Big| ||FFT(I_1)|| - ||FFT(I_2)|| \Big|^{\frac{M}{K} \times \frac{N}{K}} \tag{2}$$

$$FFT_{Final}(I_1, I_2) = L_{FFT}^{\frac{M}{1} \times \frac{N}{1}} + L_{FFT}^{\frac{M}{2} \times \frac{N}{2}} + L_{FFT}^{\frac{M}{4} \times \frac{N}{4}} \tag{3}$$

Our approach in all the experiments in the following section is to first train and test the chosen models in their default settings and do it again but with an additional F-loss term (equation 3) in the original loss function. We then assess the causal results of adding the F-loss in the loss function of each approach.

4 Experiments and Results

To assess the effectiveness of F-loss, we experiment on six SoA deep learning-based approaches for image and video reconstruction. To prove a causal result of the loss function, we only change the loss function and keep all other settings unchanged from the original work. Depending on how the SoA approaches were coded, we accordingly used the PyTorch or the Tensorflow version of the frequency loss function.

4.1 Deblurring

Our framework for deblurring is based on DeblurGAN-v2 [Kupyn et al., 2019]. The loss function used in this approach is described in equation 4. The overall loss, L, depends on three other losses. L_p is the pixel-wise loss, which, in this case, is the L_2 loss. This loss function works reasonably well in most general applications but it tends to yield over-smoothened images [Ledig et al., 2017]. To counter this over-smoothening, the content loss L_X is used which compares the output's and ground truth's high-level features calculated from pre-trained CNNs. Finally, L_{adv} is used to stabilize the training process of the GAN network by computing loss based on the generator and the discriminator networks' outputs. The difference in the FFT representations is more obvious than the difference in the spatial-image representations.

We hypothesize that F-loss, due to its operation in the frequency domain, will aid the network [Kupyn et al., 2019] in recovering the lost high-frequency components more effectively. To test this, we first trained the given model from scratch using the default hyper-parameters and training settings. The GoPro dataset [Nah et al., 2017] was used for this, as this was the same dataset that was used in the original approach. We then modified the equation 4 to include the F-loss, as shown in equation 5. We then repeated the same training procedure once again from scratch. In equation 5, L_F is the F-loss, and α is the factor value that was set as 0.5 by a trial and error process. The testing was performed on the training-testing split of the GOPRO dataset as provided by the creators. The quantitative results of this experiment are shown in table 1. The addition of F-loss improves the results in both the PSNR and SSIM metrics.

$$L = 0.5 * L_p + 0.006 * L_X + 0.01 * L_{\text{adv}} \tag{4}$$

$$L = 0.5 * L_p + 0.006 * L_X + 0.01 * L_{\text{adv}} + \alpha * L_F \tag{5}$$

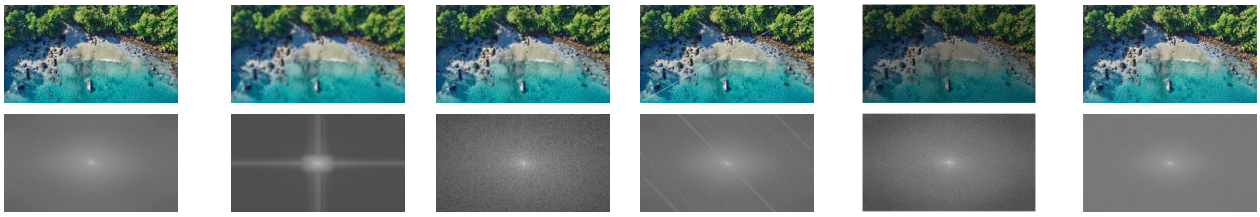


Figure 1: First row consists of the natural image and its deformed versions. Second row is the FFT of the above row. The difference between the FFTs of the natural image is more evident than the difference between the natural images themselves.

Task	Approach	PSNR	SSIM
Deblurring	DBGANv2[Kupyn et al., 2019]	29.18	0.89
Deblurring	**DBGANv2-F-loss**	**29.49**	**0.91**
Super-Resolution	SRCNN[Dong et al., 2015]	28.86	0.92
Super-Resolution	**SRCNN-F-loss**	**29.10**	**0.94**
Image-Inpainting	LBAM[Xie et al., 2019]	26.11	0.86
Image-Inpainting	**LBAM-F-loss**	**26.39**	**0.87**
Exposure correction	Chen[Chen et al., 2018]	28.60	0.767
Exposure correction	**Chen-F-loss**	**28.89**	**0.776**
Image denoising	Gaussian-Clean[Lehtinen et al., 2018])	30.30	0.87
Image denoising	**Gaussian-Clean-F-loss**	**30.80**	**0.89**
Video denoising	VideNN[Claus and van Gemert, 2019]	31.5	-
Video denoising	**VideNN-F-loss**	**32.48**	-

Table 1: Quantitative results on SoA approaches before and after adding F-loss to the loss function. This proves a quantitative causal improvement upon the addition of F-loss for a wide variety of image and video reconstruction tasks.

4.2 Super-resolution

We base our framework for super-resolution on SRCNN [Dong et al., 2015]. To perform this experiment, we first trained the SRCNN from scratch, keeping all the hyper-parameters the same as they were defined by the authors. The dataset used for this training was the BSD200 dataset [Martin et al., 2001]. We then added the F-loss to the L2 loss function. This new loss function is described in the equation 6, where L_F is the computed value of the F-loss between the output high-resolution image and the ground truth high-resolution image and α was set to 1. With this new loss function, we repeated the same training procedure from scratch. The evaluation dataset for this was the set5 dataset [Bevilacqua et al., 2012]. The quantitative results of training [Dong et al., 2015] model on set5, with and without F-loss, are shown in table 1. The addition of F-loss improves the results in both the PSNR and SSIM metrics.

$$L = \sum_{i=1}^{n} (y_{GT} - y_O)^2 + \alpha * L_F \tag{6}$$

4.3 Image Inpainting

The framework for this task is based on [Xie et al., 2019]. The LBAM approach employs a variety of loss functions, their collective formula is shown in equation 7. Here, λ_1, λ_2, λ_3, λ_4 are empirically set as $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$, $\lambda_4 = 120$. L_{l_1} is the L1 loss, which the authors call "Pixel Reconstruction Loss". L_{perc} is the perceptual loss, which assesses the high level information of the output. The L_{perc} is defined on the VGG-16 network [Simonyan and Zisserman, 2014], which is pre-trained on ImageNet dataset [Deng et al., 2009]. L_{style}

is the "style Loss" which they claim recovers detailed textures better. L_{adv} is the widely used adversarial loss [Goodfellow et al., 2014] which is computed by the discriminator network.

The LBAM model was trained from scratch using the default parameters provided by the authors. The dataset used for this experiment was the Paris StreetView dataset [Philbin et al., 2008]. This is the same dataset that was used in the original LBAM paper [Xie et al., 2019]. Next, we repeated the same training procedure but used the loss function described in equation 8. In equation 8, we add the F-loss in the original loss function (equation 7). The F-loss is controlled via the α factor, which is set empirically to 1.05. The training and testing split was the same as that used by the authors of LBAM. The quantitative results of this experiment are shown in table 1. The addition of F-loss improves the results in both the PSNR and SSIM metrics.

$$L = \lambda_1 L_{l_1} + \lambda_2 L_{adv} + \lambda_3 L_{perc} + \lambda_4 L_{style} \tag{7}$$

$$L = \lambda_1 L_{l_1} + \lambda_2 L_{adv} + \lambda_3 L_{perc} + \lambda_4 L_{style} + \alpha L_F \tag{8}$$

4.4 Exposure Correction

We would like to point the readers to this paper's parent work [Yadav et al., 2021]. That publication is focused entirely on the quantitative and qualitative improvements in an SoA exposure correction model [Chen et al., 2018] after the addition of F-loss (table 1).

4.5 Image Denoising

We choose the framework by [Lehtinen et al., 2018] as the starting point for image denoising. We first trained the given network from scratch using the default settings as provided by the authors. We trained this on the Set14 dataset [Zeyde et al., 2010]. Next, we repeated the prior process using the same model and the same dataset, but this time adding the F-loss as described in equation 9 to the original loss function. The α was set to 1. The quantitative results of this experiment are shown in table 1. The addition of F-loss improves the results in both the PSNR and SSIM metrics.

$$L = L_2 + \alpha * L_F \tag{9}$$

4.6 Video Denoising

For video denoising we start from VideNN [Claus and van Gemert, 2019]. Their model is based on two types of networks: spatial denoising and temporal denoising. Spatial denoising is built upon the architecture provided by [Zhang et al., 2017] which is a SoA CNN architecture for gaussian denoising. However, this architecture also works well for realistic signal-dependent noise. Three consecutive frames are the input for three of such spatial denoising models. For the output, these spatially denoised three frames are then fed to the temporal denoising network. The temporal denoising network then denoises the middle frame of the three frames that it receives. Both the spatial and temporal networks estimate the noise in the frames and then subtract it from the noisy input.

The L2 loss is used for both temporal and spatial denoising networks. In the previous experiment, we showed how an image-denoising network can benefit from the addition of the F-loss. Similarly, we hypothesize that the addition of the F-loss in a video-denoising network will also assist it in getting better quantitative results. To test this, we first trained the given network from scratch using the default settings as provided by the authors. We trained this on the Waterloo Exploration dataset [Ma et al., 2017]. Next, we repeated the prior process using the same model and the same dataset, but this time adding the F-loss as described in equation 10. The α was set to 1. The quantitative results of this experiment are shown in table 1. The addition of F-loss improves the results for the PSNR metric.

$$L = \sum_{i=1}^{n} (y_{groundTruth} - y_{output})^2 + \alpha * L_F \tag{10}$$

5 Conclusion, Limitations and Future Work

In this paper, we explored the applications of a novel loss function for image and video reconstruction. We show that the addition of this loss function to six SoA approaches helps them in surpassing their results. One limitations of this loss function is that the factor value α needs to be manually fine-tuned. We also hope to apply this loss function to more deep learning SoA video reconstruction approaches. In future work, we can also improve results in applications such as deblurring and super-resolution with an adjustment. The disparity between the FFT of ground truth and output mainly arises for the high-frequency components. Therefore, a desirable modification could be to create a dynamic weighted F-loss function for deblurring and super-resolution that adjusts its weightage of the high or low-frequency components. If it consistently sees a higher disparity in the high or low-frequency regions, then it assigns a higher weight to those regions. This will enable a faster convergence, and further reduce the training times of the respective models.

References

[Bevilacqua et al., 2012] Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *BMVC2012*.

[Chen et al., 2018] Chen, C., Chen, Q., Xu, J., and Koltun, V. (2018). Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300.

[Claus and van Gemert, 2019] Claus, M. and van Gemert, J. (2019). Videnn: Deep blind video denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

[Dai et al., 2016] Dai, D., Wang, Y., Chen, Y., and Van Gool, L. (2016). Is image super-resolution helpful for other vision tasks? In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE.

[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. Ieee.

[Dong et al., 2015] Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

[Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

[Goodfellow et al., 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

[Hao et al., 2018] Hao, S., Wang, W., Ye, Y., Li, E., and Bruzzone, L. (2018). A deep network architecture for super-resolution-aided hyperspectral image classification with classwise loss. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4650–4663.

[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.

[Jiang et al., 2020] Jiang, L., Dai, B., Wu, W., and Loy, C. C. (2020). Focal frequency loss for generative models. *arXiv preprint arXiv:2012.12821*.

[Johnson et al., 2016] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

[Kupyn et al., 2018] Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. (2018). Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192.

[Kupyn et al., 2019] Kupyn, O., Martyniuk, T., Wu, J., and Wang, Z. (2019). Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8878–8887.

[LeCun et al., 1995] LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

[Ledig et al., 2017] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

[Lehtinen et al., 2018] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. (2018). Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

[Ma et al., 2017] Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., and Zhang, L. (2017). Waterloo Exploration Database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016.

[Martin et al., 2001] Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE.

[Mechrez et al., 2018] Mechrez, R., Talmi, I., and Zelnik-Manor, L. (2018). The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Nah et al., 2017] Nah, S., Hyun Kim, T., and Mu Lee, K. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891.

[Philbin et al., 2008] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

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

[Tian et al., 2020] Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. (2020). Deep learning on image denoising: An overview. *Neural Networks*.

[Wang et al., 2018] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

[Xie et al., 2019] Xie, C., Liu, S., Li, C., Cheng, M.-M., Zuo, W., Liu, X., Wen, S., and Ding, E. (2019). Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8858–8867.

[Yadav et al., 2021] Yadav, O., Ghosal, K., Lutz, S., and Smolic, A. (2021). Frequency-domain loss function for deep exposure correction of dark images. *Signal, Image and Video Processing*, pages 1–8.

[Yu et al., 2018] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514.

[Yu et al., 2019] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480.

[Zeyde et al., 2010] Zeyde, R., Elad, M., and Protter, M. (2010). On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer.

[Zhang et al., 2017] Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155.

[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.

Use of Saliency Estimation in Cinematic VR Post-Production to Assist Viewer Guidance

Colm O Fearghail[†], Emin Zerman[†], Sebastian Knorr[‡], Fang-Yi Chao[†] and Aljosa Smolic[†]

[†] *V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland*
[‡] *Ernst Abbe University of Applied Sciences Jena, Germany*

Abstract

One of the challenges facing creators of virtual reality (VR) film is that viewers can choose to view the omnidirectional video content in any direction. Content creators do not have the same level of control on viewers' visual attention as they would on traditional media. This can be alleviated by estimating the visual attention during the creative process using a saliency model, which can provide a probability as to what would draw a viewer's eye. In this study, we analyse both the efficacy of omnidirectional video saliency estimation for creative processes and the potential utility of saliency methods for directors. For this, we use a convolutional neural network-based video saliency model for omnidirectional video. To assist the directors in viewer guidance, we propose a metric that provides a measure of saliency estimation in the intended viewport. We also evaluate the selected saliency model, AVS360, by comparing the output of this saliency model to the actual viewing direction. The results show that the selected saliency model can predict the viewers' visual attention well and the proposed metric can provide useful feedback for content creators regarding possible distractions in the scene.

Keywords: omnidirectional video, 360 degree video, saliency, cinematic VR

1 Introduction

Virtual reality (VR) film, also known as cinematic VR, is a form of VR entertainment that utilises among other formats omnidirectional (also known as 360-degree) video. Visual language in VR is still in development, and currently, the techniques used to relate a narrative to viewers in the form are derived from those of traditional cinema [1]. As the viewer has the freedom to look in any direction of the 360-degree environment that they are present in within the format [2], the director of the content must ensure that the narrative is observed as intended by the viewer and to do so in an immersive manner [3].

One method in which to obtain a probability of how a viewer may view a scene is through the use of computational saliency. Saliency models have been developed to evaluate what attracts the human eye within visual scenes [4]. From psychological studies, it is said that bottom-up and top-down processes take place. Bottom-up being the initial attraction based on the physical properties of the image, then the top-down process begins which relates to the task the viewer has while observing the image [5]. These computational models have also been adapted for use within 360-degree video [6].

Using saliency models to estimate the visual attention for a VR film could help the content creator to attract viewers' attention to areas which they deem important to the understanding of the narrative and gain an understanding of other competing salient areas. To allow this, the saliency models can be integrated into post-production environments, as can be seen in Fig. 1.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/27760.

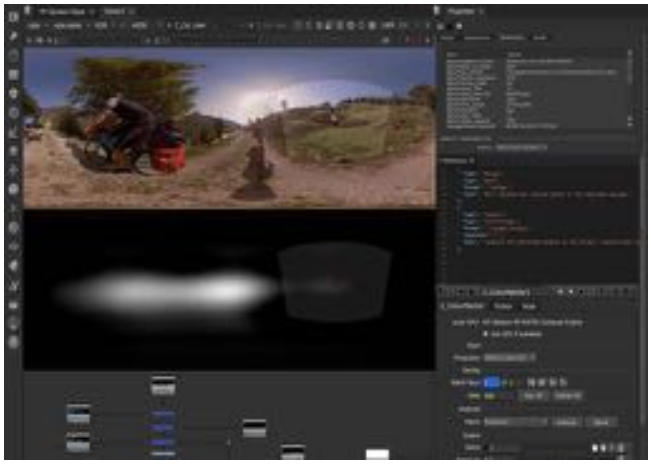


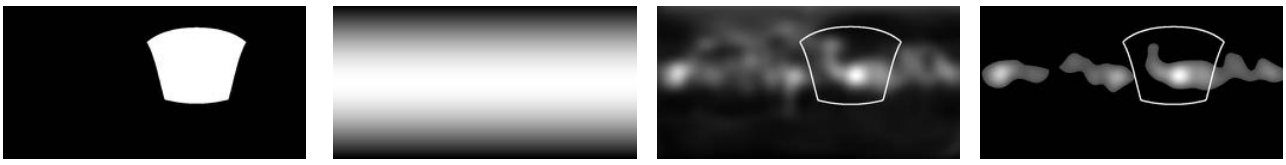
Figure 1: Mock up of a saliency estimator in a post-production environment. The director's cut and corresponding viewport are visible on the RGB image from the film frame and the saliency estimate. The saliency estimate shows that the director's cut viewport is salient, but there is also a region that could compete for viewers' attention. Use of saliency estimation algorithms allows for intervention during post-production.

This paper investigates the effectiveness of using a saliency model in post-production environments, aiming to assist the directors in viewer guidance. With this aim, we build a new metric (i.e., *viewport-based saliency ratio* - VPSR) that can help directors in concentrating the viewers' visual attention and optimising the VR film in the post-production. The proposed VPSR metric can be used with different 360-degree video saliency models. To validate the proposed VPSR metric, we used an omnidirectional video saliency model, i.e., AVS360 [6] developed in our research group for saliency prediction in ODVs (omnidirectional video), and computed the saliency on a VR film database with viewers' visual attention and annotations of director's intended viewing areas, i.e., Director's Cut database [7]. To find out the answer to *"How successful is the selected saliency model in predicting the points that attract visual attention?"*, we first measure the saliency model's output for all the frames of the omnidirectional video. For different contents, the results are then compared to ground truth visual attention to see how well the saliency prediction model performed. Secondly, to answer *"How successful is VPSR in finding frames that need attention?"*, we report the frame-wise results for the proposed viewport-based saliency ratio metric, and we measure how well the director's preferred viewport area is related to the points of saliency within the frame as predicted by the model. Given that AVS360 predicts viewers' attention with high accuracy, the results show the VPSR metric can identify the frames of the video that needs further attention. Our investigation concludes that the use of the proposed metric on saliency estimation methods can identify cases where attention guidance may fail, which can be useful for directors to take appropriate action.

2 Related Work

Among the techniques used by filmmakers in order to communicate their message to the audience are cinematography, mise-en-scène, sound, and editing [8]. In addition to this, various other methods of guiding the viewer within a VR film have been explored. Investigating the methods for guidance, Speicher *et al.* [9] found that giving the viewer an object to follow performed best. Editing from cinematic VR has also formed an area of research [10]. A comprehensive review of papers that have investigated guidance within VR and augmented reality systems can been found in [11].

In order to investigate the ability of techniques derived from traditional cinema within a 360-degree environment, Knorr *et al.* [7] developed a database which included the creator's intended viewing direction at all times throughout the film. This intended viewing direction was given the title of the "Director's Cut" (DC), and this point and corresponding viewport are named as *"DC point"* and *"DC viewport"* throughout this paper. These intended viewing directions were then compared to actual viewing directions of 20 participants to see



(a) Viewport - Ω (b) Weight map - W (c) Saliency map - S (d) Thresholded saliency map - S_{50}

Figure 2: Visualisation of the (a) viewport, (b) weight map, (c) whole saliency map for "Luther" video - Frame #4096 with overimposed viewport, and (d) thresholded saliency map of the same frame for $p = 50$ with overimposed viewport. Please refer to Eqn. 4 for the computation of the VPSR metric.

how viewers consume VR films and how it relates to directors' intentions. Further studies analysed certain elements of the devices used within the films in order to attract the viewers' attention [12] and the styles of cuts, where one scene transitions to another, and their effects on viewer behaviour within the films [13].

To anticipate the user behaviour and estimate viewers' visual attention, saliency estimation methods are developed in image processing and computer vision communities [14]. Due to their spherical nature, omnidirectional images and videos used in immersive systems and VR film are expected to have a different interaction paradigm compared to traditional images and video. How people consume omnidirectional images [15] and video [16] has been explored in the past. Many saliency estimation methods have been developed in the last 20 years [14]; however, more recent advances in the field have been made due to machine learning [17]. An example as to how these models have been used in omnidirectional images can be found in the work of Monroy *et al.* [18] referred to as SalNet360, where the spherical coordinates of the pixels are taken into account. AVS360 [6] which is the saliency model used in this study is a more recent model that caters for omnidirectional video. This model built on work completed in [19]. Development in this area has also included using audio information [20].

To investigate the use of saliency in VR films, we examined the relationship between the SalNet360 saliency estimator and the viewer fixation points at plot points in our previous study [21]. In this earlier study, we focused on plot points in particular and discussed the results for the selected frames. In this paper, differently than in [21], we use AVS360, and we aim to focus on developing a tool that could be used by directors and content creators to identify regions in the scene that could distract viewers from intended viewing areas. To the best of our knowledge, this is the first metric of its kind that will inform directors and content creators.

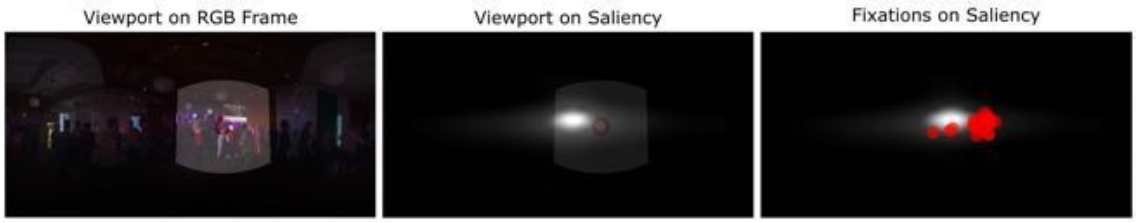
3 Proposed Metric

In this paper, we propose a new metric named *viewport-based saliency ratio* (VPSR) to address the need for a tool that allows directors to optimise their cinematic VR content during post-production. The main goal for this metric is to provide a score to describe the ratio of total probability of estimated saliency within the director's intended viewport. The secondary goal for this metric is to warn the director for possible distractions in the scene that can cause loss of viewers' attention. These distractions then can be avoided using, e.g., virtual effects during post-production.

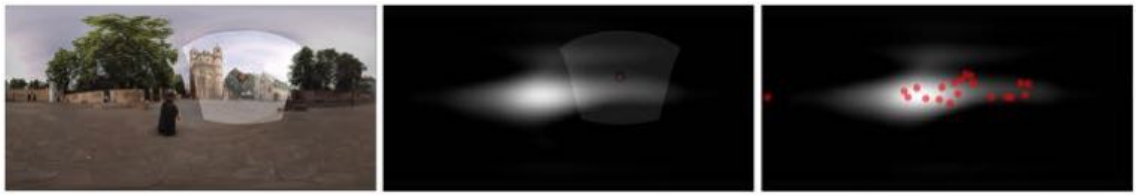
For our VPSR metric, we firstly create a viewport area around the DC point in each frame, which corresponds to the viewport area of the HMD (Head Mounted Display) used in the database study (i.e., an Oculus Rift CV1 headset). Following this, a viewport-based saliency ratio is calculated as follows:

$$\mathrm{VPSR}(S,\Omega) = \frac{\sum_{u,v\in\Omega} S(u,v)W(u,v)}{\sum_{u=1}^{M}\sum_{v=1}^{N} S(u,v)W(u,v)} \tag{1}$$

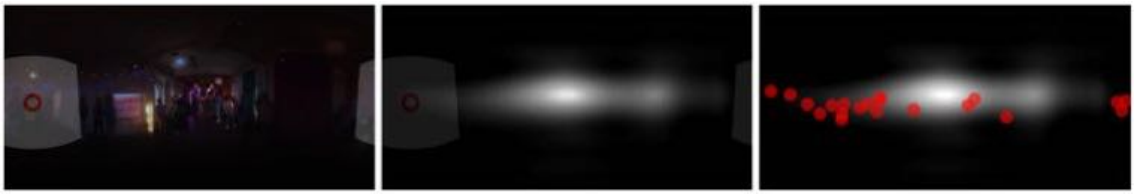
where u and v are the horizontal and vertical pixel locations of an omnidirectional frame of $M \times N$ spatial resolution, Ω is the viewport area as described above (see Fig. 2.(a)), $S(u,v)$ is the saliency map value at (u,v) location, and W is the spherical weighting map. In equirectangular projection, the originally spherical content is increasingly distorted (stretched) along the vertical direction, as we know from geographical maps. The



(a) "Jaunt" - Frame #3680 - VPSR: 0.677 - $VPSR_{50}$: 0.909



(b) "Luther" - Frame #4690 - VPSR: 0.216 - $VPSR_{50}$: 0.060



(c) "Jaunt" - Frame #0848 - VPSR: 0.018 - $VPSR_{50}$: 0

Figure 3: Visualisation of viewers' fixations and viewport corresponding to director's intention from the Director's Cut database [7]. From left to right, *(left)* the RGB film frame with DC point (red circle) and corresponding viewport (white overlay), *(middle)* saliency map with DC point (red circle) and corresponding viewport (white overlay), and *(right)* saliency map with fixation locations (red circles) for two video contents: *(a,c)* "Jaunt" and *(b)* "Luther".

weight map W counters this effect by accounting for spherical distortions, and it gives all parts of the image appropriate contribution to the metric. Here we use the same map as WS-PSNR [22] model, see Fig. 2.(b).

Looking at the distributions of the saliency values with the viewport in Fig. 2.(c) and the fixation distributions in Fig. 3, we notice that the lower saliency values might not attract a lot of viewer fixation. Therefore, we try to generalize the VPSR metric using the saliency values with highest probability. For this, we first compute the histogram of the saliency map S and divide the saliency values into B different bins of $h_i(S)$, where $i \in [1, B]$. Then, taking the highest probability values into account first, we consider the p% of the total saliency and find a threshold value τ for this p as follows:

$$\tau_p = \arg\min_{\tau} \left| \sum_{i=\tau}^{B} h_i(S) - \frac{p}{100} \sum_{i=1}^{B} h_i(S) \right| \tag{2}$$

where B is taken as $B = 256$ for this study. Afterward, using this threshold value, we compute the corresponding thresholded saliency map S_p as follows (see Fig. 2.(d)):

$$S_p(u,v) = \begin{cases} S(u,v), & \text{if } S(u,v) \geq \tau_p \\ 0, & \text{if } S(u,v) < \tau_p \end{cases} \tag{3}$$

This ensures that we always start considering the high probability values. In the last step, we compute $VPSR_p$ as below:

$$VPSR_p(S,\Omega) = \frac{\sum_{u,v\in\Omega} S_p(u,v) W(u,v)}{\sum_{u=1}^{M}\sum_{v=1}^{N} S_p(u,v) W(u,v)} \tag{4}$$

where $S_p(u,v)$ is the thresholded saliency map value at (u,v) location. We can notice that $S_{100} = S$, and Eqn. 4 is the more generic version of Eqn. 1, i.e., $VPSR_{100}$ =VPSR.

The proposed VPSR metric measures how well the DC viewport captures the estimated saliency compared to the whole saliency map. The metric is bounded between $[0, 1]$, where 0 means no saliency values are under the viewport and 1 means all are under the viewport. Sample VPSR and VPSR_{50} results are given in the captions of Fig. 3.

4 Experiments

Here, we describe the dataset used to analyse the efficacy of omnidirectional video saliency estimation for creative processes, the selected saliency estimation method, i.e., AVS360 [6], and the evaluation metrics used.

4.1 Dataset

In this paper, we use the Director's Cut database [7] to analyse how viewers' fixations relate to the estimated omnidirectional video saliency. This database contains a number of cinematic VR films and includes details from the creators as to where they intended to direct the attention of viewers. For this, creators provided their preferred viewport area throughout the films, using the *Tracker* in the commercial compositing software *Nuke*[1] from The Foundry. The centre of this viewport (i.e., *"DC point"*) is recorded as U and V coordinates, horizontally and vertically. The actual viewing directions were then collected from 20 viewers as they watched the films in a natural manner, by collecting the centre point of viewers' viewports [23]. Further details on this technical process can be read in [7]. Fig. 3 visualises the RGB frames, viewers' fixations, and the estimated saliency maps for three different frames. The first column shows the director's intended viewport overlayed on the RGB frame, the second column shows the director's intended viewport overlayed on the estimated saliency map, and the third column shows participants' fixation points plotted over the saliency map.

For our analysis, we selected four of the films from the Director's Cut database: *"DB"*, *"Jaunt"*, *"Luther"*, and *"Vaude"*. These films had the greatest amount of details as provided by the films' creators, and they also had a range of different lighting and guiding devices used within them.

4.2 Saliency estimation method

To investigate the use of omnidirectional video saliency on VR films and creative processes, we selected the AVS360 model [6] as one of the recent saliency models, the implementation of which is publicly available. This model is composed of two 3D residual networks (ResNets) to encode visual and audio cues. The first one is embedded with a spherical representation technique to extract 360° visual features, and the second one extracts the features of audio using the log mel-spectrogram. While this can take spatial audio into account, the DC database was created with videos using mono sound. The AVS360 model was used as is, without any retraining on the DC database. Interested readers are referred to the original paper [6] for further training details.

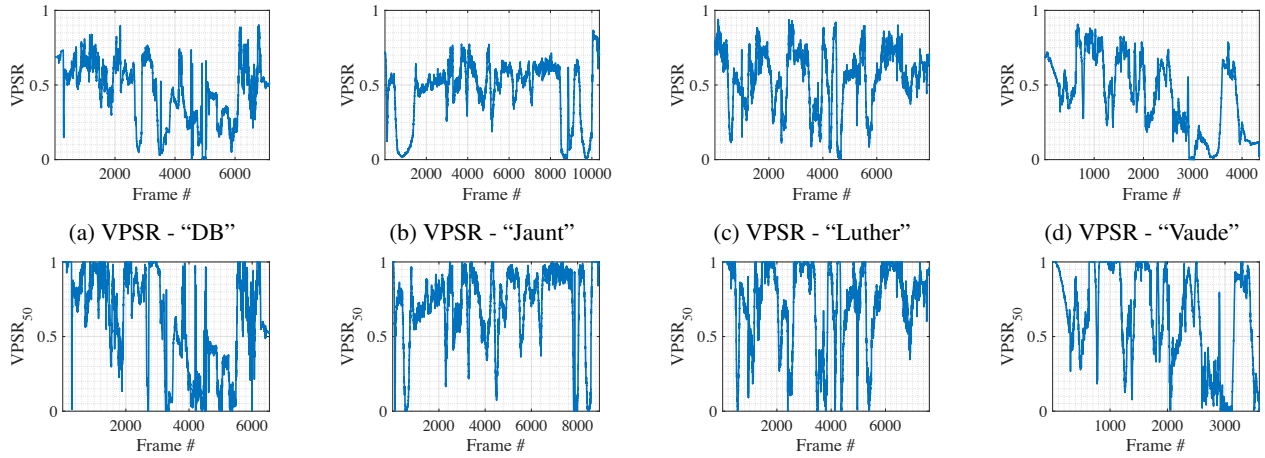
4.3 Evaluation metrics

To evaluate how well the AVS360 model predicts the regions that attract visual attention, we use two saliency evaluation metrics: area under curve (AUC) and normalized scanpath saliency (NSS). Both AUC and NSS are location-based metrics, and they are computed using the ground truth fixation points and estimated saliency map. To compute AUC, the evaluation task was reframed as classification task and the area under the receiver operating characteristic (ROC) curve is computed by finding the true positive and false positive rates. NSS on the other hand first normalises the saliency map (i.e., saliency map is shifted to a mean of zero with standard deviation of one) and estimates the average of the normalised saliency. Additional detail on the metrics used can be found in [24]. To compute these metrics, we used open source implementations for NSS[2] and AUC[3] [25].

[1] https://www.foundry.com/products/nuke

[2] https://sites.google.com/site/saliencyevaluation/evaluation-measures

[3] http://www.saliencytoolbox.net/



(a) VPSR - "DB" (b) VPSR - "Jaunt" (c) VPSR - "Luther" (d) VPSR - "Vaude"

(e) $VPSR_{50}$ - "DB" (f) $VPSR_{50}$ - "Jaunt" (g) $VPSR_{50}$ - "Luther" (h) $VPSR_{50}$ - "Vaude"

Figure 4: Frame-wise evaluation of the relationship between saliency estimation and directors' intention for each video in terms of VPSR.

5 Analysis and Discussion

5.1 Validating the use of AVS360

We first analyse how well AVS360 predicts the ground truth viewing directions. For this, the AUC and NSS metrics are calculated, and the mean AUC and NSS metric results are reported in Table 1.

As we can see from Table 1, both AUC and NSS values support the hypothesis that the selected saliency estimation method, AVS360, can predict the fixation locations well. The AUC score is defined in the range of $[0,1]$, and a high AUC score (e.g., 0.8594 as in Table 1) indicates that the estimated saliency map predicts the distribution of the fixations well. The NSS score shows how large the saliency values correspond to the fixation locations, and having NSS scores ~ 2.5691 means that the saliency values corresponding to fixation locations are 2.5σ away from the mean of the saliency map. That is, the estimated saliency map yields high values at fixation locations. Both of these observations show that the AVS360 model can predict salient regions well. Furthermore, AVS360 can identify locations that might divert visual attention.

5.2 Frame-wise VPSR results

The generic VPSR metric given in Eqn. 4 enables directors to fine-tune the VPSR results by modifying the p value between $[1,100]$. To validate VPSR metric and to show how a change in p affects the results, in this sub-section, we provide the frame-wise results for the proposed VPSR metric for two different cases: considering the whole saliency map ($p = 100$) and considering the highest probabilities that sum up to 50% ($p = 50$).

Sample VPSR metric results were provided in the captions of Fig. 3 along with sample frames. These sample results show that a VPSR value of 0.667 corresponds to a very good overlap between the DC viewport and the estimated saliency while a VPSR value of 0.018 indicates poor correspondence. The values are more intuitive for $VPSR_{50}$ as it yields both higher and lower values for these examples. Fig. 4 shows the overall

Table 1: Mean AUC and NSS metric results across all frames comparing saliency maps and viewers' viewing directions.

Film	**"DB"**	**"Jaunt"**	**"Luther"**	**"Vaude"**	**Overall**
$AUC_{Viewers}$	0.8940	0.9264	0.9346	0.8594	0.9036
$NSS_{Viewers}$	1.7367	2.7249	2.7531	2.5691	2.4459

results and allows analysis of how well saliency prediction and directors' intent agree. We can identify dips, which indicate areas that may require intervention to keep viewers' attention. Overall, the VPSR metric was higher for "Luther" compared to other contents. The graphs for VPSR and $VPSR_{50}$ show similar characteristics, while $VPSR_{50}$ has larger swings; therefore, it might provide more intuitive results for the director.

6 Conclusion

In this paper, we proposed a metric that allows directors to optimise their cinematic VR content for viewer guidance. To demonstrate how this metric is capable of yielding useful scores for directors, we used the AVS360 saliency estimation method on an omnidirectional video dataset. We first validated that AVS360 predicts viewers' attention well, and then we presented frame-wise VPSR results. The visual results along with the frame-wise results show that the VPSR metric is indicative of how well the intended viewports could retain viewers' attention.

The results indicate that the AVS360 model and the VPSR metric could form part of a plug-in that will notify the director of regions of possible distractions within the film. The directors will be presented with frame-wise VPSR results as shown in Fig. 4 and they can identify the dips in VPSR values (e.g., dips in visual attention) without checking the saliency estimation results for all the frames manually. With this information the director could then alter the film set during the production or use visual effects (VFX) in post-production accordingly. The VFX option could even be done in an adaptable manner should the viewers' attention stray.

References

[1] K. Dooley, "Storytelling with virtual reality in 360-degrees: a new screen grammar," *Studies in Australasian Cinema*, vol. 11, no. 3, pp. 161–171, 2017.

[2] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[3] J. Mateer, "Directing for cinematic virtual reality: How the traditional film director's craft applies to immersive environments and notions of presence," *Journal of Media Practice*, vol. 18, no. 1, 2017.

[4] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1153–1160.

[5] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[6] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 355–358.

[7] S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic, "Director's cut: A combined dataset for visual attention analysis in cinematic VR content," in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*. ACM, 2018, p. 3.

[8] M. Vosmeer and B. Schouten, "Project orpheus a research study into 360° cinematic VR," in *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, 2017.

[9] M. Speicher, C. Rosenberg, D. Degraen, F. Daiber, and A. Krúger, "Exploring visual guidance in 360-degree videos," in *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, 2019, pp. 1–12.

[10] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia, "Movie editing and cognitive event segmentation in virtual reality video," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[11] S. Rothe, D. Buschek, and H. Hußmann, "Guidance in cinematic virtual reality-taxonomy, research status and challenges," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 19, 2019.

[12] C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut - analysis of aspects of interactive storytelling for VR films," in *International Conference on Interactive Digital Storytelling*. Springer, 2018, pp. 308–322.

[13] C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut - analysis of VR film cuts for interactive storytelling," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2018.

[14] L. Itti and A. Borji, "Computational models: Bottom-up and top-down aspects," in *The Oxford Handbook of Attention*, A. C. Nober and S. Kastner, Eds. Oxford University Press, 2014.

[15] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017.

[16] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018.

[17] K. Zhang, Z. Chen, and S. Liu, "A spatial-temporal recurrent neural network for video saliency prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 572–587, 2020.

[18] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "SalNet360: Saliency maps for omni-directional images with CNN," *Signal Processing: Image Communication*, vol. 69, pp. 26–34, 2018.

[19] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 01–04.

[20] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[21] C. O. Fearghail, S. Knorr, and A. Smolic, "Analysis of intended viewing area vs estimated saliency on narrative plot structures in VR film," in *International Conference on 3D Immersion*, 2019. [Online]. Available: https://v-sense.scss.tcd.ie/?attachment_id=4339

[22] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, Sept 2017.

[23] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *10th International Conference on Quality of Multimedia Experience (QoMEX)*, Sardinia, Italy, May 2018.

[24] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[25] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

More efficient Geospatial ML modelling techniques for identifying man-made features in Aerial Ortho-imagery

Samuele Buosi, Shubham Sonarghare, John McDonald and Tim McCarthy

Maynooth University

Abstract

Deep learning techniques are used to achieve state-of-art accuracy in semantic segmentation on aerial ortho-imagery datasets. These algorithms are known to be efficient in terms of accuracy but at the expense of computational power required for training and subsequent inference operations. In this paper we strive to achieve a comparable performance but with lower floating point operations per second (FLOPS) and less training time. With this in mind, we chose to evaluate the EfficientNet-B0 network configured with 5.3 millions parameters and 0.39 billion FLOPS as a feature extractor operating inside a U-net architecture, achieving accuracy levels (mean F1 score of 0.869) comparable to a state-of-the-art deep learning architecture (U-net with Resnet50 as backbone) configured with 25.6 million parameters and 4.1 billion FLOPS which achieved a mean F1 score of 0.87. These promising results demonstrate that employing EfficientNet as the feature extractor in semantic segmentation on aerial ortho-imagery can be an effective strategy, in achieving higher performance results in terms of computational power, especially when running these networks on the edge.

Keywords: Deep Learning, Supervised Image Segmentation, semantic segmentation, ortho-imagery, Deep convolutional neural network

1 Introduction

Over the past decade, advances in Machine Learning (ML) and in particular Deep Learning (DL) algorithms have resulted in significant advances in Computer Vision. One of the key applications is Semantic Segmentation which is used in a number of applications including; Robotic Localisation, Autonomous Driving, Scene Understanding and, building High-Definition Maps [Kemker et al., 2018].

In terms of geospatial applications, unmanned aerial vehicles (UAVs) are playing an increasing role in data gathering and mapping our real world environments. These robotic aerial data gathering platforms are now commonly found across the globe, collecting large volumes of data that require automated processing such as feature extraction to be carried out on the fly. Such requirement demands both computationally inexpensive and high accuracy feature extraction techniques [Ammour et al., 2017].

Most common and well-known traditional techniques in computer vision like Support Vector Machines, [Waske and Benediktsson, 2007], and Random Forests, [Pal, 2005], often result in less accurate outputs compared to the DL techniques that produce significantly improved accuracy but at the expense of resources required to train and carry out subsequent inference [O'Mahony et al., 2020]. In this paper we investigate the potential for EfficientNet family, [Tan and Le, 2019], to help reduce this expense in extracting man-made features in UAV aerial imagery. We investigate this hypothesis using an U-Net architecture, [Ronneberger et al., 2015], with an EfficientNet-B0, [Tan and Le, 2019] feature extractor. To assess the performance of the resulting architecture we utilise the International Society for Photogrammetry and Remote Sensing (ISPRS) benchmark dataset [ISPRS, 2016].

2 Prior Work

Recent developments in aerial robotic data gathering platforms, such as UAVs, now enable the rapid capture of aerial imagery at higher spatial-temporal resolutions as well as lower costs. In parallel, emerging developments in

contemporary DL algorithms in automating the data processing and feature extraction has resulted in new data products and information services for applications including; urban planning, land cover classification, Emergency Response, etc. [Ammour et al. 2017].

It is possible to generate an orthophoto from overlapping aerial imagery that is geometrically corrected (orthorectified) so it can be used to measure true distances and dimensions. The process of orthorectification enables various real-world phenomena and distortion such as topographic relief, lens distortion and camera orientation to be corrected [Habib et al., 2007].

Semantic Segmentation is an important algorithm that can assign a class to each pixel of a given image where the classes are defined *A Priori*. Semantic Segmentation applied to ortho-imagery is very useful and important because of its ability to detect and categorise one or more classes in the ortho-image [Liu et al., 2018]. Traditional image segmentation methods include; Watershed, Graph Cuts and Random Forests which have been used to classify high-resolution aerial images [Meyer and Beucher, 1990; Boykov and Jolly, 2001; and Pal, 2005]. However, DL techniques involving convolutional neural networks have proven to be more efficient and effective in extracting features from images compared to these more traditional approaches [Deng et al., 2009]. DL methods perform well even for semantic segmentation due to their ability to automatically extract features. For example, in 2015, there was a 20% relative improvement to 62.2% mean Intersection over Union (IoU) using a Fully Convolution Networks (FCNs based on the PASCAL VOC 2012 benchmark dataset compared to the state-of–the-art techniques of that time [Long et al., 2015].

There are many Neural Network architectures that utilise CNNs for semantic segmentation tasks e.g., U-net [Ronneberger et al., 2015], LinkNet [Chaurasia and Culurciello, 2017], Feature Pyramid Networks [Li et al., 2019]. As an example, [Wu et al., 2018] uses U-net [Ronneberger et al., 2015] for automatically segmenting building features from aerial imagery. Similarly, [Boonpook, et al., 2018] uses SegNet [Vijay et al., 2016] to extract building features from UAV images for riverbank monitoring. One of the novelties of these architectures is their compatibility and adaptability with a range of feature extractors. For example, one can use VGG [Simonyan and Zisserman, 2015] as the feature extractor in a U-net architecture [Ronneberger et al., 2015] or use ResNet [He et al., 2016] inside a LinkNet [Chaurasia and Culurciello, 2017]. The performance of these networks completely depends on the performance of the feature extractor in combination with how the architecture combines these features to segment the objects under observation. More recently, ScasNet [Liu et al., 2018] which utilized Resnet [He et al., 2016] as a feature extractor, achieved one of the best results with an overall accuracy of 91.1% on the ISPRS Potsdam benchmark dataset [Liu et al., 2018]. With a more complex feature extractor is it possible to achieve higher performance with respect to accuracy in resulting object segmentation, but this also increases the number of parameters to train. This gives rise to computationally more expensive requirements since high-performing DL techniques require relatively large volumes of training data to train models with a high number of parameters.

In this paper, we investigate the potential for a more efficient and scalable Semantic Segmentation Neural Network architecture that allows a comparable level of performance to be achieved similar to the actual state-of-art applied to ortho-imagery from the ISPRS Potsdam benchmark dataset. To this end, we employ a combination of a U-net architecture [Ronneberger et al., 2015], an EfficientNet [Tan and Le, 2019] feature extractor and focal/dice loss [Lin et al., 2020, Deng et al., 2018].

3 Technical Description

The main drawback of the majority of CNNs are their tendency to down-scale or reduce the spatial resolution of the features along the depth of the network which is not ideal in a segmentation context.

To overcome down-sampling of the spatial resolution, many Fully Convolutional Neural Networks have been suggested like Segnet [Vijay et al., 2016], U-net [Ronneberger et al., 2015]. We chose a U-net architecture with an EfficientNet-b0 as the feature extractor, after an initial assessment based on literature review, for this study.

The U-Net architecture is a CNN widely used for Semantic Segmentation. The original network consists of an encoder path and a decoder path that gives the U-shaped architecture. The Encoder part is composed by repeated convolution layers, each followed by a rectified linear unit (ReLU) layer and a maximum pooling layer. The decoder part is composed by sequence of up-convolutions and concatenations.

Figure 1: Overall U-net architecture using EfficientNet-b0

The U-net architecture readapted with the EfficientNet-B0 as the encoder is detailed in Figure 1. The EfficientNet is a family of Convolutional Neural Networks developed in the context of AutoML where the authors have investigated a possible solution for neural network (NN) scaling for efficiency [Tan and Le, 2019]. Tan and Le, [Tan and Le, 2019], created a first baseline EfficientNet-B0 inspired by a MnasNet and scaled up to the B7 network using their new compound scaling method, optimizing both accuracy and FLOPS at the same time. As a result, the network is faster and smaller compared to the other major networks used based on the ImageNet benchmark dataset [Tan and Le, 2019]. Specifically, EfficentNet-B0 uses 4.9 times less parameters and 11 times less FLOPS compared to ResNet-50 while providing 77.1 % as Top-1 accuracy on ImageNet compared to 76.1% of ResNet-50 [He et al., 2016]. Figure 2 shows the EfficientNet-B0 architecture.

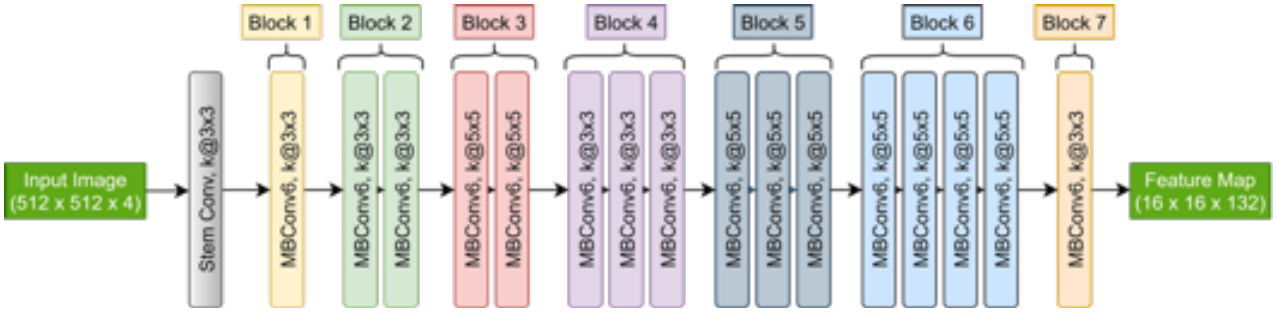


Figure 2: Architecture of EfficientNet-B0 as feature extractor

Along with the architecture, it also important to carefully select the loss function which will penalize the network for incorrect predictions and detections. Standard cross-entropy loss is calculated as the average of per-pixel loss. This poses a huge issue when the number of foreground pixels are far less than the number of background pixels. Although, weighted cross entropy loss helps alleviate this problem, it does not result in a significant improvement. To overcome this issue, we used a combination of a focal and dice loss. While the focal loss helps in learning hard negative examples and addresses the issue of class imbalance, dice loss helps to learn better class boundaries [Lin et al., 2020, Deng et al., 2018].

The Dice Loss is defined by

$$TP(c) = \sum_{i=1}^{N} p_i(c) g_i(c)$$

$$FN(c) = \sum_{i=1}^{N} (1 - p_i(c)) g_i(c)$$

$$FP(c) = \sum_{i=1}^{N} p_i(c)(1 - g_i(c))$$

$$\mathcal{L}_{Dice} = C - \sum_{c=0}^{C-1} \frac{2TP(c)}{2TP(c)+FP(c)+FN(c)} \quad (1)$$

where C is the total number of classes, N is the total number of pixels, $p_i(c)$ is the predicted class of the pixel, $g_i(c)$ is the ground truth class of the pixel. TP, FP and FN are respectively the true positives, false positives, false negatives of a particular class. The Focal Loss is defined by

$$\mathcal{L}_{Focal} = -\lambda \frac{1}{N} \sum_{c=0}^{C-1} \sum_{i=1}^{N} g_i(c)(1 - p_i(c))^{\gamma} log(p_i(c)) \quad (2)$$

The focusing parameter γ was set to 2 and the weighting factor λ was set to 0.25 in our experiment. Thus, the total loss is given by,

$$\mathcal{L}_{DF} = \mathcal{L}_{Dice} + \mathcal{L}_{Focal}$$

$$= C - \sum_{c=0}^{C-1} \frac{2TP(c)}{2TP(c)+FP(c)+FN(c)} - \lambda \frac{1}{N} \sum_{c=0}^{C-1} \sum_{i=1}^{N} g_i(c)(1 - p_i(c))^{\gamma} log(p_i(c)) \quad (3)$$

4 Experiments

4.1 Implementation

We implemented U-net architecture using Tensorflow 2.3.1 with CUDA 10.1 support. Training images are read on the fly and randomly augmented using Tensorflow data API. We did our performance tests using a graphics processing unit (GPU) NVIDIA GeForce GTX 1650 with 4 GB of GPU memory.

4.2 Benchmark Dataset

We applied and studied the performance of the architecture described in section 3 with the ISPRS Potsdam benchmark dataset [ISPRS, 2016]. This benchmark dataset contains 38 ortho-images of same size of 6000 x 6000 pixels generated from cropping a larger orthophoto at a ground sampling distance (GSD) of 5 cm. Each ortho-image in the dataset consist of 4 channels IRRGB (Infrared, Red, Green, Blue) and for each ortho-image, there is a corresponding Digital Surface Model (DSM), representing elevation and normalised DSM (nDSM) data. The ground truth labels are also provided for training purposes for 24 of these 38 ortho-images. An example of the dataset is detailed in Figure 3 where a ISPRS RGB patch is overlapped with the ground truth. The ground truth colour map used for ISPRS classes/objects is listed in Table 1.

Figure 3: Labels overlapped on a RGB ortho-image crop from ISPRS Potsdam dataset

Colour	Class
White	Impervious Surfaces
Blue	Buildings
Cyan	Low Vegetation
Green	Trees
Yellow	Car
Red	Clutter

Table 1: ISPRS colour and class definition

4.3 Training and Evaluation

For the experiments, we pre-processed the raw ISPRS Potsdam dataset and generated 4681 patches of 512x512 pixels each with the infrared (IR), red (R), green (G), and normalized digital surface model (nDSM) band. Every patch has the correlated mask in a different folder with the same patch name in .tif format. For training, we used an 80/20 split so, 80% of all the 4681 patches was used for train the model and the remaining 20% patches was used for validation purposes. Data was also normalized, and data augmentation was applied, which consisted of random rotation of 90°, vertical and horizontal flips with a probability of 0.5. We choose a batch size of 4 due to our memory constraints. We did not use the Transfer Learning technique because most of the common pre-trained weights are based on RGB images, but in this case, we have 4 channels corresponding to IR, RG and nDSM data. Hence, we initialized the network with Xavier initialization [Glorot and Bengio, 2010]. The initial learning rate (LR) was set to 0.001 with a learning rate scheduler that monitored the validation loss. The LR was set to decrease by a factor of 0.1 every 5 epochs if the validation loss doesn't reduce. The minimum LR was set to 1e-15. The optimizer chosen was Adam [Kingma and Ba, 2015].

We trained two models using the ISPRS Potsdam dataset and created a comparison table (Table 2) with the F1 score metric (2) per class and reporting the number of parameters and FLOPS required. All the models are based on the same U-net architecture but with a different feature extractor. We chose to compare EfficientNet B0 with ResNet50 because these two architectures have comparable performances [Tan and Le, 2019]

We assessed quantitative performance of the two models based on the F1 score applied to all the six classes as,

$$F1 = 2\frac{Precision*Recall}{Precision+Recall} \tag{4}$$

where, Precision and Recall are defined by:

$$Precision = \frac{TruePositives}{TruePositives+FalsePositives} \tag{5}$$

$$Recall = \frac{TruePositives}{TruePositives+FalseNegatives} \tag{6}$$

4.4 Results and Analysis

We generated predictions for each of the fourteen ortho-images contained in the ISPRS Potsdam test dataset and compared to the ground truth calculating the metrics for both architectures. We also produced qualitative results as shown in Figure 4 where we show the IRRG image, the ground truth, the prediction with Resnet50 and the prediction with EfficientNet-B0 based on an ortho-image from the ISPRS Potsdam test dataset.

As seen from Table 2, EffientNet-b0 resulted in almost the same weighted F1 score as ResNet-50 but with 4.9x less parameters and 11x less FLOPS. This resulted in comparable performance when comparing EfficientNet-b0 to ResNet-50 but with significantly less computational overhead.

Architecture	Num. of parameters	FLOPS	Weighted Mean F1	F1-Scores					
				Impervious Surfaces	Buildings	Low Vegetation	Trees	Car	Clutter
U-net + EfficientNet -B0	5.3M	0.39B	0.869	0.89	0.95	0.82	0.83	0.89	0.41
U-net + ResNet50	25.6M	4.1B	0.87	0.89	0.95	0.82	0.82	0.88	0.45

Table 2: Model comparison

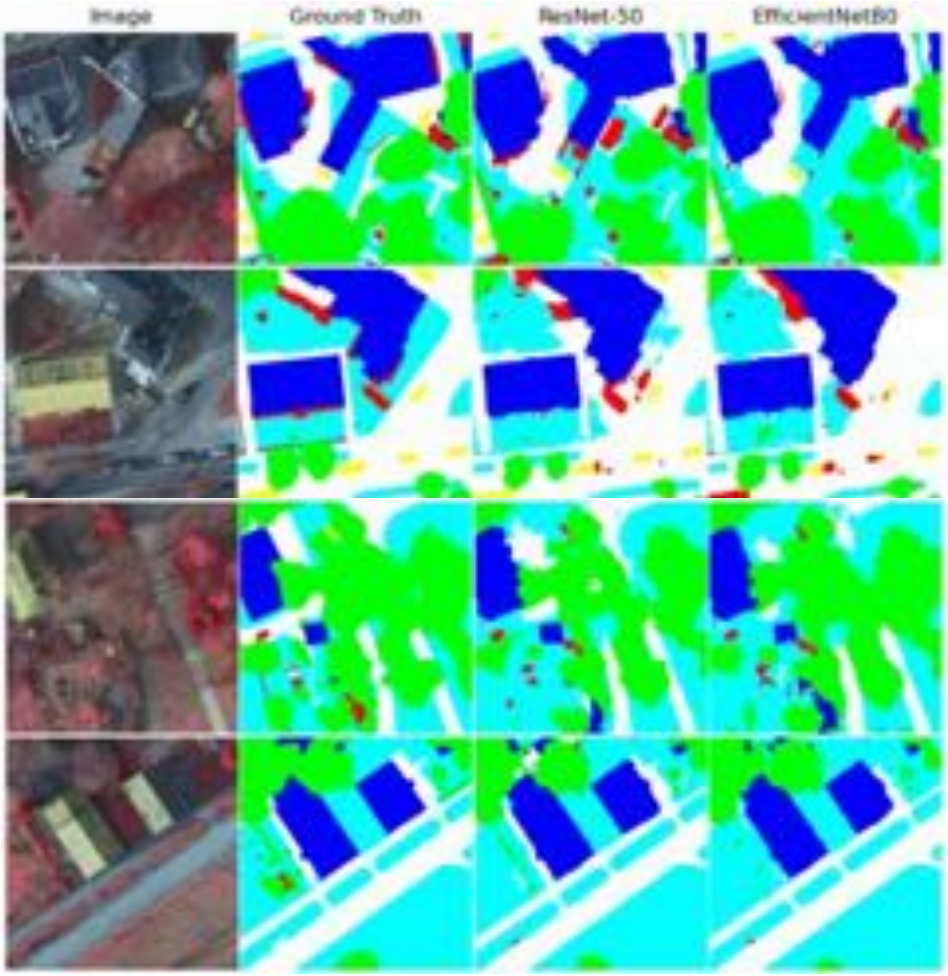


Figure 4: Qualitative comparison side by side of the inference from the models on the ISPRS Potsdam test set.

5 Conclusions

In this paper, we investigated a more efficient Neural Network architecture that can achieve state-of-art performance on Semantic Segmentation applied to ortho-imagery, captured using UAVs. We reviewed a Neural Network based on a U-net architecture but modifying the features extractor with the new EfficientNet-B0. We were not interested in accuracy alone, but also examining the possibility of reducing the computational power required by the common architecture ResNet50. Initial results are promising and scalable. Further experimentation could be conducted on testing and evaluating the robustness and versatility of these architectures using different datasets and comparing the results also with other well-known Semantic Segmentation architectures.

Acknowledgements

This material is based upon works supported by U-Flyte (Unmanned Aircraft Systems Flight Research) 17/SPP/3460 which is funded under the Science Foundation Ireland Strategic Partnership Programme

References

[Ammour et al. 2017] Ammour, N., Alhichri, H., Bazi, Y., Benjdira, B., Alajlan, N., & Zuair, M. (2017). Deep learning approach for car detection in UAV imagery. *Remote Sensing*, *9*(4). https://doi.org/10.3390/rs9040312

[Boykov and Jolly, 2001] Boykov, Y.Y.; Jolly, M.P. Interactive graph cuts for optimal boundary and region segmentation of objects in ND images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Vancouver, BC, Canada, 7–14 July 2001; pp. 105–112.

[Boonpook, et al., 2018] Boonpook, W., Tan, Y., Ye, Y., Torteeka, P., Torsri, K., & Dong, S. (2018). A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring. Sensors (Switzerland), 18(11). https://doi.org/10.3390/s18113921

[Chaurasia and Culurciello, 2017] Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. 2017 IEEE Visual Communications and Image Processing, VCIP 2017, 2017-Janua, 1–4. https://doi.org/10.1109/VCIP.2017.8305148

[Deng et al., 2018] Deng R.; Shen C.; Liu S.; Wang H.; Liu X., "Learning to predict crisp boundaries," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11210 LNCS, pp. 570–586, 2018.

[Deng et al., 2009] Deng J., Dong W., Socher R., Li L., Li K. and L. F.-F. (2009). ImageNet: A Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[Glorot and Bengio, 2010] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks Xavier. AISTATS , volume 9 of JMLR Proceedings, page 249-256. JMLR.org.

[Habib et al., 2007] Habib, A. F., Kim, E. M., & Kim, C. J. (2007). New methodologies for true orthophoto generation. Photogrammetric Engineering and Remote Sensing, 73(1), 25–36. https://doi.org/10.14358/PERS.73.1.25

[He et al., 2016] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 770–778. https://doi.org/10.1109/CVPR.2016.90

[Huang et al., 2017] Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

[ISPRS, 2016] International society for photogrammetry and remote sensing. 2D Semantic Labeling Challenge. Available at: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/

[Kingma and Ba, 2015] Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–15.

[Kemker et al., 2018] Kemker, R., Salvaggio, C., & Kanan, C. (2018). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *145*, 60–77. https://doi.org/10.1016/j.isprsjprs.2018.04.014

[Li et al., 2019] Li, X., Lai, T., Wang, S., Chen, Q., Yang, C., & Chen, R. (2019). Weighted feature pyramid networks for object detection. Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019, 1500–1504. https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00217

[Lin et al., 2020] Lin T. Y.; Goyal P.; Girshick R.; He K.; Dollar P., "Focal Loss for Dense Object Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, 2020.

[Liu et al., 2018] Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., & Pan, C. (2018). Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS Journal of Photogrammetry and Remote Sensing, 145, 78–95. https://doi.org/10.1016/j.isprsjprs.2017.12.007

[Long et al., 2015] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

[Marmanis et al., 2016] Marmanis D., Datcu M., Esch T., and Stilla U., "Deep Learning Earth Observation Classification using ImageNet Pretrained Networks," IEEE Trans. Geosci. Remote Sens. Lett., vol. 13, no. 1, pp. 105–109, 2016.

[Meyer and Beucher, 1990] Meyer, F.; Beucher, S. Morphological segmentation. J. Vis. Commun. Image R. 1990, 1, 21–46.

[O'Mahony et al., 2020] O'Mahony N.; Campbell S.; Carvalho A.; Harapanahalli S.; Hernandezo G.; Krpalkova L.; Riordan D.; Walsh J., "Deep Learning vs. Traditional Computer Vision," Adv. Intell. Syst. Comput., vol. 943, no. Cv, pp. 128–144, 2020.

[Pal, 2005] Pal, M. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 2005, 26, 217–222.

[Peng et al., 2017] Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. arXiv 2017, arXiv:1703.02719.

[Ronneberger et al., 2015] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv 2015, arXiv:1505.04597.

[Simonyan and Zisserman, 2015] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14.

[Tan and Le, 2019] Tan M. and Le Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." in Proceedings of the 36th International Conference on Machine Learning, 2019.

[Vijay et al., 2016] Vijay, B.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495.

[Waske and Benediktsson, 2007] Waske, B., & Benediktsson, J. A. (2007). Fusion of support vector machines for classification of multisensor data. IEEE Transactions on Geoscience and Remote Sensing, 45(12), 3858–3866. https://doi.org/10.1109/TGRS.2007.898446

[Wu et al., 2018] Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., & Shibasaki, R. (2018). Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks. Remote Sensing, 10(3), 1–18. https://doi.org/10.3390/rs10030407

Automated Ki-67 proliferation scoring from histopathology images using Mobile and Cloud technology

Miranda J.E McConnell[1], Richard Gault*[1], Stephanie G. Craig[2], David Cutting[1], Austen Rainer[1], and Jacqueline James[2,3]

[1]*School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast*
[2]*Precision Medicine Centre, Patrick G Johnston Centre for Cancer Research, Queen's University, Belfast*
[3]*Belfast Health and Social Care Trust, Belfast, Northern Ireland*

Abstract

The Ki-67 protein is associated with cell proliferation and is a clinical marker for breast cancer tumour aggressiveness. The percentage of immunopositive cells present in a histological image stained for Ki-67 expression informs the proliferation index quantifying tumour aggressiveness. This calculation is frequently carried out through manual assessment that is time consuming and susceptible to human error. Automated image analysis tools for Ki-67 breast cancer images may have a significant impact if they could be integrated in to clinical and digital pathology workflows by reducing workload for pathologists, as well as improving efficiency and accuracy. This work presents the development of a deep learning based model for automated calculation of Ki-67 proliferation scores from stained histological images. The resulting computational model predicts cell types (immunopositive vs immunonegative) with 96% accuracy, the Ki-67 index category with 88% accuracy and the Ki-67 index with lower RMSE than the state of the art models. The predicted mask from the model provides a transparent explanation of the computational decision making. Moreover, the computational model is hosted on a cloud platform and can be utilised through a mobile application designed for this investigation. The proof-of-concept mobile application has the potential to make an impact in many communities, especially in low and middle income countries where there are currently insufficient resources, namely a lack of expensive digital scanners, to support digital pathology in the fields of medicine and education.

Keywords: Deep learning, image segmentation, Ki-67 proliferation score, mobile technology, cloud computing

1 Introduction

In recent years, the Ki-67 protein has been investigated as a clinical marker for breast cancer tumour aggressiveness [Yerushalmi et al., 2010, Inwald et al., 2013]. The biomarker is a nuclear protein associated with cell proliferation; the increase in the number of cells as a result of cell growth and division. Pathologists use scoring systems to estimate a proliferation index; low (<10%), borderline (10-20%), and high (>20%). A higher proliferation index indicates that more cells are undergoing cell division, which can signify a more aggressive tumour.

Currently, there are several challenges incorporating automated Ki-67 proliferation index measurements in to clinical and digital pathology workflows without using a digital scanner. Measuring the proliferation index manually is time consuming and sensitive to variance. This variance could be due to differences in staining protocols, digitisation equipment, staining compounds or slide preparation, which can create variabilities in

*richard.gault@qub.ac.uk

image quality and colour across datasets. The cell nuclei are also subject to variance in terms of structure, shape, colour and intensity [Joseph et al., 2019]. Diagnostics from histopathology images usually rely on a visual assessment of the cell slides by a pathologist, which can imply an inherent element of interpretation with consequent subjectivity and possible human error. This manual process can be time-consuming and susceptible to human error, there is a motivation to introduce computational methods to encode the expertise of the decision making process.

Logistically, scanners required for digitising slides are expensive and don't offer portability. Advances in mobile phone camera technology have shown the capability to take satisfactory resolution images of microscopy cell slides. [Hernández-Neuta et al., 2019] acknowledge that the image sensors within a smartphone's camera module are sensitive enough for many diagnostically relevant applications. There is the potential for the adaptation of smartphones as imaging read-out platforms that could be used for on-site image acquisition, real-time analysis, management of the generated results at the user's convenience, and data transfer from the site of detection to other healthcare professionals.

This highlights an opportunity for digital pathology tools (and digital analysis) to be shared with the world through mobile devices, thanks to the portability and technological features they offer. Not only would this have a positive impact in the practice of pathology, but it would particularly benefit communities in low and middle income countries, where the current option of using an expensive scanner is not feasible. Furthermore, connecting global experts through a common platform that facilitates ease of communication and knowledge transfer would support the digital pathology community as a whole but specifically those working in low and middle income countries.

This paper presents a novel investigation in to the automated Ki-67 proliferation scoring of histological images. This work also presents the development of a prototype mobile application that would enable multiple users to interact on a single platform and analyse images stored on or captured by the device. Section 2 will outline the current research in the area of automated Ki-67 proliferation scoring. Section 3 will outline the proposed methodology and computational system along with the experimental protocols used to evaluate the computational model. Section 4 provides an overview of the results before the findings of this work are discussed in Section 5 in the context of existing work. The conclusion (Section 6) summarises the main contributions of this work and highlights future opportunities for research.

2 Background

The advances and success of deep learning methodologies in the area of image processing combined with the quantity and quality of image data in the digital pathology domain has led to a surge in the development of deep learning solutions to support digital pathology analysis. In particular in the area of Ki-67 proliferation scoring, a number of approaches have been considered.

Proliferation Tumour Marker Network (PTM-NET) [Joseph et al., 2019] is a four layer convolutional neural network (CNN) that performs instance segmentation on cells before identifying Ki-67 immunopositivity in supplementary analysis. PTM-NET predicts immunopositive and immunonegative cells with an accuracy of 70% and 88% respectively. The performance of the algorithm in calculating the proliferation index is not presented but the work shows how a relatively simple CNN can be used to provide reasonable accuracy in identifying Ki-67 expression. A more complex solution using a similar aproach is PathoNet [Negahbani et al., 2021] which utilises U-Net [Ronneberger et al., 2015] as the foundation for its modelling framework with the first layer and convolutional layers replaced by a residual dilated inception module that reduces model complexity. The model predicts Ki-67 immunopositive and immunonegative cells with 85% and 75% accuracy respectively. Ki-67 proliferation index scoring was achieved with an root-mean-squared error (RMSE) of 0.62 which was in keeping with alternate benchmark models. Cell segmentation in the predicted masks is carried out using the Watershed algorithm [Atta-Fosu et al., 2016]. The findings highlight the potential for U-net to form the basis of modelling efforts in this area, although the variance in the staining and imaging of the samples may result in the cell segmentation algorithm performing to a lower standard than desired.

Instead of taking an instance segmentation approach to cell detection, [Geread et al., 2019] proposed a novel unsupervised colour separation model before distinguishing cells through post processing and nuclei detection algorithms. The model achieved a classification accuracy of 92.5%. This unsupervised approach avoids human error in the subjective labelling and ground truth annotation of slides. However, it is difficult to say how well the colour separation model would generalise across laboratories where staining protocols can lead to very different colour profiles in images. It is desirable in the current study to prioritise generalisability over absolute performance on a given dataset due to the aims to support digital pathology analysis in low and middle income countries that would be spread across different laboratories.

3 Methodology and experimentation

This section will first outline in Section 3.1 the development of the computational model for detecting Ki-67 expression and the dataset used in this endeavour. The mobile application that was simultaneously developed to house the computational analysis is outlined in Section 3.2.

3.1 Computational modelling

3.1.1 Data

Annotated cell slides of microscopic biopsy images of malignant breast tumours containing Ki-67 protein expression were obtained from [Negahbani et al., 2021]. Each raw image is downsized from a starting dimension of 1228×1228 to a final dimension of 256×256. Segmentation masks for each image are generated using the OpenCV-Python library. An example of the raw images and its corresponding mask are illustrated in Figure 1. The staining protocol results in cells that express the Ki-67 protein appearing as a dark brown colour (Figure 1 (a)). Adaptive Gaussian thresholding is first applied to the images to separate cells from the background and create a binary image with background pixels in black and cell pixels in grey. This image is then used to draw contours around each cell. Ground-truth annotations are contained in a corresponding JSON file comprised of co-ordinates for the annotated nuclei, as well as their classification (Ki-67 immunopositive or immunonegative). The centre position of each contour is compared with the JSON file to locate all the cells identified as being immunopositive. Any matches have their pixels changed from grey to white (Figure 1 (b)). This forms the ground truth masks for training of the deep learning model.

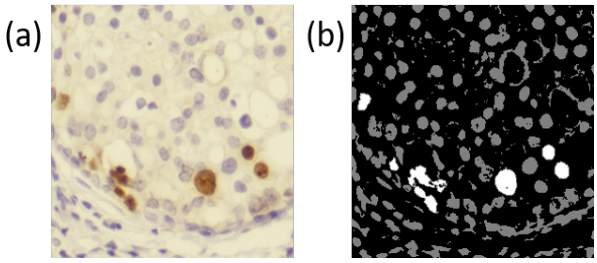


Figure 1: (a) shows an example raw image and (b) shows the resultant ground truth mask generated from the JSON annotation.

3.1.2 Experimental overview and model outline

The instance segmentation in this study is achieved through the application of transferred learning on the U-Net model [Ronneberger et al., 2015], with a pretrained ResNet encoder (specifically ResNet50 [He et al., 2016]) and ImageNet [Deng et al., 2009] weights developed in PyTorch (version 1.8.1). The optimiser is Adam [Kingma and Ba, 2014] and the learning rate is set to 0.001. All weights are considered trainable. The model is trained for 10 epochs. Preliminary analysis found that the model plateaued at sufficient accuracy around 9 epochs. The train, validation and test split of data is 70%-15%-15% respectively. To perform model evaluation through a variety of metrics, the mask is decoded to establish the number of Ki-67 immunopositive and immunonegative cells. This is carried out in a reverse process to the JSON-to-mask encoding used in the ground-truth annotations (Section 3.1.1). Gaussian thresholding is applied to the mask and contours are extracted before the label for each cell is documented for use in the evaluation process. A copy of the model implementation can be found: `https://github.com/richardgault/Automated-Ki-67-proliferation-scoring`.

3.2 Mobile application and cloud hosting

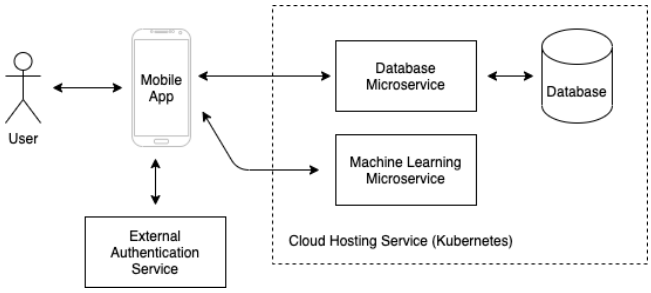


Figure 2: Illustration of the software architecture

To support end-user access a mobile application has been implemented which is available on both Android and iOS platforms. The application communicates to services, such as the machine learning platform which performs the analysis, through HTTP requests made over the Internet. The services themselves including the machine learning elements and database storage are packaged as microservice containers using the industry standard Docker which allows them to be easily run on any cloud hosting provider. For the purposes of our implementation they are deployed using a kubernetes hosting cluster. The services provide an Application Programming Interface (API) to which the application can connect and make requests to process, store, or retrieve data. The generalised architecture is shown in Figure 2. Such an approach allows easy remote access from any device with a network connection including cellular data and also offers the potential for other implementations beyond our application to make direct remote use of the hosted services, i.e. a third party could integrate our cloud based processing in to another platform for gathering images.

4 Results

4.1 Evaluation of model performance

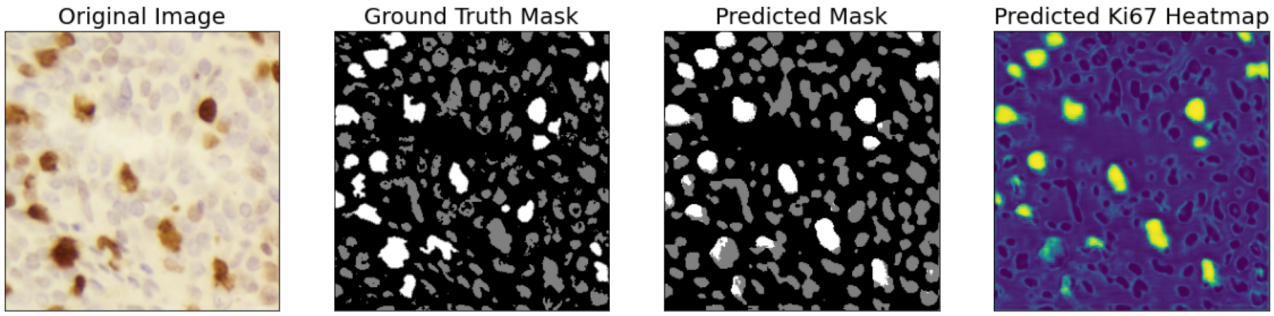


Figure 3: Example of the predicted mask and heat map relative to the raw image and ground truth mask.

Figure 3 shows an illustration of the predicted mask and heatmap produced by the model relative to the original image and its ground truth mask. Table 1 provides an overview of the training, validation and test performances of the model. The performance is consistently high and comparable across all datasets and measures. It is notable that the predicted mask generally estimates cells that were completely solid as desired whereas the ground truth mask sometimes included "background" (i.e. black) pixels in the middle of some of the immunonegative cells (shown in grey in the masks of Figure 3). This is likely an artifact introduced by the mask generation approach that was outlined in Section 3.1.1 when the light coloured portions of cells have been mislabelled as background following the adaptive Gaussian thresholding process. These mismatches between the ground truth and predicted masks result in reduced Intersection over Union (IoU) scores for these isolated regions despite the prediction being meaningful and appropriate. This is important to remember when considering the IoU score.

Table 1: Evaluation of Model performance

Dataset	Accuracy	Precision	Recall	f1-score	IoU
Train	0.9575	0.9371	0.9353	0.9362	0.8813
Validation	0.9584	0.9385	0.9367	0.9376	0.8901
Test	0.9609	0.9421	0.9403	0.9412	0.8961

Table 2: Comparison of model performance

Model	RMSE
Mod. DeepLabv3-Mobilenetv2	0.050
Mod.DeepLabv3-Xception	0.063
Mod. FCRN-A	0.067
Mod. FCRN-B	0.069
PathoNet	0.062
Proposed model	**0.045**

The Ki-67 index score is calculated as the number of immunopositive cells relative to the total number of cells in the sample. Previous studies [Negahbani et al., 2021] compared the performance of existing models in their ability to accurately calculate the Ki-67 index of a sample. These figures are presented in Table 2 alongside the model presented in this work. The results show that the proposed modelling approach has performed slightly better than the existing models. However, strong conclusions should not be drawn from this table as the test set used in the current analysis is likely to be different than that used in [Negahbani et al., 2021] despite using the same data; also, previous models were trained on 3-class classification (immunopositive vs immunonegative vs lymphocyte) which extends the binary classification considered in this work.

For diagnostic purposes and to inform prognosis, the Ki-67 index is categorised in to low (<10%), borderline (10-20%) and high (>20%). The model was able to predict the correct category with 88% accuracy. Table 3 shows the confusion matrix for the category prediction and Table 4 provides the associated metrics for the model's performance. It is clear from Tables 3 and 4 that the model performs particularly well in the extreme cases of low and high Ki-67 expression. The accuracy of the category prediction is lower than the accuracy of the individual cell prediction because misclassification of an individual cell has relatively little impact given the high number of cells the model considers. Each image contains multiple cell types and typically a large number of cells. Therefore the impact of misclassifying an image in to the correct category has a greater impact on the results since the total number of images is vastly less than the total number of cells, which the metrics in Table 1 consider. The results show that the model can accurately identify cells, classify the presence of Ki-67 expression and accurately categorise the Ki-67 index in to an appropriate category for diagnosis and prognosis.

Table 3: Confusion matrix for Ki-67 category

		Predicted		
		<10%	10-20%	>20%
Actual	<10%	270	25	2
	10-20%	9	54	6
	>20%	0	4	25

Table 4: Ki-67 Index metrics per category

Category	precision	recall	f1-score
<10%	0.968	0.909	0.937
10-20%	0.651	0.783	0.711
>20%	0.758	0.862	0.806

4.2 Evaluation of the Mobile Application

The mobile application is a multi-user system authentication to ensure access control. The user interface is designed with a consistent dark mode colour scheme to improve visual ergonomics by reducing user eye strain, adjusting brightness according to lighting conditions and facilitating screen use in dark environments – all while conserving battery power by reducing the use of light pixels. Intuitive and minimal functionality is presented to the user to support ease of use.

A new image can be taken using the mobile device camera or an existing one can be selected from storage (Figure 4 (a)). Multiple images can be selected at one time to enable batch analysis. The image is then displayed on the user interface (Figure 4 (b)) and the button to analyse the image is enabled. Once pressed, all images to be analysed are sent as an image stream in an HTTP POST multi-part form request to the machine learning microservice for analysis. The JSON response is returned to the mobile application, where it is then deserialized to obtain values such as the number of immunonegative cells, number of immunopositive cells and consequently the Ki-67 proliferation index as a percentage. The results are returned to the user nearly instantaneously when there is strong internet connectivity. Asynchronous communication handling of the mobile application means that normal operation of the mobile application is undisturbed if there is an unforeseen delay in the returning of results.

A predicted mask is also returned with the model predictions providing a visual result for transparency and

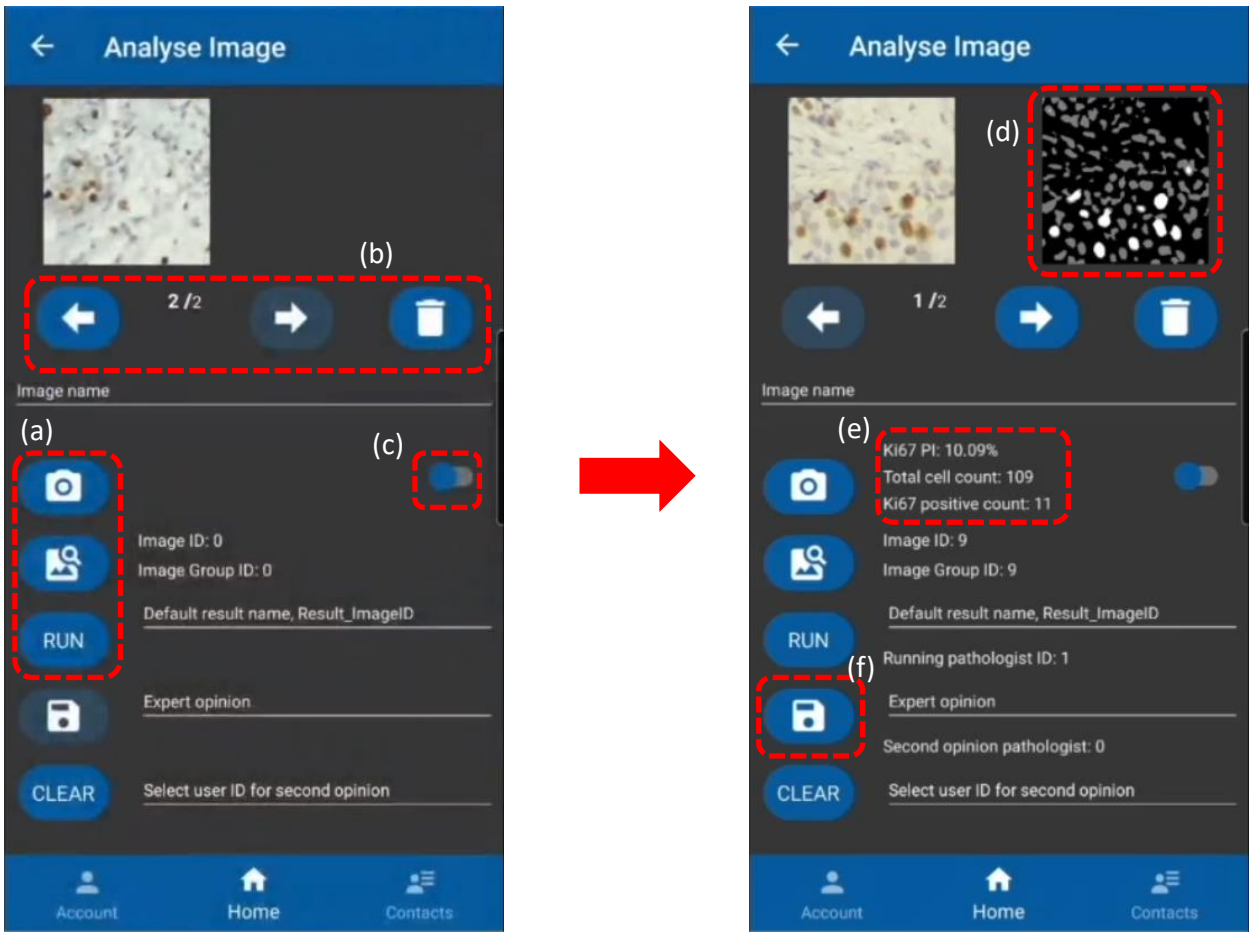


Figure 4: Screen shots of the mobile's analyse image functionality: (a) Images imported by taking a new image with the device's camera or from file (individually or as a batch). Images can be processed by selecting "RUN". (b) Batches of images can be previewed and removed from the user's selection as desired. (c) If the user wishes to first provide a blind review of the results they can toggle the result viewer to hide or show the modelling results. (d) Predicted masks are displayed along with (e) key information regarding the Ki-67 proliferation index. (f) The results can now be saved (function now enabled) for future records. Optionally, the image(s) can be reviewed by the current user and/or assigned to another user for a second opinion.

traceability of the model's decision making for the pathologist to examine as well as the Ki-67 statistics (Figure 4 (d) and (e) respectively). The application has been designed to facilitate both single and batch image analysis. Results can be saved to the database (Figure 4 (f)). The user running the analysis can assign the result to another user in their favourite contacts list (determined by the user), so that they can get a second opinion.

The mobile application is evaluated through automated testing. The NUnit framework (Version 3.12.0) is used for the view-model unit tests. Tests have been written using the Arrange, Act, Assert (AAA) pattern, which involves initialising objects and setting data values, invoking the method under test with the arranged parameters and verifying that it behaves as expected. Unit tests for the mobile application view models had 71% coverage using Rider IDE from JetBrains with all tests passing. Moq (Version 4.16.10) is used as a mocking framework to emulate responses from interfaces, which is done during the Arrange part of the tests. Some of the classes used in the code implementation do not inherit interfaces and therefore cannot be mocked. This meant that some of the code was not testable, but the code coverage shows that a majority of the logic of the mobile application code for the view models was still able to be tested. The fact that all tests passed verifies that the logic of the code functions successfully and as expected. Integration tests for checking the connection to each request handler were implemented and all successfully passed. A YAML file is used to configure GitLab Continuous Integration/Continuous Deployment. It allows the database microservice to be built and tested

automatically each time a code change is pushed to the codebase. This verified that the microservice could receive HTTP requests.

5 Discussion

The developed computational model for automated Ki-67 proliferation scoring is designed to improve analysis workflows in digital pathology by providing a proof-of-concept solution that can be developed further. The model is embedded in a computational system that allows users to intuitively and quickly run Ki-67 analysis through a mobile application. The cloud based computation enables scalability and utilisation of computational resources that would not be available locally on the mobile device.

The computational model has been able to correctly classify 96% of cells with very strong coincidence with the ground-truth as captured by the IoU score of 0.89. The quality of the predicted masks provides pathologists with confidence, transparency and traceability in the computational decision making process. Future work is needed to extend the model's exposure to more datasets collected from different international laboratories and exposure to other tumour types than the breast cancer samples considered in the present work.

The mobile application has the potential to make an impact in many communities, especially low and middle income countries where there are currently insufficient resources, namely expensive digital scanners, to support digital pathology in the fields of medicine and education. Microscope mounts for mobile devices are available that allow the device's camera to be positioned at the lens of the microscope. This enables the user to capture and analyse the field of view through the application. In terms of supporting low and middle income countries, the multi-user mobile application provides the baseline infrastructure for further development that make it feasible to connect global experts if they are using the platform. The authentication system implemented means that existing email and authentication methods are retained and the application can be easily scaled to a large audience. The second reviewer functionality implemented in the app provides the framework to connect pathologists and researchers with global experts in their designated area for further opinions. The cloud hosted analysis module also supports scalability and easy redeployment. There is also the potential for the system to address challenges of educating medical students in digital pathology that were highlighted in [Fontelo et al., 2012]. For example, the computational model could be used to test a student's annotating skills by comparing the model's predicted mask against the student's annotations and use the model's mask as feedback for the student in both a qualitative and quantitative way.

Future work for this system entails the refinement of the mobile application to enable enriched user features, such as institutional/company specific areas, and complete automated and user testing. Although the results of the Ki-67 prediction area high in all areas and the IoU of the predicted masks is also high, the computational model needs to be evaluated on independent datasets from other sources. Staining and imaging protocols vary around the world leading to diverse colour profiles and image quality. Consequently, further refinement will likely be needed to ensure the model is robust before the model is used in real-world applications.

6 Conclusion

The proof of concept mobile application presented in this work provides accurate identification of Ki-67 immunopositive and immunonegative cells in histopathology images. The model can accurately support a pathologist or researcher in the scoring of Ki-67 proliferation as evidenced through the high performance in individual cell classification, IoU score and KI-67 index categorisation. The computational system efficiently produces information that is transparent and essential in many standard digital pathology analysis pipelines. Extensions of this work could provide sufficient impact in low and middle income countries where state-of-the-art scanning and computational analysis resources are not widely available. This cloud based system paves the way for scalable and global solutions that could connect analysts with leading experts using the platform. Future work is planned to enhance the functionality of the mobile system and refine the computational modelling to enable its generalisation and robustness to variation across centres.

Acknowledgments

The authors would like to thank Farzin Negahbani (and colleagues from Shiraz Histopathological Imaging Dataset Center, Shiraz University of Medical Sciences) for providing the image dataset used in this work.

References

[Atta-Fosu et al., 2016] Atta-Fosu, T., Guo, W., Jeter, D., Mizutani, C. M., Stopczynski, N., and Sousa-Neves, R. (2016). 3d clumped cell segmentation using curvature based seeded watershed. *Journal of imaging*, 2(4):31.

[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. Ieee.

[Fontelo et al., 2012] Fontelo, P., Faustorilla, J., Gavino, A., and Marcelo, A. (2012). Digital pathology–implementation challenges in low-resource countries. *Analytical Cellular Pathology*, 35(1):31–36.

[Geread et al., 2019] Geread, R. S., Morreale, P., Dony, R. D., Brouwer, E., Wood, G. A., Androutsos, D., and Khademi, A. (2019). Ihc color histograms for unsupervised ki67 proliferation index calculation. *Frontiers in bioengineering and biotechnology*, 7:226.

[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.

[Hernández-Neuta et al., 2019] Hernández-Neuta, I., Neumann, F., Brightmeyer, J., Ba Tis, T., Madaboosi, N., Wei, Q., Ozcan, A., and Nilsson, M. (2019). Smartphone-based clinical diagnostics: towards democratization of evidence-based health care. *Journal of internal medicine*, 285(1):19–39.

[Inwald et al., 2013] Inwald, E., Klinkhammer-Schalke, M., Hofstädter, F., Zeman, F., Koller, M., Gerstenhauer, M., and Ortmann, O. (2013). Ki-67 is a prognostic parameter in breast cancer patients: results of a large population-based cohort of a cancer registry. *Breast cancer research and treatment*, 139(2):539–552.

[Joseph et al., 2019] Joseph, J., Roudier, M. P., Narayanan, P. L., Augulis, R., Ros, V. R., Pritchard, A., Gerrard, J., Laurinavicius, A., Harrington, E. A., Barrett, J. C., et al. (2019). Proliferation tumour marker network (ptm-net) for the identification of tumour region in ki67 stained breast cancer whole slide images. *Scientific reports*, 9(1):1–12.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Negahbani et al., 2021] Negahbani, F., Sabzi, R., Jahromi, B. P., Firouzabadi, D., Movahedi, F., Shirazi, M. K., Majidi, S., and Dehghanian, A. (2021). Pathonet introduced as a deep neural network backend for evaluation of ki-67 and tumor-infiltrating lymphocytes in breast cancer. *Scientific Reports*, 11(1):1–13.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Yerushalmi et al., 2010] Yerushalmi, R., Woods, R., Ravdin, P. M., Hayes, M. M., and Gelmon, K. A. (2010). Ki67 in breast cancer: prognostic and predictive potential. *The lancet oncology*, 11(2):174–183.

Identifying Pathological Facial Weakness using Fuzzy Inference

Victoria Porter*[1], Eliza Przewozniak[2], Richard Gault†[1], Mark McDonald[3], and Omar Uribe[3]

[1]*Queen's University, Belfast, UK*
[2]*Citi, Budapest, Hungary*
[3]*Syntrillo, INC, Delaware, USA*

Abstract

Stroke is the second largest cause of death and disability-adjusted life-years in the world. Minimising the time to treatment for patients is extremely important. Facial weakness is a core symptom that medical professionals consider when identifying cases of stroke. This is a subjective assessment of asymmetry in the face. Due to this subjectivity, it is challenging to articulate the decision making process of a neurologist. This work presents novel computational approaches to accurately model the detection of pathological facial weakness from images of people with and without pathological facial weakness. Instance segmentation is first used to isolate key facial features that inform the decision making of a fuzzy inference system. This proof of concept study shows the feasibility of automated feature extraction and the effectiveness of fuzzy inference systems in identifying facial weakness. Furthermore, the transparent nature of the instance segmentation model and the fuzzy rule base has enabled the model to be compared against the real-world decision-making process of a neurologist. The findings motivate future investigations to develop fuzzy inference systems to detect other common deficits of stroke including limb weakness and drift as well as dysarthria.

Keywords: Stroke, facial weakness, instance segmentation, fuzzy inference

1 Introduction

In the most recent Global Burden of Diseases, Injuries, and Risk Factors Study (GBD), stroke was the 2nd largest cause of death globally and contributed to the 2nd highest disability-adjusted life-years (DALYs) worldwide [Johnson et al., 2019]. Although the age-standardised rate of deaths from stroke have declined since 1990, treatment and post-stroke care is a significant economic burden with the treatment cost for severe stroke being more that double that required for a mild stroke [Roger et al., 2011]. It is anticipated to have an even greater global impact with the ageing population [Katan and Luft, 2018]. Ideally, the number and severity of cases would be managed by preventative measures however the high burden of stroke globally suggests that such strategies are not widely used or are ineffective. Consequently, there is a need to optimize the treatment of acute stroke in order to reduce disability and cost.

Stroke therapy can be highly effective if administered early. Reducing the time to treatment by as little as fifteen minutes can save one month of disability-free life [Meretoja et al., 2014]. This time-sensitive nature of treatment efficacy is perhaps best expressed in the neurologists' adage "time is brain". Unfortunately, 1/3 of patients have a significant delay in treatment [Kamal et al., 2017]. A major contributing factor to this delay is the failure of patients, first-responders, and paramedics to recognize signs of stroke [Lachkhem et al., 2018]. Stroke signs are patterns of neurological function used by both medical professionals and the general public to identify stroke. Some signs can be easily identified by most people while other signs are difficult to

*vporter03@qub.ac.uk
†richard.gault@qub.ac.uk

detect without extensive training [Brandler et al., 2014]. The three features most commonly used to identify stroke in the field are unilateral or one-sided facial weakness, unilateral arm weakness, and slurred speech [Hurwitz et al., 2005]. Brandler et al. [2014] found that with simple instructions untrained individuals recognized over 95% of cases of arm weakness and slurred speech. In contrast, only 74% of facial weakness cases were correctly detected. This discrepancy may be due to the fact that people without any neurological disease often have some degree of facial asymmetry at rest and while smiling. Learning the proper threshold of normal vs abnormal facial asymmetry likely requires specialized training. Similarly, paramedics were found to correctly classify facial weakness only 82% of the time [Nor et al., 2004]. Even among physicians with neurologic training, the inter-rater reliability for certain components of the neurologic exam is less than 60% [Hansen et al., 1994], and thus detection of common signs of stroke in the field varies substantially [Nor et al., 2004, Josephson et al., 2006, Meyer and Lyden, 2009]. Both facial weakness and dysarthria have poor inter-rater reliability [Josephson et al., 2006, Meyer and Lyden, 2009]. Unfortunately, such clinical expertise is scarce and often not immediately accessible at the time that a stroke occurs. As a result, many patients either fail to receive timely treatment or are ineligible for acute stroke therapy.

A technology that can accurately detect abnormal facial asymmetry could support the diagnostic capabilities of non-experts and reduce delays in stroke treatment. Two major challenges of identifying pathological facial weakness are that 1) the decision-making process is subjective and 2) the categories of "normal" and "abnormal" are ambiguous. To deal with these challenges, we propose implementing a fuzzy inference system (FIS) that specialises in handling vague or imprecise inputs. Additionally, the fuzzy rule base can be interpreted by humans, allowing for direct comparison with human decision making. This work will investigate whether a FIS can be used to detect the presence or absence of pathological facial weakness.

Section 2 will outline the models considered and Section 3 of this paper will provide an outline of the datasets used in this investigation. Section 4 will provide an outline of the results which are discussed in more detail in Section 5. A final conclusion and summary of this investigation is provided in Section 6.

2 Methods

The proposed system will automatically extract facial features using instance segmentation and then predict the presence or absence of facial weakness. Details of each stage are presented in Sections 2.1 and 2.2 respectively.

2.1 Instance Segmentation of Facial Features

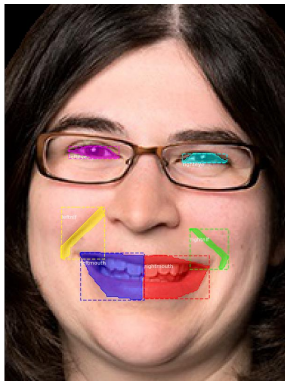


Figure 1: Example annotation of the ROI.

Three facial regions widely used by specialists when rating facial weakness, namely the mouth, eye and nasolabial fold (NLF) (Figure 1), are considered. The left and right portions of each region are considered independently resulting in 6 regions of interest (ROI) for each person. Mask R-CNN is used as the basis of the instance segmentation of the facial features [He et al., 2017]. The model is trained through transfer learning using the datasets outlined in Section 3 with a base model of ResNet101 [He et al., 2016] using the pretrained weights derived from training with the MS COCO dataset [Lin et al., 2014]. The implementation was conducted in Google Colab using Python 3.7, a single Tesla K80 GPU, learning rate of 0.001, a batch size of 2, a minimum detection confidence of 0.9 and 150 training epochs. The goal of the instance segmentation component of the system is to accurately extract the facial features ROI to aid with the subsequent classification of whether facial weakness is present or not present. When more than one instance of a region is predicted, only the predicted region with the highest confidence value is considered as the model's prediction since all images should have only one of each region. The model's performance will be evaluated using the Intersection over Union (IoU) metric.

In this investigation the aim is to detect facial weakness in a binary manner (weakness vs no weakness) in keeping with the Cincinnati Prehospital Stroke Scale (CPSS) [Kothari et al., 1999]. When facial weakness is present it is anticipated that there will be asymmetry in the facial structure in some or all of the ROI. The degree of facial symmetry is captured by the ratio (R_E, R_N, R_M) of the number of pixels in the corresponding left and right regions for the eyes, NLFs and mouth by Equation 1,where P_{X_l} and P_{X_r} are the number of pixels in the regions on the left and right side respectively and $X \in \{E, N, M\}$. The ratio is bound in the range [0,1] with perfect symmetry having a value of 1 and complete asymmetry being 0. The eye, NLF and mouth ratios are fed in to the decision making component of the system.

$$R_X = \frac{\min(P_{X_l}, P_{X_r})}{\max(P_{X_l}, P_{X_r})} \quad (1)$$

2.2 Detecting facial weakness through Fuzzy Inference

A Mamdani style FIS is implemented in Python using the skfuzzy package to model the decision making of neurologists for facial weakness detection. The membership functions and fuzzy rules are prescribed by the investigator and the resultant model referred to as the Prescribed Fuzzy Inference System (PFIS). For each input ratio (eye, NLF and mouth) the model has three membership functions representing "low", "medium" and "high" symmetry. The "medium" (or Med) membership function for each input is a Gaussian centred around 0.76, 0.55 and 0.63 for the eye, NLF and mouth inputs respectively and with a 0.8 standard deviation. The "low" and "high" membership functions are trapezoidal in nature and are respectively maximum (1) at 1 standard deviation below and above the Gaussian centred points mentioned above. The two output variables represent the presence (1) or absence (0) of weakness. The consequent fuzzy sets have Sigmoid membership functions that peaked at 0 and 1 respectively. These parameters for the FIS were prescribed heuristically. The Fuzzy rule base consists of 10 rules. A single rule used OR logic to specify that "low" symmetry in any of the inputs implies there is weakness present. All remaining 9 rules are detailed in Table 1 and use AND logic to combine the membership functions for the inputs R_E, R_N and R_M corresponding to the eye, mouth and NLF ratios respectively.

Table 1: AND rules from the PFIS

MF R_E	***MF*** R_N	***MF*** R_M	***Weakness***
Low	Low	Low	1
Med	Med	Med	1
High	High	High	0
High	High	Med	0
High	Med	High	0
High	Med	Med	0
Med	High	High	0
Med	High	Med	0
Med	Med	High	0

The PFIS is benchmarked against a Multi-layer perceptron (MLP) with 10 neurons in the hidden layer and trained for 1000 epochs with a batch size of 200 using binary cross-enthropy as a loss function and the Adam Optimiser. The outputs of the PFIS and MLP are rounded to the nearest integer to obtain a classification prediction of weakness (1) or no weakness (0).

3 Dataset

Two datasets are used in this study; the FEI face database and a facial weakness dataset. Both datasets contain images of individual faces. The ground truth masks for the six ROI were annotated using the VGG Annotator Tool [Dutta et al., 2016, Dutta and Zisserman, 2019]. All annotations were conducted by someone with no clinical experience to ensure that unconscious bias was not introduced in images where facial weakness was present. The annotator was advised where each region was located in a sample of images and completed annotations were reviewed for quality control. Each dataset and it's usage in this work will be detailed in Sections 3.1 and 3.2.

3.1 FEI face database

The full FEI database contains 200 individuals photographed at various facial angles, expressions and contrast [do Amaral and Thomaz, 2008]. Only the 200 fully front facing images are used as faces turned to one side often have a restricted view of the ROI leading to an apparent facial asymmetry without there actually being

asymmetry present. Given the potential use case for this system to detect facial weakness with a first responder it is a fair pre-condition to request the model to only make decisions when the person is facing straight on in the picture. This dataset is used to train the Mask R-CNN. To increase the quantity and diversity of the training dataset, augmentations were applied to the images. All images were converted to greyscale and shifted a random amount of pixels in the vertical and horizontal direction and 98% of the time an additional augmentation was applied to the image quality (blurring, contrast adjustment, etc). All augmentations were carried out using the imgaug Python package.

3.2 Weakness Dataset

The facial weakness dataset contains 203 open-sourced images containing faces with possible signs of weakness and no signs of weakness. The precise pathology of the cases of apparent facial weakness are not known but likely contain cases of stroke as well as other conditions with similar symptoms, e.g. Bell's Palsy. The people in the images vary in ethnicity, age, and sex. Three board-certified neurologists were asked to blindly rate the images from 1 (very likely no weakness is present) through to 5 (very likely weakness is present). The ground truth is taken to be the modal score of the three experts. From these ratings images were grouped in to no weakness (<3), weakness (>3) or in determinant (=3 or split decision). Two images were categorised as in determinant and were omitted from the remainder of the analysis. In total, 90 images were classed as no weakness and 111 as weakness. Somes images had to be removed from the analysis as ROIs were obscured by hair or facial angle. This left 80 no weakness images and 107 weakness images (47 left sided, 60 right sided). This dataset is used to test the instance segmentation model and the PFIS. Additionally, 80% of the ground truth masks are used to calculate the symmetry ratios (Equation 1) for each region and subsequently train the MLP in the decision making task. The remaining 20% of images are used for testing the MLP with classes evenly balanced in the train/test split.

4 Experimentation and Results

A number of experiments are proposed to evaluate each component of the system (instance segmentation and decision making) as well as the combined system as a whole. In this section each experiment will be outlined and the results presented immediately thereafter.

4.1 Experiment 1:Analysis of Instance Segmentation

Transfer learning is applied to the Mask R-CNN model described in Section 2.1 using the 200 "healthy" images from the FEI dataset. The model is then evaluated using the Facial weakness dataset. Note that the model has only been trained on the ROI (eyes, NLFs, mouth) and has not previously been exposed to cases where the regions are asymmetric.

Table 2 shows the IoU for the facial weakness dataset when considering the images with no weakness, and images with weakness independently. Mask R-CNN was able to accurately identify the eye and mouth regions in cases with no weakness but had difficulty distinguishing the NLFs in all cases. The model particularly struggled in identifying the mouth regions when weakness is present. This is likely caused by a distortion in the shape of the mouth when weakness is present, which is significantly different from samples in the training dataset. However, the IoU is not a perfect evaluation on the usefulness of the Mask R-CNN in identifying facial features. An illustration of the strong predictive mask can be seen in Figure 2 (a). The ROI, in particular the NLF, is not always clearly defined. Indeed in the case of Figure 2 (b) the NLF predicted mask only partially overlaps with the ground truth mask. However, on closer inspection the predicted NLFs are meaningful and may be in agreement with a different annotator. Future studies should consider multi-annotators and inter-rater variability to get a more robust ground truth definition.

Table 2: IoU results for Mask R-CNN in each region of interest

Data	Left Eye	Right Eye	Left NLF	Right NLF	Left Mouth	Right Mouth
Images: No weakness	0.74	0.73	0.54	0.54	0.77	0.74
Images: Right weakness	0.71	0.73	0.55	0.50	0.55	0.57
Images: Left weakness	0.71	0.69	0.49	0.54	0.55	0.53

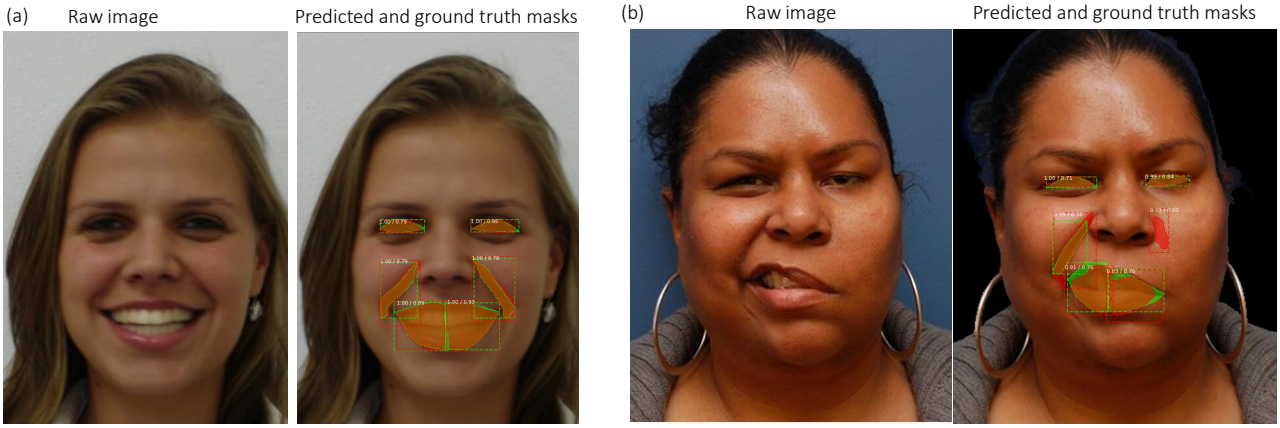


Figure 2: Illustration of the Mask R-CNN's mask prediction (red) compared with the ground truth (green) in (a) a case of no weakness and (b) a case of weakness.

4.2 Experiment 2: Detecting facial weakness using PFIS

The ability of the PFIS to accurately classify facial weakness is carried out in two stages. Firstly, the PFIS will receive inputs from the Mask R-CNN output from Experiment 1 and will make a prediction for each image. Secondly, to achieve a true analysis of the PFIS' ability to act as a decision making system without inheriting errors from Mask R-CNN, the ground truth masks for the facial weakness dataset are used to derive the symmetric ratios (Equation 1) for the three input regions and the performance of the PFIS is evaluated. In both cases the PFIS will aim to classify an image as either having weakness or no weakness.

The PFIS predicted all cases perfectly when making the decision based on ground truth masks. When using the predicted masks from the Mask R-CNN, the model was 79% accurate and had a precision of 83%, sensitivity of 89% and specificity of 75%. The results show that the PFIS is completely effective in its decision making when supplied with accurate features. Even though the Mask R-CNN provided imperfect features the PFIS was still able to achieve a very strong performance as sometimes this imperfection resulted in a desirable outcome. Given the significance of the facial weakness decision being made it is important that good fortune is not relied upon to make the decision and thus the results further motivate the need to enhance the instance segmentation component of the system.

4.3 Experiment 3: Detecting facial weakness using MLP

The MLP is used as a data-driven alternative to the PFIS and will be evaluated in an analogous setup to Experiment 2. The MLP obtained an accuracy of 92%, a precision of 100%, sensitivity of 90% and specificity of 100% when making a decision based on the ground truth masks while it had an accuracy of 83%, a precision of 100%, sensitivity of 81% and specificity of 100% when making a decision based on the predicted masks from the Mask R-CNN. The results show that the PFIS performs better when using the ground truth masks to inform the input data. Furthermore, when predictive masks are used, the PFIS will tend to be over cautious in its prediction compared with the MLP that would tend to have more false-negatives than false-positives. In the context of identifying potential facial weakness it is more desirable to favour the PFIS which is more likely to highlight potential cases for further consideration than overlook cases of facial weakness.

4.4 Experiment 4:Explainability of the PFIS

Unlike the black box decision making of the MLP, the PFIS contains semantically meaningful information and interpretable fuzzy rules. This experiment aims to evaluate the degree of agreement between the PFIS and an experienced neurologist to ascertain whether or not the PFIS encodes the decision making of the expert. An experienced neurologist was asked to comment on the relevance of the membership functions and inference rules to clinical decision making. A sample of images with the 5 highest membership values to the membership functions “low”, “medium” and “high” in each of the regions were blindly presented to the neurologist. The specialist was asked to rate the appropriateness of a number of statements describing each feature on a scale from 0-100 which took the format: “To what degree would you agree that the symmetry of the [eyes/mouth/NLF] is [high/medium/low]?”. This value was scaled to the unit interval $[0, 1]$ for comparison with the membership value of each input. This value will be referred to as the descriptor agreement value. Additionally, the specialist was asked to comment on the relevance of each of the inference rules from the PFIS to the real world decision making process. This qualitative analysis is designed to see how comparable the PFIS model is to the subjective decision making process.

In general, there is a significant difference between the descriptor agreement values and the corresponding membership values (Figure 3). However, in all cases, the specialist’s agreement values are closest to the membership values in the extreme cases when there is either “high” symmetry or “low” symmetry in the image. As the notion of “medium” symmetry is not well defined it is unsurprising that in each region the expert’s opinion of medium symmetry is most different from the membership values to the fuzzy set “medium”.

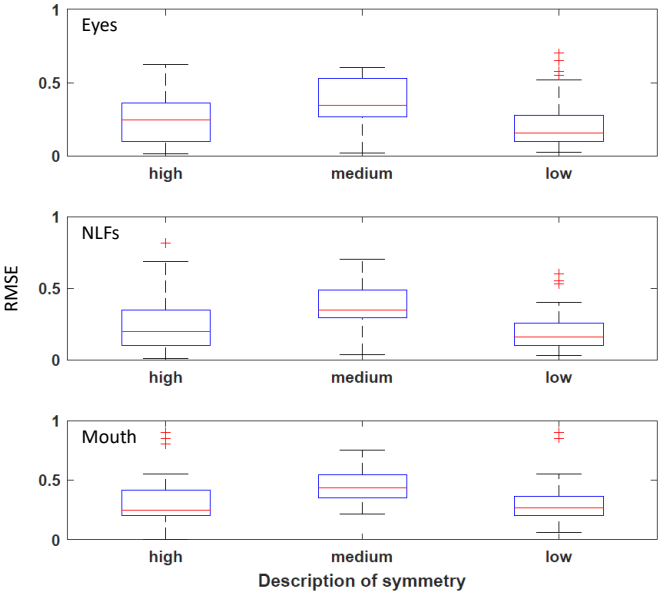


Figure 3: Spread of the RMSE between the descriptor agreement values and the values of the membership functions for the eyes, NLFs, and mouth of the PFIS.

The neurologist commented that the inference rules used in the PFIS are clinically relevant. It is possible however that the rule “high eye symmetry" and “high mouth symmetry" and “moderate NLF symmetry" implies no weakness, may not be applicable in some cases. In particular, the expert noted that when making decisions about weakness in the upper portion of the face that eye symmetry is comprised of both the size of the eye opening and eye brow elevation; the latter is not accounted for in this current study. Future work will investigate how to extract features for eye brow elevation from facial images and their inclusion in the decision making process for facial weakness.

5 Discussion

This work presents a proof-of-concept feature extraction and decision making system for the identification of pathological facial weakness. The results show that facial weakness can be accurately identified by the PFIS; particularly when accurate feature information is used in the decision making process. The findings of this work are relevant to screening tools for stroke such as the CPSS. One significant strength of the PFIS is the transparency of the decision making process. The neurologist noted that it is clinically relevant to characterise symmetry in the eyes, NLF and mouth, into three distinct categories that align with the membership functions used in the PFIS. The classification of “weakness" vs “no weakness" is clinically described as “abnormal" or “normal". Moreover, neurologists naturally interpret images in terms of the degree of asymmetry rather than the degree of symmetry as used in this work. Finally, the three descriptive categories/membership functions of “low", “medium” and “high" symmetry would traditionally be referred to as “mild", “normal" and “severe" in

relation to asymmetry. The canonical isomorphism between these sets of naming conventions mean the findings would hold if labelling were changed.

In comparison to a similar study [Zhuang et al., 2018] which used penalized Linear Discriminant Analysis (pLDA) to classify the input images as normal, left-sided weakness or right-sided weakness, the models presented in this paper appear to outperform the previous findings. In the present paper only 3 input parameters are used in the decision making compared with the 68 coordinates of facial landmarks used in [Zhuang et al., 2018]. The input features in this work rely on the pixel area of each predicted mask. Although this resulted in good results, it is a primitive and course approach to calculate asymmetry. A more robust approach would be to take in to consideration the shape of the predicted masks as well as the area. A more sophisticated approach to measuring facial symmetry will be considered in future work. As this previous study considered 3 class classification it is not appropriate to do a direct comparison with the models in this work. The distinction between left and right sided weakness is omitted in this work as it is not required for commonly used stroke screening tools but could provide an interesting line of investigation in future studies.

The precise pathology of facial weakness is not known in the open source image set and it likely contains more conditions than just stroke; such as Bell's Palsy. In particular, upper facial weakness, including eye symmetry, is only applicable in a small number of stroke cases. Future work will aim to refine the modelling approach presented in this work to facilitate the distinction between cases of stroke and other conditions that mimic stroke. Such investigations would require the collection of novel data including a comprehensive neurological examination to determine the pathology of the facial weakness. The images used in the facial weakness dataset contain some heterogeneity in terms of sex, age, ethnicity and facial weakness. The images also vary by dimension and background scenes. Moreover, the faces in the images differed in size as well as mild orientation differences. To implement the detection of facial weakness in a clinically relevant setting there are a number of factors that also need to be considered. The models have not been exposed to images where the person is lying down. In this scenario the facial features may become sunken backward and become more challenging to distinguish. Future work is required to increase the quantity, diversity and realism of the training and test datasets.

6 Conclusion

The extraction of facial features using Mask R-CNN and the subsequent detection of facial weakness by the PFIS shows that it is feasible to accurately automate the detection of facial weakness. This decision making is transparent, traceable and aligns with current clinical practices and expertise. The findings from this work show that Mask R-CNN and the PFIS could be effective methods for identifying facial weakness relevant for the early detection of stroke. Future work will investigate the use of fuzzy modelling with image and video data of patients with known stroke.

Acknowledgements

This work was partly funded through a PricewaterhouseCoopers (PwC) degree apprenticeship and the IAESTE program.

References

[Brandler et al., 2014] Brandler, E. S., Sharma, M., Sinert, R. H., and Levine, S. R. (2014). Prehospital stroke scales in urban environments: a systematic review. *Neurology*, 82(24):2241–2249.

[do Amaral and Thomaz, 2008] do Amaral, V. and Thomaz, C. E. (2008). Normalizacao espacial de imagens frontais de face.

[Dutta et al., 2016] Dutta, A., Gupta, A., and Zissermann, A. (2016). VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/. Version: 2.0.10, Accessed: 01/09/2020.

[Dutta and Zisserman, 2019] Dutta, A. and Zisserman, A. (2019). The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. ACM.

[Hansen et al., 1994] Hansen, M., Sindrup, S., Christensen, P., Olsen, N., Kristensen, O., and Friis, M. (1994). Interobserver variation in the evaluation of neurological signs: observer dependent factors. *Acta Neurologica Scandinavica*, 90(3):145–149.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

[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.

[Hurwitz et al., 2005] Hurwitz, A. S., Brice, J. H., Overby, B. A., and Evenson, K. R. (2005). Directed use of the cincinnati prehospital stroke scale by laypersons. *Prehospital Emergency Care*, 9(3):292–296.

[Johnson et al., 2019] Johnson, C. O. et al. (2019). Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(5):439–458.

[Josephson et al., 2006] Josephson, S. A., Hills, N. K., and Johnston, S. C. (2006). Nih stroke scale reliability in ratings from a large sample of clinicians. *Cerebrovascular diseases*, 22(5-6):389–395.

[Kamal et al., 2017] Kamal, N. et al. (2017). Delays in door-to-needle times and their impact on treatment time and outcomes in get with the guidelines-stroke. *Stroke*, 48(4):946–954.

[Katan and Luft, 2018] Katan, M. and Luft, A. (2018). Global burden of stroke. *Seminars in neurology*, 38(2):208–211.

[Kothari et al., 1999] Kothari, R. U., Pancioli, A., Liu, T., Brott, T., and Broderick, J. (1999). Cincinnati prehospital stroke scale: reproducibility and validity. *Annals of emergency medicine*, 33(4):373–378.

[Lachkhem et al., 2018] Lachkhem, Y., Rican, S., and Minvielle, É. (2018). Understanding delays in acute stroke care: a systematic review of reviews. *The European Journal of Public Health*, 28(3):426–433.

[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. *European conference on computer vision*, pages 740–755.

[Meretoja et al., 2014] Meretoja, A., Keshtkaran, M., Saver, J. L., Tatlisumak, T., Parsons, M. W., Kaste, M., Davis, S. M., Donnan, G. A., and Churilov, L. (2014). Stroke thrombolysis: save a minute, save a day. *Stroke*, 45(4):1053–1058.

[Meyer and Lyden, 2009] Meyer, B. C. and Lyden, P. D. (2009). The modified national institutes of health stroke scale: its time has come. *International Journal of Stroke*, 4(4):267–273.

[Nor et al., 2004] Nor, A. M., McAllister, C., Louw, S., Dyker, A., Davis, M., Jenkinson, D., and Ford, G. (2004). Agreement between ambulance paramedic-and physician-recorded neurological signs with face arm speech test (fast) in acute stroke patients. *Stroke*, 35(6):1355–1359.

[Roger et al., 2011] Roger, V. L. et al. (2011). Heart disease and stroke statistics—2011 update: a report from the american heart association. *Circulation*, 123(4):e18–e209.

[Zhuang et al., 2018] Zhuang, Y., Uribe, O., Mcdonald, M., Lin, I., Arteaga, D., Dalrymple, W., Worrall, B., Southerland, A., and Rohde, G. (2018). Pathological facial weakness detection using computational image analysis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 261–264. IEEE.

Video-Based Hand Pose Estimation for Abnormal Behaviour Detection

Fiona Marshall, Shuai Zhang and Bryan Scotney

School of Computing, Ulster University

Abstract

Hand gesture recognition, using skeletal hand keypoints estimated from depth sensors, is an active field of research. Unfortunately, many other potential hand keypoint applications are precluded by the limited range of depth sensors offering accurate hand pose estimation. Video-based hand pose estimation offers the potential to provide non-intrusive monitoring of hand movements. However, due to occlusions and the complex structure of the hand, hand poses predicted from video frequently contain many erroneous keypoints, hampering the detection and recognition of hand movements. We propose a method to significantly improve the detection of abnormal hand movements and introduce a novel video-based dataset containing normal and abnormal hand movements.

Keywords: activity recognition, skeletal joints, hand keypoints, data cleaning, hand pose estimation

1. Introduction

Video-based human activity recognition has been the focus of much interest over recent years and has applications in fields as diverse as security, sports, and healthcare. Skeletal keypoints, estimated from video data using algorithms such as OpenPose [Wei et al., 2016], can be used to create efficient activity recognition models. In addition to body keypoints, OpenPose can estimate hand keypoints [Simon et al., 2017], shown in Figure 1, opening new possibilities for the automatic recognition of hand movements. However, accurate hand pose estimation is generally more challenging than body pose estimation due to the frequent occlusions of parts of the hand and the complex skeletal structure of the hand, leading to the generation of inaccurate and often physically impossible hand poses. Hand keypoints can be hidden from a camera's view by other parts of the hand, body or objects held in the hand.

As OpenPose predicts keypoints independently for each frame in a video, there is a slight variation in the predicted location of keypoints between frames even when a person remains stationary. In extreme cases, mainly when a limb is partially occluded, the algorithm may erroneously generate sequences in which a stationary limb appears to move, as illustrated in Figure 2. However, erroneous finger joint keypoints movement is relatively common due to the size and occlusion of the phalanges. As fingers are frequently occluded during normal movement, unrealistic estimated hand poses and rapid inter-frame movement are common. Applying a temporal filter to smooth keypoint locations is an efficient way to remedy most erroneous limb movements [Han et al., 2017], but is ineffective for finger keypoints due to the frequency of erroneous predictions.

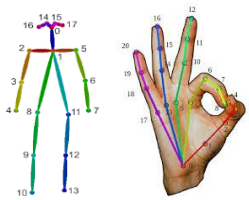


Figure 1: OpenPose estimates 18 body keypoints and 21 hand keypoints from video.

To date, most existing research on the automatic recognition of hand movements has been focused on the emblematic hand gestures used for human-computer interaction or sign language interpretation [Cheng et al., 2016]. In these cases, the subject typically provides precise gestures, helping to ensure that the hand pose is captured clearly, aiding automatic interpretation. Less explored are hand movements where clear hand signals are not provided, such as those observed when monitoring for abnormal or agitated behaviours within the home. We investigate whether keypoints extracted from video data, captured by an RGB camera situated at an unobtrusive distance from the subject, such as above a TV screen in

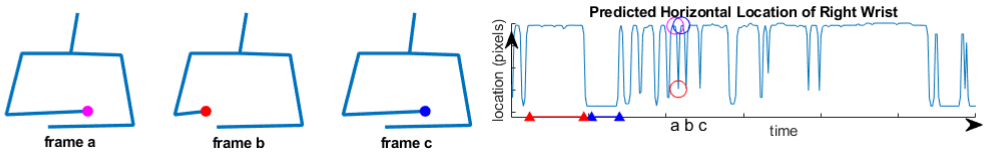


Figure 2: Estimated poses of a subject with folded arms, the incorrect poses create a false impression of movement.

the living room, can provide sufficiently accurate detail to detect abnormal hand movements. We explore whether including hand keypoints with body keypoints improves the detection of abnormal hand movements and considers approaches for cleaning noisy hand keypoint data. Finally, we present a novel video-based dataset of settled and agitated hand movements. The dataset, realistically unbalanced, has been collected in a simulated home environment contains mainly settled behaviour interspersed with a small number of short periods of agitated hand movement.

2. Related Research

Representing a person's movement by a sequence of skeletal keypoints is a widely researched activity recognition approach [Han et al., 2017]. Hand and body keypoints can be extracted from data captured by a 3D sensor or video camera. 3D sensors able to locate hand keypoints include the Leap Motion[1] Controller and Intel RealSense[2]. However, both 3D sensors, when used for hand tracking, have a range of less than 60cm, rendering them unsuitable for monitoring behaviour within a home environment. RGB cameras are widely available, non-intrusive, and provide a rich source of data. In addition to capturing the whole body, they can provide detailed information about hand movements. 2D body keypoints obtained from video have been shown to be as effective for activity recognition as 3D body keypoints [Marshall et al., 2019]. While algorithms that estimate body keypoints from video have been trained on massive datasets, there are no hand datasets with annotated keypoints of comparable size due to the difficulty of creating annotated or synthetic hand datasets. Instead, the OpenPose hand keypoint detector employs weakly supervised learning to train a hand keypoint detector using a small, labelled training dataset and a series of unlabelled images of a single hand from multiple views [Simon et al., 2017]. Similar to OpenPose but faster, Google's MediaPipe[3] also offers image-based hand pose extraction. However, as MediaPipe is still in the early stages of development and subject to changes, this study is based on OpenPose keypoints. Moreover, we expected occluded joints to present similar challenges across all hand pose estimation models.

Dynamic hand gestures consist of hand movements that evolve over multiple frames. Whilst some recognition approaches are based entirely on keypoint location [Nguyen et al., 2019], many approaches construct new low-level frame-based features. Handcrafted features include fingertip angle, the distance between fingers, motion, rotation and elevation from individual or multiple keypoints [Marin et al., 2014; De Smedt et al., 2017]. Traditional learning approaches reduce the sequence of frames to a single feature vector using Fisher Vectors, temporal pyramids [De Smedt et al., 2017] or neural networks [Nguyen et al., 2019]. Non-deep models used for classifying dynamic hand gestures include Support Vector Machines (SVM) [Nguyen et al., 2019; De Smedt et al., 2017] and Random Forest [Canavan et al., 2017]. Similar to whole-body activity recognition, deep dynamic hand movement recognition approaches include Convolutional Neural Networks [Devineau et al., 2018], Recurrent Neural Networks [Avola et al., 2019], and Graph Convolutional Networks [Li et al., 2019]. These deep learning approaches enable both temporal and spatial connections to be retained. Whilst successful for hand gesture classification, these approaches have been used only with hand keypoints captured by a 3D sensor close to the hand. Four widely used, publicly available, labelled 3D hand gesture datasets provide 3D depth images and keypoint locations of hand gestures: [Avola et al., 2019; Boulahia et al., 2017; Marin et al., 2014; De Smedt et al., 2017]. The datasets, created for hand gesture recognition tasks, contain precise,

[1] Leap Motion https://www.ultraleap.com/product/leap-motion-controller/
[2] Intel RealSense: https://www.intelrealsense.com/
[3] MediaPipe Hands: https://google.github.io/mediapipe/solutions/hands.html

predetermined hand gestures, where the hand is directly in front of the sensor. In contrast to previous research, our study focuses on recognising non-gesture hand movements symptomatic of agitation.

3. Data cleaning of estimated wrist and hand keypoint locations

Erroneous keypoint locations can result in a false prediction of movement. Figure 2 illustrates a situation where, although the subject is stationary with their arms folded, the predicted wrist locations in frames a-c suggest movement. Due to the high proportion of erroneous keypoints generated by the hand pose detection algorithm and the potential for long sequences of hand and wrist occlusion, sequences of wrist and hand keypoints that erroneously suggest movement are common and can be challenging to clean. For example, in Figure 2, it is unclear whether the true position of the wrist corresponds to the locations predicted between the red or the blue triangles. Three approaches are considered for identifying erroneous sequences of hand and wrist keypoints: two rules-based methods specific to sequences of hand poses and a generic outlier detection method.

For each frame, the OpenPose body keypoint detector predicts 18 body keypoints, whilst the hand keypoint detector predicts 21 hand keypoints (four keypoints per finger and thumb and one palm keypoint) for each hand. A confidence score is generated for each keypoint. For a single 2D frame, *f*, the location of the 60 joint locations, *j*, is denoted by $P_f = \{p_{f,0}, \dots, p_{f,59}\}$, in the original image coordinate space, where $p_{f,j} = \{x_{f,j}, y_{f,j}, c_{f,j}\}$ for the j^{th} keypoint, and c is a confidence score in the range [0,1]. Temporal sequences of keypoints over *n* frames are denoted $S = \{P_1, P_2, \dots P_n\}$. All body keypoint locations, except for the wrists, are discarded. In this study, video data is captured at 30 frames per second. All sequences are non-overlapping and of ten seconds (300 frames) duration.

As OpenPose detects hand and body pose independently for each frame, even non-occluded keypoint locations move slightly between frames. Therefore, instead of using every keypoint in a sequence to calculate movement, movement is considered over ten frames. Each ten-second sequence is split into 30 non-overlapping ten-frame windows from which the largest Euclidean distance between any two keypoints in each window is calculated. *Ten-frame displacement* is illustrated in Figure 5a.

$$\textit{ten-frame displacement of joint } j = \max\left(\sqrt{(x_{a,j} - x_{b,j})^2 + (y_{a,j} - y_{b,j})^2}\right), \forall \text{ frames } a \text{ and } b \text{ within the window} \quad [1]$$

Restraint of unrealistic palm or wrist movement: OpenPose estimates wrist locations using the body detector model, whilst the palm is estimated using the hand detection model. As the wrist and palm are closely linked physiologically, we would expect a strong correlation between their locations and movement. Large movement of the wrist when the palm is mainly stationary, or vice-versa, as illustrated in Figure 3, is unrealistic. When the sum of all 30 *ten-frame displacements* of a wrist sequence are greater than twice that of the palm, or vice-versa, unrealistic movement is suppressed by fixing the keypoint with the most movement to the initial frame location for the entire sequence.

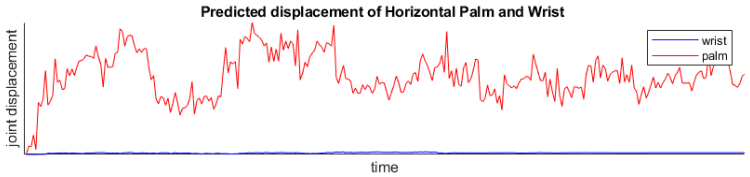


Figure 3:The difference between the extent of palm and wrist movement suggests erroneous keypoint predictions.

Replacement of unrealistic hand poses: unrealistic finger keypoint locations, such as those shown in Figure 4, are common due to the occlusion of fingers. The physiology of the hand is used to create criteria for a realistic hand pose. Each hand pose is checked to ensure that it fulfils all the criteria. Any poses considered to be unrealistic are removed and imputed using linear interpolation between the frames. The four categories of unrealistic hand poses are:

a. Crooked fingers: joints within a finger are bent in different directions, as shown in Figure 4a, where the blue and pink fingers are bent in different directions.
b. Joint angles with small angles: joints that close at an angle of less than 10^o are considered unrealistic. The red finger in Figure 4b has an unrealistic joint.

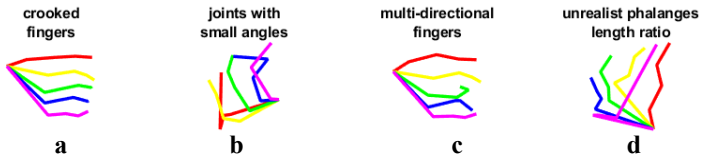


Figure 4: Four categories of unrealistic hand poses are identified.

c. Multi-directional fingers: the direction of adjacent fingertips changes more than once. Figure 4c illustrates a predicted hand pose where the directions of the yellow, green, blue, and pink fingertips alternate.
d. Unrealistic phalange length ratio: a phalange is more than three times the length of any corresponding phalange. The third pink phalange in figure 4d is disproportionately longer than those in the other fingers.

If a hand pose is unrealistic according to any criteria, all the keypoints from the unrealistic hand are removed and replaced using linear interpolation of the keypoints from adjacent frames. As dynamic information may be lost by interpolating missing data between the remaining poses, especially in sequences with many unrealistic hand poses, interpolated hand pose locations are adjusted to track wrist movement. If all hand poses in a sequence are considered unrealistic, the first pose is replicated for every frame of the sequence, suppressing all hand movement. Suppressing movement in this way is acceptable, as a moving hand is likely to be correctly estimated at some point in the sequence.

Outlier Detection: Outlier detection is widely used for signal data. Outliers are defined as keypoints more than three scaled median absolute deviations [MAD] from the median. The horizontal and vertical keypoint locations are considered separately. A sequence, $S_{j,d}$ of n frames of the j^{th} keypoint is denoted $S_{j,d} = \{q_{1,j} \dots, q_{n,j}\}$, where q can represent the horizonal or vertical dimensions of the point p.

$$\text{MAD} = (median(|S_{j,d} - median(S_{j,d})|) \quad for\ i = 1,2, \dots n \qquad [2]$$

Outlying hand and wrist keypoint locations are replaced using linear interpolation between adjacent frames.

A Savitzky-Golay filter (with polynomial order three and temporal window of length 7) is applied to each keypoint to smooth the keypoint locations. The Savitzky-Golay filter was selected due to its ability to smooth noisy signals with large frequency spans whilst maintaining the shape and height of the waveform peaks [Schafer, 2011], as is indicative of agitated movement.

4. Machine Learning Model and Feature Creation

Twelve displacement features are created from hand and wrist keypoints to describe the subject's movement during each ten-second sequence to detect abnormal movement. All features are normalised in the range [0,1].

Dominant hand: As abnormal movement can occur in either hand, features are created from a single dominant hand. The dominant hand is deemed the one that moves the most, calculated from the sum of all 30 *ten-frame displacements* of each wrist. Where there is only a small amount of wrist movement (the sum of all 30 *ten-frame displacements* is less than twice the distance between shoulder keypoints), the dominant hand is determined using the sum of *ten-frame displacements* of the thumb and the sum of *ten-frame displacements* of the mean fingertip locations.

Displacement of keypoints (8 features): Displacement features are created from the hand and wrist keypoints of the dominant hand. Three features are created from the thirty *ten-frame displacements* of the mean of the 21 hand locations in each frame: sum of *ten-frame displacements,* standard deviation of *ten-frame displacements,* largest *ten-frame displacement.* The *sum of ten-frame displacements* is also found for the wrist, thumb, fingertips, and palm keypoint location. A final feature, *300-frame displacement* for the entire 10-second sequence, is calculated from the largest Euclidean distance between any two keypoints in the sequence.

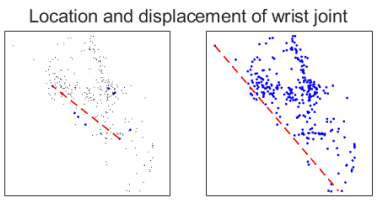


Figure 5. The maximum distance between keypoint locations over different periods of time can describe joint movement. a) ten frames and b) 300 frames.

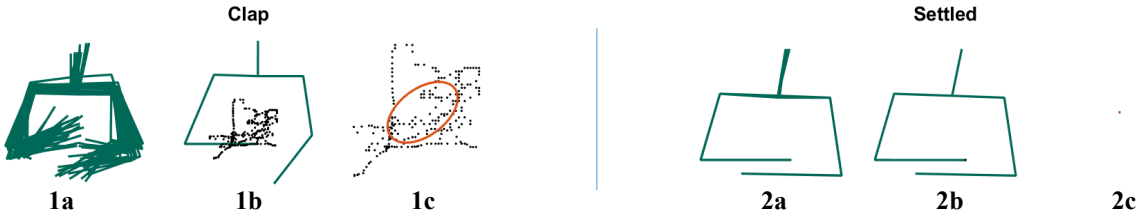


Figure 6. A covariance ellipse can be used to describe movement. Whilst the wrist movement of clapping is seen the large ellipse 1c (plotted in red), no movement, as in 2c, produces a tiny ellipse.

Figure 5 illustrates the keypoint locations of a wrist a) over a ten-frame window and b) over the entire ten-second (300 frames) sequence where wrist joint locations in each frame are shown with a blue point and displacement a red line.

Distribution of Keypoint Locations (3 features): Three covariance features are created from the wrist location, describing the distribution of keypoint locations within a sequence: the *horizontal variance of keypoint location*, the *vertical variance of keypoint location*, and the *direction of the eigenvectors*. The absolute value of direction is used to prevent differentiation between left- and right-handed movement. Figure 6 illustrates how covariance of keypoint locations can help detect movement. In Figures 6.1a and 6.2a, the movement of the limbs throughout the sequence is shown. In Figures 6.1b and 6.2b, only the wrist keypoints are used to describe movement. Figure 6.1c and 6.2c show how a covariance ellipse can describe the range and direction of movement of a keypoint. Where no movement is detected, such as in Figure 6.2c, the ellipse is represented by a single point.

Variation in Confidence Scores (1 feature): Occlusion of some hand keypoints is inevitable with movement. Therefore, confidence scores provided by the OpenPose algorithm can indicate hand movement, as changes in confidence may occur because of occlusions. Since a low confidence score indicates occlusion and a high score indicates non-occlusion, a confidence score that varies significantly throughout a sequence suggests hand movement. Movement is detected by locating peaks in the average confidence scores for the fingertips and thumbs, as shown in Figure 7. Peaks are found using the *findpeaks* MATLAB function (with prominence=0.4 determined empirically). The total number of peaks detected is used as a feature.

Classification: Support Vector Machines were used to classify the behaviours, as they are suitable for use with data containing imbalanced class distributions. Leave-one-out cross-subject validation was used. Accuracy was compared using F1 scores due to the imbalanced classes and the importance of false negative and false positive detections of abnormal behaviour.

$$\text{F1 score} = \frac{true\ positive}{true\ positive + \frac{1}{2}(false\ positive + false\ negative)} \quad [3]$$

5. Dataset

Whilst there is a small number of RGB hand gesture datasets, we are not aware of any datasets containing natural, non-gesture hand movements. To facilitate this study, we have collected a small dataset of settled behaviour, where subjects are relaxing whilst watching TV, interspersed with various repetitive hand movements. Eight healthy participants were recorded using an RGB video camera whilst seated in front of a TV for around 35 minutes. The camera was set above the TV, directly facing the participant at approximately two metres distance. Video images were captured at 30 frames per second. The setup is shown in Figure 8.

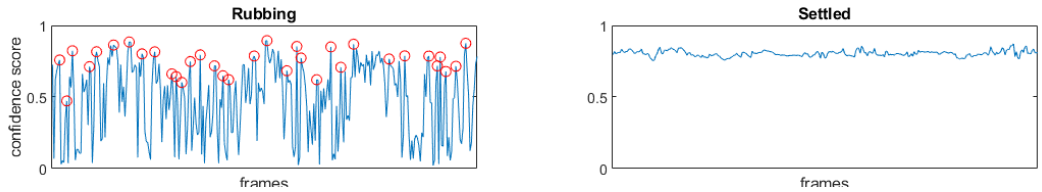


Figure 8: With a camera placed above the TV screen, the subject is recorded whilst seated, watching TV. OpenPose is used to estimate hand and body keypoints.

Participants were asked to remain seated throughout the entire recording. An audio alarm was used every three minutes to remind participants to demonstrate a repetitive hand movement for at least ten seconds. Five types of hand movements are studied: picking, scratching, rubbing, wringing, and clapping. The first four movements are indicative of common types of repetitive behaviours [Cullen et al., 2005] in people living with dementia. Clapping - a clearly defined repetitive action - is included as a benchmark activity. The participants decided which movement that they wanted to demonstrate on each occasion. For the remainder of the time, participants relaxed, seated in front of the TV. Whilst relaxed, although remaining seated, participants moved normally. Normal movements included rubbing the face, checking the phone, stretching, and blowing the nose. No participant sat completely still whilst relaxing. The study was approved by the relevant Ulster University Faculty Research Ethics Filter Committee.

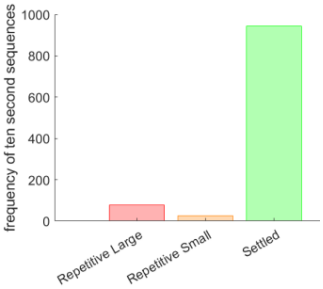
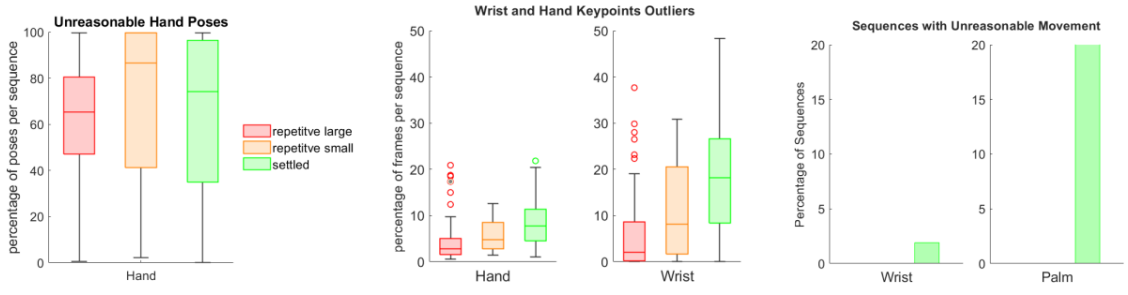


Figure 9: Distribution of types of behaviour in the dataset.

Estimation of body and hand keypoints: If a body keypoint is occluded, OpenPose returns a missing value for that keypoint. If a wrist keypoint is missing, all the keypoints from the corresponding hand will also be missing. However, if the wrist keypoint is predicted, OpenPose predicts all the corresponding hand keypoints, regardless of whether any hand keypoints are occluded. As parts of the hand are frequently occluded from camera view, OpenPose must regularly predict the location of unseen hand keypoints. Only hand and upper body keypoints were used in this study.

Data Annotation: Each frame of the video data was annotated manually by the first author. Movements were divided into four behavioural classes: *settled, normal, small repetitive movements*, and *large repetitive movements*. Small repetitive movements are finger-only movements, including using a mobile phone, changing television channels with remote control, rubbing fingers, and small picking movements. While the difference between types of behaviours is incremental, dividing the behaviours into different classes is necessary for detecting abnormal behaviours. As the subjects are recorded continuously, the data contains behaviours from the four classes and the transitions between classes. In this study, only sequences containing ten continuous seconds of the same behaviour are considered; the detection of abnormal hand movements from continuous data will be considered in a future study.

Keypoint location data are divided into ten-second (300 frames) sequences. Sequences of behaviour lasting less than 10 seconds were discarded. To ensure that sequences do not contain a mixture of behaviours, additional two



a **b** **c**

Figure 10: a-b) Proportion of hand poses replaced from each sequence. c) number of keypoint sequences replaced.

seconds of data was discarded each time a behaviour changed. A total of 1093 discrete ten-second sequences were created. As most normal behaviour, such as checking the time or rubbing the face, lasts less than ten seconds, only three ten-second sequences of normal behaviour were captured, which were also discarded due to insufficient observations. Additionally, as the dataset was collected to realistic movement suitable for recognising and detecting abnormal behaviour, most sequences contain settled behaviour, as shown in Figure 9.

6. Results and Evaluation

The effectiveness of replacing unrealistic hand poses, restraining unrealistic wrist and hand movements, and outlier detection approaches to cleaning noisy hand pose estimations were compared using the same features and classifier. Additionally, movement features were created from only the wrist keypoints estimated from the OpenPose body keypoint detection algorithm to discover whether including hand keypoint locations with body keypoints improves the model's ability to detect abnormal hand movements.

As shown in Figure 10, unrealistic hand poses were identified in sequences in all three types of behaviour, highlighting the importance of cleaning the hand pose data. However, unrealistic wrist or palm movement was identified only in the sequences of settled behaviour where there should not have been much movement. Outlier detection identified unrealistic keypoint locations for both hand and wrist joint sequences in all types of behaviour, although they were most frequently identified were in settled behaviour.

The F1 accuracies in Table 1 show the performance of different combinations of the keypoint cleaning approaches, highlighting that cleaning the hand keypoint data improves the ability of the model to recognise repetitive hand movements. Furthermore, the rules-based cleaning approaches can increase accuracy more than using outlier detection. The confusion matrices in Figure 11 show that whilst only 50% of the small repetitive hand movements were detected even after the keypoints had been cleaned, this is a ten-fold improvement on the number of keypoints detected for the uncleaned sequences. Combining outlier detection with the rules-based cleaning methods resulted in a lower detection accuracy than using only the rules-based methods, suggesting that outlier detection is less precise than the rules-based approaches. Additionally, the approach based only on wrist keypoints achieved a slightly higher accuracy than the approach which included uncleaned hand pose keypoints (64% compared to 62%), indicating that some cleaning of hand pose keypoints is essential if they are to be included in the model.

7. Conclusion and Future Research

Whilst the dataset used is small, this study demonstrates the potential for using a video-based approach to detect abnormal hand movements. We have shown that keypoints extracted from video data, when cleaned appropriately, can provide sufficient detail to recognise most repetitive hand movements. Despite the large numbers of unrealistic hand poses and keypoint sequences, including cleaned hand keypoints with the body keypoints improved the recognition of abnormal hand movements, especially for the small repetitive hand movements. Furthermore, we have demonstrated that whilst both generic and rules-based approaches to cleaning hand keypoint data can improve detection of abnormal hand movements, the rules-based approach resulted in higher accuracy. We plan to build on this study by creating a system that automatically detects abnormal, agitated hand movements from keypoints estimated from continuous video data. Furthermore, we aim to enlarge the dataset. A larger dataset would enable us to understand better the ability of our approach to generalise to new subjects, as well as to investigate the use of deep classification models

	Keypoint Cleaning Method							
	Hand and Wrist Keypoints (12 features)						Wrist keypoints only (7 features)	
Unrealistic poses replaced (UP)		✓			✓	✓		
Wrist and Palm Restrained (WP)			✓		✓	✓		
Outlier Detection (OD)				✓		✓		✓
Macro F1 accuracy	0.62	0.83	0.73	0.69	**0.84**	0.78	0.64	0.64

Table 1: Comparison of Results

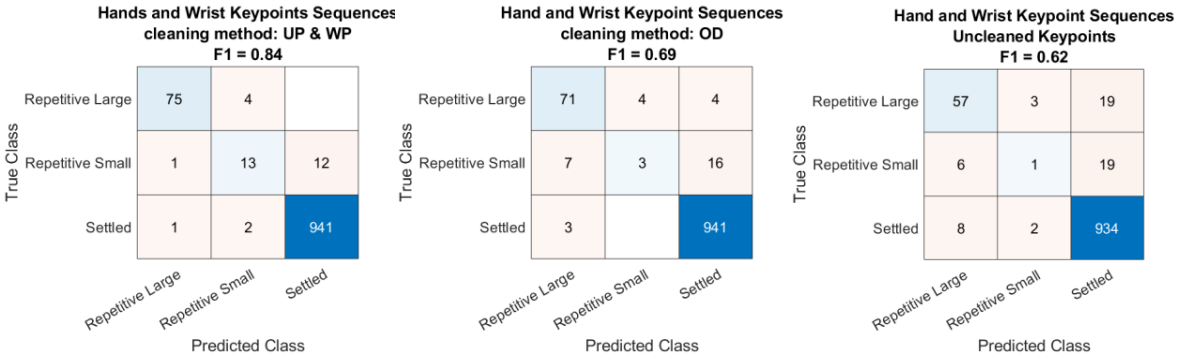


Figure 11: Confusion Matrices of results for different keypoint cleaning approaches.

8. References

[Avola et al., 2019] Avola, Danilo et al. 2019. "Exploiting Recurrent Neural Networks and Leap Motion Controller for the Recognition of Sign Language and Semaphoric Hand Gestures." *IEEE Transactions on Multimedia* 21(1): 234–45.

[Boulahia et al., 2017] Boulahia, Said Yacine et al. 2017. "Dynamic Hand Gesture Recognition Based on 3D Pattern Assembled Trajectories." *2016 23rd International Conference on Pattern Recognition (ICPR)*.

[Canavn et al., 2017] Canavan, Shaun et al. 2017. "Hand Gesture Recognition Using a Skeleton-Based Feature Representation with a Random Regression Forest." In *2017 IEEE International Conference on Image Processing (ICIP)*, , 1–5.

[Cheng et al., 2016] Cheng, Hong, Lu Yang, and Zicheng Liu. 2016. "Survey on 3D Hand Gesture Recognition." *IEEE Transactions on Circuits and Systems for Video Technology* 26(9): 1659–73.

[Cullen et al., 2005] Cullen, Breda et al. 2005. "Repetitive Behaviour in Alzheimer's Disease: Description, Correlates and Functions." *International Journal of Geriatric Psychiatry* 20(7): 686–93.

[Devineau, G. et al., 2018] Devineau, Guillaume, Fabien Moutarde, Wang Xi, and Jie Yang. 2018. "Deep Learning for Hand Gesture Recognition on Skeletal Data." *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*: 106–13.

[Han, et al., 2017] Han, Fei, Brian Reily, William Hoff, and Hao Zhang. 2017. "Space-Time Representation of People Based on 3D Skeletal Data: A Review." *Computer Vision and Image Understanding* 158: 85–105.

[Li et al., 2019] Li, Yong et al. 2019. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Dynamic Hand Gesture Recognition." *Eurasip Journal on Image and Video Processing* 2019(1).

[Marin et al., 2014] Marin, Guilo, Fabio Dominio, and Pietro Zanuttigh. 2014. "Hand Gesture Recognition with Leap Motion and Kinect Devices." *International Conference on Image Processing(ICIP)*: 1565–69.

[Marshall et al., 2019] Marshall, Fiona, Shuai Zhang, and Bryan Scotney. 2019. "Comparison of Activity Recognition Using 2D and 3D Skeletal Joint Data." In *Irish Machine Vision and Image Processing*, , p13-20.

[MediaPipe, n.d.] MediaPipe. "MediaPipe Hands." https://google.github.io/mediapipe/solutions/hands.

[Nguyen et al., 2019] Nguyen, Xuan Son, Luc Brun, Olivier Lézoray, and Sébastien Bougleux. 2019. "Skeleton-Based Hand Gesture Recognition by Learning SPD Matrices with Neural Networks." *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*.

[Schafer 2011] Schafer, Ronald W. 2011. "What Is a Savitzky-Golay Filter? [Lecture Notes]." *EEE Signal Processing Magazine* 28, n(July): 111–17.

[Simon et al., 2017] Simon, Tomas, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping." *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua: 4645–53.

[De Smedt et al., 2017] De Smedt, Q. et al. 2017. "SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset." *Eurographics Workshop on 3D Object Retrieval, EG 3DOR* 2017-April: 33–38.

[Wei,S et all., 2016] Wei, Shih En, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. "Convolutional Pose Machines." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem: 4724–32.

Billboard Detection in the Wild

Miss Sayali Avinash Chavan, Dr. Dermot Kerr, Prof. Sonya Coleman, Mr. Hussein Khader
Intelligent Systems Research Centre
School of Computing, Engineering & Intelligent Systems
University of Ulster
Northern Ireland, United Kingdom

Abstract

Advertising has a huge impact on modern life hence its analysis is very important. Billboard detection in a scene is very challenging given the outdoor position of billboards and the changing nature of a board's size, scale, and the angle at which it is viewed by oncoming traffic or pedestrians. Hence the requirement to detect the billboard and determine the visibility to consumers is a very difficult task. In this paper, we propose a system which will not only detect a billboard but also classify the different types of billboard panels. There exists a number of different types of billboard such as Street Furniture, Roadside, in-Mall, or Spectacular to name a few, however here we focus solely on Street Furniture and Roadside. For this, the Tensorflow object detection API is used with the Single Shot Multibox Detector (SSD) architecture. SSD is chosen because of its high-speed computation and ability to eliminate false-positive cases. In this paper, we demonstrate SSD's detection performance by fine-tuning hyperparameters and illustrate this using a dataset of billboards in the wild.

Keywords: Object Detection, Single Shot Detector (SSD), Deep Learning, Billboard, Image Processing

1. Introduction

Billboards are the only media which cannot escape the attention of the pedestrians and drivers of any moving vehicle. They are also considered a more reliable source of advertising than advertisements seen online [Borisova, O. and Martynova, A. 2017]. Advertisements influence a person's mind and this ultimately increases the profitability of the company who has placed them. Traditionally, outdoor impressions have been measured using traffic counts and daily circulations to calculate total reach and quantify consumer viewing of billboard advertising. More recently, with the advancement of Spatial and GIS information, along with SDKs and millions of applications, location based mobile data are widely used to measure audience impressions and give insight. While eye-tracking technology offers huge advancements in outdoor impression measurement, considering the high cost of equipment and wearables combined with the sample size, its use still poses difficulties around accurate measurement of billboard visibility. Visibility factors take the physical visibility characteristics of billboards into consideration which are independent of an individual's visual attention and are therefore more suited for obtaining an overall measurement of billboard visibility [Wilson, R. T. and Casper, J. 2016].

When dealing with detection of objects or structures outdoors or '*in the wild*', we need to consider factors such as the object location, the viewing angle of the billboard, whether the object is illuminated, if it is static or digital signage, if a car is in motion then what is the contact zone, occlusion by trees or many other factors. Visibility also can be obstructed by the weather conditions, glare from any light source, relative position of object and observer distance, brightness of background etc. All such factors, which can be considered as Visibility Adjustment Indices (VIAs), play an important role in determining the impact and value of an advertisement impression, which will affect the marketable value of a particular billboard at a particular location. For example, if a billboard is partially covered by a tree from the pedestrians on the other side of the road or it is not visible due to illumination issues, this could potentially reduce it's value to an advertiser. In the case of high billboard visibility such as placement of a unit beside traffic lights, it will most likely be viewed by drivers as well as pedestrians, and therefore will have

increased value [Wilson, R. T. and Casper, J. 2016]. Hence the VAIs play a key role in advertising billboard value. In this paper we present a dataset which contains static billboard images captured to include a range of these possible visual variations. We then determine the billboard's location by manually annotating a bounding box, and subsequently training a convolutional neural network to automatically recognise the billboard. Detection performance is evaluated and visual performance demonstrated.

2. State of Art

Computer vision is commonly used for understanding the external world through the use of algorithms. Common applications include understanding image data in order to identify and classify objects, object tracking, vehicle position monitoring, lane tracking and night time lane marking recognition [Li, Y. et al. 2016]. Much research has also focussed on developing algorithms to model human attention and saliency; saliency detection is an automatic process of locating the key parts of an image without any prior knowledge. Issues with current saliency approaches are that models use contrast and colour for low-level saliency cues [Krishna, O. and Aizawa, K. 2018].

Computer vision techniques have been previously used to specifically detect advertising boards. Such approaches have been based on Canny edge detection and morphological operations in order to determine the rectangular area of a billboard. However, many of these approaches used datasets containing minimal background noise and therefore are unable to generalise for the detection of unknown billboards in a range of dynamic scenes and environments [Rahmat, R.F. et al. 2019]. Another study demonstrated that to automatically detect regions that may be billboards, planar object detection can initially be used to locate and describe such objects. Planar object detection involves gathering individual object level information from an image then classifying which one of those objects is a billboard [Liang, P. et al. 2018]. In [Watve, A.K. and Sural, S. 2007] the focus was on the detection of advertising billboards on a soccer field, and this was achieved using the Fast Fourier Transform (FFT). The approach considered field detection, baseline detection, occlusion by players, image rectification, advertising board height detection and advertising board extraction using image data collected with a high resolution camera to provide the required clarity and high frequency colour intensities. The research in [Cai, G., Chen, L. and Li, J. 2003] focussed on advertising detection in sport TV using Hough transforms and geometric features of text to extract information from live high quality images and showed promising results. Another example is detecting advertisements from buildings in order to determine if they are in accordance with rules and regulations using computer vision techniques such as image rectification and segmentation in order to find the coordinates of the billboard object [Bochkarev, K. and Smirnov, E. 2019].

All these existing approaches are based on localisation and classification processes typically at pixel level. This process is very slow and has localisation problems when there are multiple objects in the scene. Additionally, considering real time use with low quality, blurry or occluded images, these methods have several limitations. There are various feature detectors available, however they are not capable of handling large amounts of data in real time applications. Hence Convolutional Neural Networks have become popular for working with large image datasets and high speed object detection and recognition [Shi, W., Bao, S. and Tan, D. 2019].

When neural networks are built using a number of convolutional layers in its model, these are known as Convolutional Neural Networks (CNNs). CNNs detect patterns in the image data and produce highly accurate predictions which are often measured as a percentage of correct classifications. Previous work using CNNs specifically for object detection has seen them applied to a wide range of applications ranging across medical applications, robotics, industry, wildlife detection, and geo-tagging. Most of the modern neural network architectures are derivatives of the famous ImageNet competition on supervised learning in computer vision in 2010, for example AlexNet, VGG, GoogLeNet, NiN, DenseNet and ResNet [Alom, M. Z. et al. 2018]. The best feature of a CNN is the capability to use transfer learning where the pre-trained model transfers the weights of its learned network to initiate the process of fine-tuning for another (unseen) dataset. There is a wide range of existing pre-

trained neural network models like YOLO, SSD MobileNet, Faster R-CNN ResNet, and R-FCN ResNet which use various open-source frameworks such as TensorFlow Object detection API, PyTorch, Microsoft cognitive toolkit, Keras, OpenCV, and DDN Library [Rahmat, R. F. et al. 2019]. The ADNet architecture is specifically designed to detect advertising instances from video frames and uses Microsoft COCO dataset to train its network [Hossari, M. et al. 2018]. One of the most recent examples of research based on billboard detection includes a comparative study of a SSD model vs YOLO (You Only Look Once) which showed promising results [Morera, Á. et al. 2020]. Hence we will utilise SSD in this study.

3. Methodology

The chosen methodology focusses on using a large dataset of annotated billboard images to train a convolutional neural network in order to recognise different classes of billboard. This section describes the dataset, the architecture of SSD and the experimental setup.

3.1 Dataset

The given dataset (see examples in Figure 1) consists of high-resolution images of real-world billboards with various background scenes. For each image the billboard is positioned in a different geographical location, contains differing advertisement content, the billboard is subject to various changes in orientations, varying positions from where the image was taken, has been captured over a wide range of times and thus is subject to daily and seasonal variations. Within each image there may be multiple background objects such as roads, pedestrians, vehicles, buildings, trees etc.

Figure 1. Selection of different images representing the different billboard classes and imaging conditions from the dataset: (a)-(c) Roadside Billboards; (d)-(f) Street Furniture Billboards

The billboards can be approximately grouped into two classes based on their overall size and location: Roadside Billboards (see Figure 1(a)-(c)) are approximately 325x250 cm and can be placed on the pavements , or major

intersections , mounted on one or two posts; Street Furniture billboards (see Figure 1(d)-(f)) are approximately 169 x 111 cm and often placed alongside road central reservations or pedestrian areas and mounted using a single post or a Base . The only obvious distinguishing feature between the billboard classes is that the Roadside Billboards are square shaped, and Street Furniture Billboards are rectangular shaped**.** All billboards have a frame and include the advertising company logo.

Billboards have been manually labelled using the LabelImage software to annotate a bounding box around the billboard or billboards within each image. In all cases the bounding box includes the entirety of the billboard frame as well as partial representation of the mounting posts when present. All bounding boxes retain a small proportion of background information. In total 1052 images were annotated, 532 images containing Roadside Billboards and 520 images containing Street Furniture billboards, and labelled with the appropriate class label. This provided an annotated PASCAL VOC format dataset.

3.2 SSD Convolutional Neural Network

The Single Shot Detector (SSD) [Liu, W. et al. 2016] is a feed-forward convolutional network that predicts bounding boxes and the classes directly from feature maps in one single pass, hence why it is known as the Single Shot Detector. The SSD detector is composed of 2 parts as illustrated in Figure 2: extraction of feature maps, and application of convolution filters to detect objects. In most cases the early network layers are a standard VGG-16 network [Simonyan, K., & Zisserman, A. 2014] which has been truncated prior to any classification layers used to extract the feature maps; the remaining network structure is composed of six additional convolutional feature layers that are appended to the end of the truncated VGG-16 network.

After passing through the VGG-16 layers we obtain a feature layer with a number of bounding boxes corresponding to object region locations. These bounding boxes may be different sizes and aspect ratios as a vertical rectangle is more fit for Street Furniture billboard, and a square is more fit for a Roadside billboard. For each bounding box the class score is computed along with offsets which correspond to the distance from the original bounding box shape.

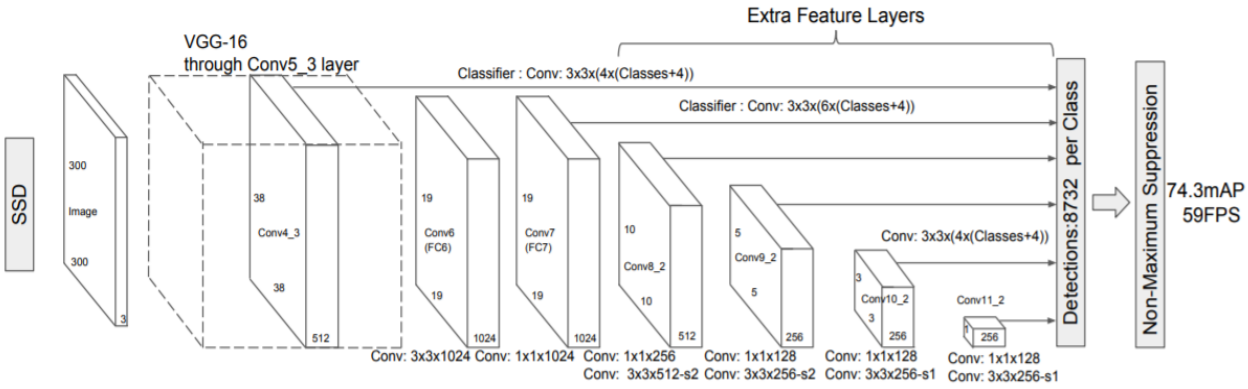


Figure 2. Architecture of SSD [Liu, W. et al. 2016]

To further improve detection, the feature maps go through the remaining network convolution feature layers where the location and class scores are computed using small convolution filters. SSD applies 3 × 3 convolution filters for each location to make predictions. The 3 × 3 convolution filters compute the results just like the regular CNN filters. SDD uses multiple layers to detect objects at different scales because as the spatial dimension is reduced the feature map resolution also reduces. In SSD, the lower resolution layers are used to detect objects at a larger scale and the higher resolution layers are used to detect objects at a smaller scale. Multi-scale feature maps have been shown to improve accuracy significantly [Liu, W. et al. 2016].

During SSD model training, the loss function is calculated using values obtained from the labelled, predicted and offset categories. The loss function is the sum of a classification loss and localization loss controlled by cross validation further comparing with matched bounding boxes as shown in equation 1:

$$L = \frac{1}{N}(Lc + \aleph Ll) \quad (1)$$

where, L is the loss, Lc is the classification loss, Ll is the localisation loss, N is the number of matched values of the bounding box, and $\aleph$ is the cross validation calculated balanced weight between two losses.

4. Experimental Setup and Results

We used the annotated image dataset described in Section 3.1 with the Tensorflow object detection API [Abadi, M. *et al.* 2016] and is used with pre-trained SSD mobileNet [Howard, A. G. *et al.* 2017]. The dataset consist of 1052 images is divided into two parts: 926 images (88.02%) of the images used as the training set and 126 images (11.98%) which are unseen during training and used to test the resulting network. In order to train the system, we tuned a number of hyperparameters, including batch-size, step-size and learning rate. Batch-size was varied from 24 to 1, step-size was varied from 25,000 to 75,000, and learning rate was varied between 0.001 and 0.004. Overall, the optimal parameters were found to be batch-size = 1, step-size = 75,000 and learning rate = 0.004.

Performance evaluation was conducted in parallel with the hyperparameter tuning to determine the optimal parameters. To do so, we use metrics such as precision, recall, average recall (AR), mean average precision (mAP), intersection over union (IoU) and loss each described as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}\ , Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Precision quantifies the number of positive class predictions that actually belong to the positive class
- Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.
- AR is the average of correctness of category prediction over IoU's.
- mAP calculates the average precision for each class based on the network's predicted bounding values and is combined with IoU, defined as the Area of the overlap divided by the area of the union of a predicted bounding box to adjust the accuracy if the match.
- Loss is the localisation loss calculated as the difference between the ground truth bounding box value and the predicted bounding box.

Step size	25,000	50,000	**75,000**
Loss	6.156	6.774	**6.022**
AR	62,71	53.88	**64.90**
mAP	54.03	49.65	**59.79**
mAP@.50IOU	80.01	71.19	**83.55**
mAP@.75IOU	63.21	65.37	**73.04**
(a) Learning rate = 0.004			

Step size	25,000	50,000	75,000
Loss	6.406	6.253	6.642
AR	59.67	61.02	59.47
mAP	45.25	54.67	50.09
mAP@.50IOU	78.88	80.55	75.41
mAP@.75IOU	48.37	67.71	62.51
(b) Learning Rate = 0.001			

Table 1: SSD output values with respect to hyperparameter changes with (a) learning rate of 0.004 and (b) learning rate of 0.001

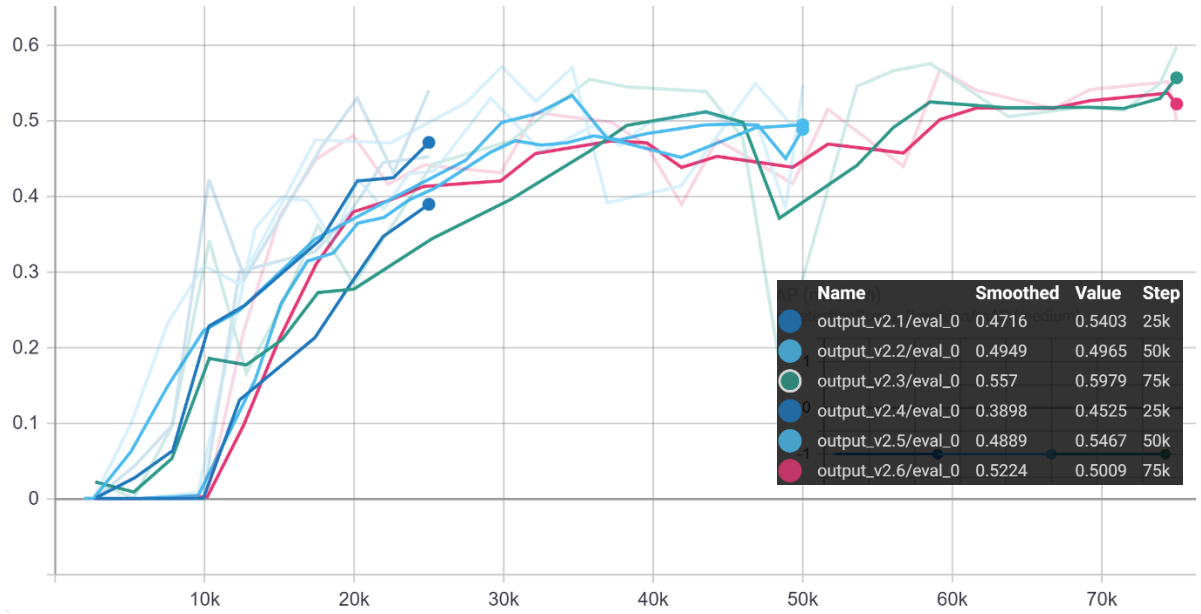


Figure 3: mAP values with respect to step size

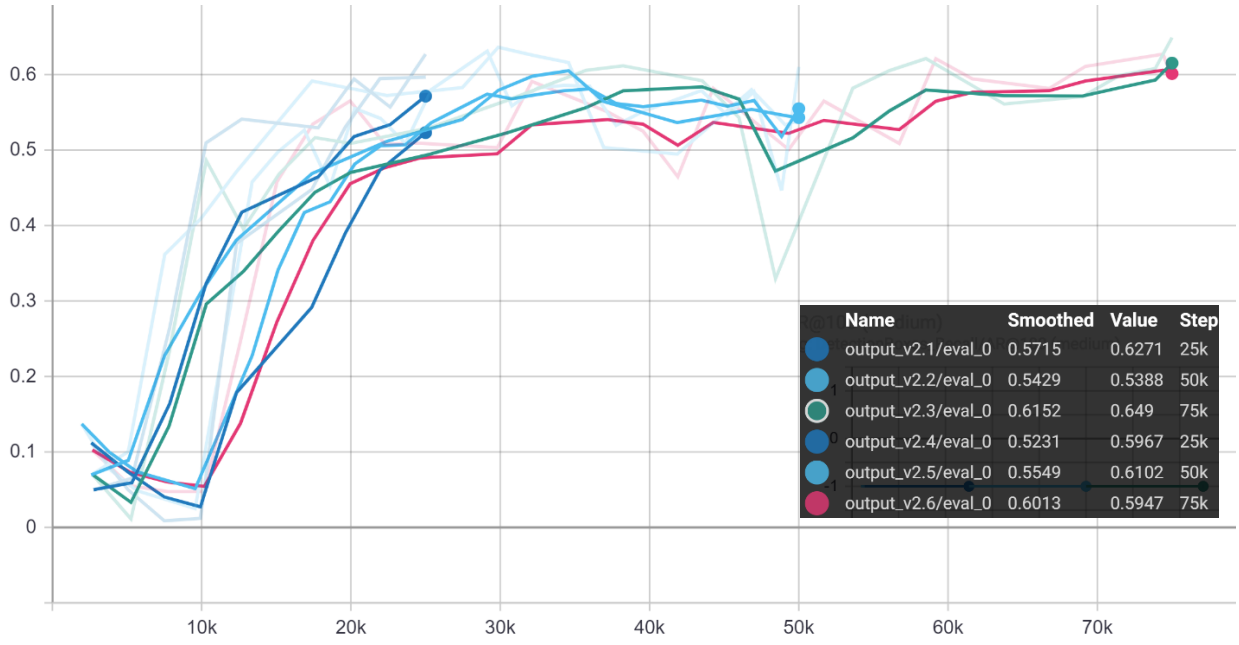


Figure 4: AR values with respect to step size

Observing the highest AR values in Figure 3 and Figure 4 alongside the mAP in Table 1 we can conclude that the highest accuracy at 59.79 mAP approx. 60% is obtained when the learning rate is set to 0.004 and step size is set to 75,000 resulting in the lowest loss value in Table 1(a). Once training is completed the test set is used to validate performance and visual examples of the network in detecting the two classes of billboards Roadside and Street Furniture are shown in the Figure 5.

Figure 5. Examples of the final network performance in detecting the two classes of billboards using unseen testing images: (a)-(c) Roadside Billboards; (d)-(f) Street Furniture Billboards

5. Conclusion

This paper presents an approach to detecting advertising billboards in outdoor environments. Using transfer learning with SSD the outdoor billboards were successfully detected with 60% training accuracy. However, in testing there are some cases when billboards were not detected as either of the two classes resulting in missed detection. We are currently exploring increasing the training dataset size and augmenting the dataset with additional variations of billboards to improve the detection rate. We are also exploring the use of different segmentation masks other than bounding boxes to determine if they can improve detection performance. Additionally, we will consider the use of other deep learning architectures such as RCNN, RFCN and YOLO.

Acknowledgements

We would like to express gratitude towards Digital Natives for providing the dataset and funding the PhD Scholarship at Ulster University.

References

[Abadi, M. *et al.* 2016] Abadi, M. et al. (2016) 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems'. http://arxiv.org/abs/1603.04467.

[Alom, M. Z. et al. 2018] Alom, M. Z. et al. (2018) 'The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches'. http://arxiv.org/abs/1803.01164.

[Bochkarev, K. and Smirnov, E. 2019] Bochkarev, K. and Smirnov, E. (2019) 'Detecting advertising on building façades with computer vision', *Procedia Computer Science*, 156, pp. 338–346. doi: 10.1016/j.procs.2019.08.210.

[Borisova, O. and Martynova, A. 2017] Borisova, O. and Martynova, A. (2017) 'Comparing the Effectiveness of Outdoor Advertising with Internet Advertising', Jamk, (September), p. 85.

[Cai, G., Chen, L. and Li, J. 2003] Cai, G., Chen, L. and Li, J. (2003) 'Billboard advertising detection in sport TV', Proceedings - 7th International Symposium on Signal Processing and Its Applications, ISSPA 2003, 1, pp. 537–540. doi: 10.1109/ISSPA.2003.1224759.

[Hossari, M. et al. 2018] Hossari, M. et al. (2018) 'ADNet: A deep network for detecting adverts', CEUR Workshop Proceedings, 2259, pp. 45–53.

[Howard, A. G. *et al.* 2017] Howard, A. G. *et al.* (2017) 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications'. http://arxiv.org/abs/1704.04861.

[Krishna, O. and Aizawa, K. 2018] Krishna, O. and Aizawa, K. (2018) 'Billboard, Saliency Detection in Street Videos for Adults and Elderly|", *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2326–2330.

[Li, Y. et al. 2016] Li, Y. et al. (2016) 'Nighttime lane markings recognition based on Canny detection and Hough transform', 2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016, pp. 411–415. doi: 10.1109/RCAR.2016.7784064.

[Liang, P. et al. 2018] Liang, P. et al. (2018) 'Planar object tracking in the wild: A benchmark', Proceedings - IEEE International Conference on Robotics and Automation, pp. 651–658. doi: 10.1109/ICRA.2018.8461037.

[Liu, W. et al. 2016] Liu, W. et al. (2016) 'SSD: Single shot multibox detector', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9905 LNCS, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[Morera, Á. et al. 2020] Morera, Á. et al. (2020) 'SSD vs. Yolo for detection of outdoor urban advertising panels under multiple variabilities', Sensors (Switzerland), 20(16), pp. 1–23. doi: 10.3390/s20164587.

[Morera, Á. et al. (2019)] Morera, Á. et al. (2019) 'Robust detection of outdoor urban advertising panels in static images', Communications in Computer and Information Science, 1047(June), pp. 246–256. doi: 10.1007/978-3-030-24299-2_21.

[Rahmat, R. F. et al. 2019] Rahmat, R. F. et al. (2019) 'Advertisement billboard detection and geotagging system with inductive transfer learning in deep convolutional neural network', Telkomnika (Telecommunication Computing Electronics and Control), 17(5), pp. 2659–2666. doi: 10.12928/TELKOMNIKA.v17i5.11276.

[Rahmat, R. F. et al. 2019] Rahmat, R. F. et al. (2019) 'Android-based automatic detection and measurement system of highway billboard for tax calculation in Indonesia', Indonesian Journal of Electrical Engineering and Computer Science, 14(2), pp. 877–886. doi: 10.11591/ijeecs.v14.i2.pp877-886.

[Shi, W., Bao, S. and Tan, D. 2019] Shi, W., Bao, S. and Tan, D. (2019) 'FFESSD: An accurate and efficient single-shot detector for target detection', Applied Sciences (Switzerland), 9(20). doi: 10.3390/app9204276.

[Simonyan, K., & Zisserman, A. 2014] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[Watve, A. K. and Sural, S. 2007] Watve, A. K. and Sural, S. (2007) 'Detection of on-field advertisement billboards from soccer telecasts', pp. 12–17. doi: 10.1049/cp:20060494.

[Wilson, R. T. and Casper, J. 2016] Wilson, R. T. and Casper, J. (2016) 'The role of location and visual saliency in capturing attention to outdoor advertising: How location attributes increase the likelihood for a driver to notice a billboard ad', Journal of Advertising Research, 56(3), pp. 259–273. doi: 10.2501/JAR-2016-020.

Context Aware Object Geotagging

Chao-Jung Liu[◇], Matej Ulicny[◇], Michael Manzke[◇] & Rozenn Dahyot[⋆]

ADAPT Research Centre, [◇]Trinity College Dublin, [⋆]Maynooth University, Ireland

Abstract

Localization of street objects from images has gained a lot of attention in the recent years. We propose an approach to improve asset geolocation from street view imagery by enhancing quality of the metadata associated with the images using Structure from Motion. The predicted object geolocation is further refined by imposing contextual geographic information extracted from OpenStreetMap. Our pipeline is validated experimentally against the state of the art approaches for geotagging traffic lights.

Keywords: Structure from Motion, street view imagery, OpenStreetMap

1 Introduction

Monitoring public assets is a labour-consuming task and for many decades, solutions collecting street view imagery have been routinely deployed in combination with computer vision-based approaches for object detection and recognition in images [8]. Nowadays, street view images are available in massive amounts (e.g.: Mapillary[1], Google Street View (GSV)[2]) and additional information about the scene can be further extracted by machine learning techniques. Krylov et al. [14, 15] have employed deep learning modules for segmenting objects of interest (e.g. poles) in images and estimating their distance from the camera, and a Markov Random Field (MRF) is then used as a decision module to provide a usable list of the GPS coordinates of the assets of interest, limiting duplicates by reconciling detection from multiple view images.

Figure 1: Example of a street view image with its metadata overlaid. The image is captured with Dioptra app[3].

The MRF conveniently merges information extracted from images and their metadata i.e. their associated camera location (GPS) and bearing information (cf. Fig. 1). Currently, the pipeline of Krylov et al. assumes that the metadata associated with the camera view pose is noiseless, however, it is not always the case (e.g. due to GPS receiver imprecision) and consequently, this noise affects the accuracy of the geo-location of the assets found. In this paper, we propose to improve that pipeline by (1) denoising the camera metadata using Structure from Motion (SfM) and (2) using contextual information extracted from Open Street Map (OSM)[4] to push the predictions to a more probable area where the objects should be situated based on road and building

[1] https://www.mapillary.com/

[2] https://developers.google.com/maps/documentation/streetview/overview

[3] https://play.google.com/store/apps/details?id=com.glidelinesystems.dioptra

[4] https://www.openstreetmap.org/

locations. Fig. 2 summarizes our contributions and our approach has been validated for traffic light geolocation (c.f. Sec.4).

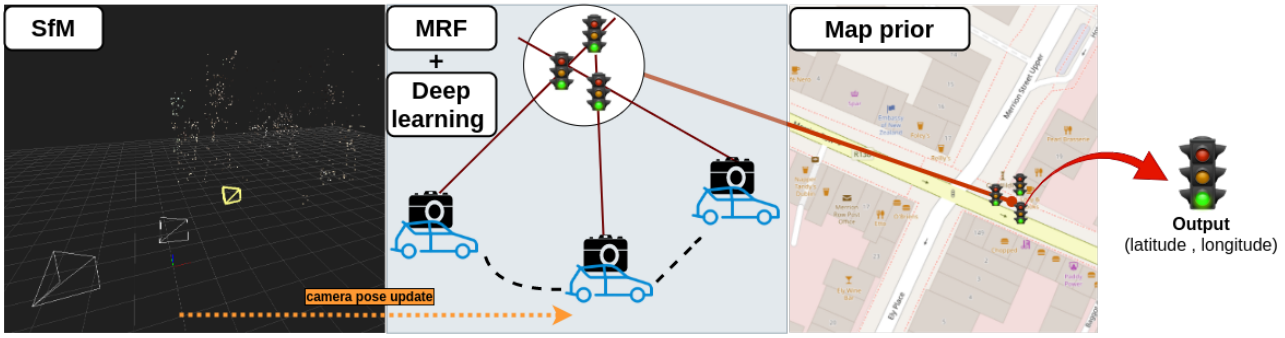


Figure 2: Pre-processing: SfM aims to de-noise camera medadata (i.e. poses) used as an input of the MRF. Post-processing: the Map prior module refines the result from MRF using contextual information from OSM.

2 State of the Art

2.1 Camera geolocalization

Various Simultaneous Localization and Mapping (SLAM) and SfM techniques have been proposed to infer 3D points and to estimate the motion from a set of images [13, 27, 7, 16, 10]. Bundle adjustment (BA) is integrating matched points within a sequence of images and finding a solution simultaneously optimal with respect to both camera parameters and 3D points. Agarwal et al. [2] is first to propose the bundle adjustment that is used in the structure from motion. The trajectory of camera pose estimation is based on relative measurements, error accumulation over time thus leads to drift. Lhuillier [17] proposed to use GPS geotag in the bundle adjustment optimization. A similar problem is the camera re-localization [29, 1]. A GPS tag and SfM technique are used to geo-localize a street view image by estimating its relative pose against images from a database. Bresson et al. [6] and Kendall et and al. [11] proposed to employ a CNN (Convolution Neural Network) features to estimate camera pose transformation.

2.2 Object geotagging

Qin et al. [24, 23] proposed to estimate the instance-level depth of objects in images as an alternative to pixel-wise depth estimation. They found out the latter (obtained by minimising the mean error for all pixels) sacrifices the accuracy of certain local areas in images. Bertoni et al. [5] employed a prior knowledge of the average height of humans to perform pedestrian localization. Qu et al. [25] proposed to detect and locate traffic signs from a monocular video stream. They relied on bundle adjustment with image GPS geo-tag to reconstruct a sparse point cloud as a 3D map, then align it with several landmarks from the 3D city model generated by Soheilian et al. [26].

Wegner et al. [28] proposed a probabilistic model to locate trees. They employed multiple modalities, including aerial view semantic segmentation, street view detection, map information as well as the tree distance prior. Information is fused into a conditional random field (CRF) to predict the positions of trees. However, identical features may be mismatched in case the recurring objects sit nearby. To solve this issue, Nassar et al. [20, 21] employed the soft geometry constraint on geo-location of camera pose to identify a same object that appears in two views. They concatenate camera pose information together with image features and decode them using a CNN. The same object in first view can be re-identified in the second view.

Nassar et al. [19] extend the method by constructing a graph from detected bounding boxes across the multi-views, feed the graph to a GNN [12] and let the GNN identify the same objects across different views. Hebbalaguppe. et al. [9] predicted bounding boxes around street objects, which was followed by the two-view

Figure 3: Input image representation consists of 8 overlapping rectilinear views split from a 360° field of view panorama. The image shown above is acquired from Mapillary API[1].

epipolar constraint to reconstruct 3D feature points from the two observed scenes. However, the 3D feature point does not necessarily fall inside the target bounding box. Krylov et al. [15] employed the camera pose from multiple views as a soft constraint and used semantic segmentation of images alongside a monocular depth estimator to extract the information (bearing and depth) about objects of interest, and feeds the obtained information into an MRF model that predicts their locations.

3 Methods

We present camera calibration using the SfM technique in Section 3.1, which provides a higher quality information to be used as an input to the MRF presented in Section 3.2. Section 3.3 proposes a post-processing method to refine the MRF predictions.

3.1 Structure from Motion: using optical observation to denoise on GPS data

The input represents a set of N panoramic street view images (360° field of view) captured with their metadata in an area of interest. Accurate camera geo-location is a key to accurately geo-locate objects in the scene. The GPS position in the metadata is inherently noisy, which lowers the accuracy for predicting object positions. To get a better estimate of the GPS coordinates associated with each camera position, we propose to tune each of the camera positions with a conventional 3D reconstruction pipeline [3], followed by bundle adjustment [2]. To ease image matching, we split the 360° panorama views into 8 overlapping rectilinear views: each view covers a 90-degree field of view and is overlapped by 45 degrees in the horizontal direction. Each view is then considered as an image captured by a pinhole camera, free of distortion (see Figure 3).

We aim to find all possible matching features extracted from our images and perform camera calibration to adjust the camera pose from image metadata.

We note the set of rectilinear views $\mathcal{V} = \{v_1^{(i)}, \cdots, v_8^{(i)}\}_{i=1,\cdots,N}$ where $v_1^{(i)}, \cdots, v_8^{(i)}$ corresponds to rectilinear views associated with panorama $i = 1, \cdots, N$ ($N = 112$ in our experiment). Suppose two views are matched by their detected features. The epipolar constraint with 5 point algorithm [3] is applied to find the essential matrix E which establishes the geometry relationship between two views. E can be further decomposed into translation and rotation matrix, noted as R and τ, respectively. They can be put together as a transformation

matrix $\Theta \in SE(3)$

$$\Theta = \begin{pmatrix} \mathrm{R} & \tau \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{4\times 4} \quad \text{with} \quad \mathrm{R} \in SO(3) \text{ and } \tau \in \mathbb{R}^3. \tag{1}$$

Each calibrated view in $\mathcal{V}$ is associated with $\Theta = (\mathrm{R}, \tau)$, these parameters can be estimated by minimizing the re-projection error from 3D feature space to 2D image plane within a bundle of images.

3.2 Object geolocation with MRF

The MRF model performs binary decisions on the nodes of a 2D graph, each node corresponding to an intersection between two rays. The rays correspond to rays (in 2D) with origins the camera GPS coordinates and with directions the bearings associated with the segmented object of interest (the pixel in the middle of the segmented object is chosen for the bearing information). The objects of interest are segmented using a deep learning pipeline that also estimates their distances (from the camera) [15]. Each camera view provides one or many rays shooting to the objects of interest. The MRF model is optimised to perform a binary decision for each node concerning its occupancy by the object of interest (i.e. 0 = no object, 1 = object present). For more information, please refer to [15]. Our contribution in this paper is in providing more accurate GPS coordinates for the camera positions (than originally available in the image metadata) thanks to SfM, hence improving the geo-location of the nodes on this MRF and ultimately improving the accuracy of GPS coordinates for the objects of interest.

3.3 Post-processing

Because of the inaccuracies of the rays that define the MRF nodes, the same object may be associated with multiple nodes (Fig. 4 left) located in the same vicinity on the MRF graph. To resolve this issue Krylov et al. [15] added a hierarchical clustering step after optimising the MRF to merge close positive sites together. The final position is the average of sites in the cluster. However, we have observed that some of the positive

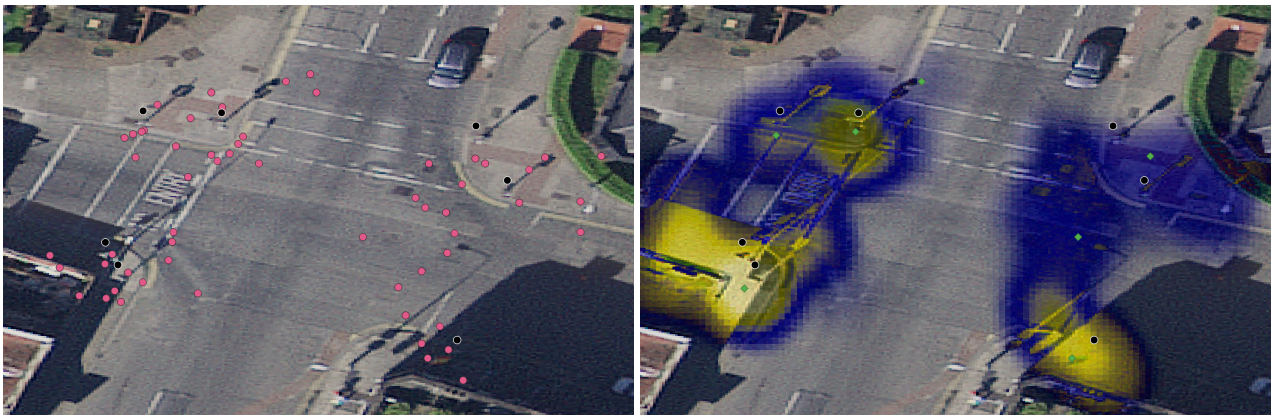


Figure 4: On the left figure, the red dots are positive intersections from MRF. The location of ground truth are shown in black dots. On the right figure, the green dots are the result after clustering process. Probability density function is applied to demonstrate the points' density of the intersections. The color code from blue to yellow means the number of intersections from small to many.

sites were situated at improbable areas, for example, in the middle of the road. Therefore, we propose here to use OSM data to act as a useful prior for an area. As our objects of interest (e.g. traffic lights) are static and are located on a side of the road, we apply the following rule: *the object of interest can not be located in the middle of the road, or around the edge of a building*. The OSM data has building and road classes represented by polygons and lines, respectively. A Normal kernel is fitted at each OSM node (cf. Fig. 5). Suppose a cluster C containing n positive sites $C = [c_1, c_2, ..., c_n]$, $W(x)$ is the function to query the weight that corresponds to the

particular site in C and depends on the OSM nodes N_x within the close proximity of the site x. The position P (equation 2) can be refined using a weighted average where certain sites are penalized with small weights.

$$W(x) = 1 - \min\left(1, \sum_{\mu,\sigma \in N_x} -\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)\right) \quad \text{and} \quad P = \frac{\sum_{i=1}^{n} W(c_i) \times c_i}{\sum_{i=1}^{n} W(c_i)} \tag{2}$$

The σ stands for the standard deviation in meters and the μ is the node centre obtained from the OSM data. The σ is set to 2 meters for roads and 1 meter for building edges. The resolution of the Gaussian grid is 25 centimetres.

3.4 Implementation

SURF descriptors [4] are used and in each view, 6,000 features are extracted. We employ FLANN (Fast Library for Approximate Nearest neighbours [18]) to match the SURF features between rectilinear views. RANSAC is used to remove outliers. We use the OpenSfM [5] to calibrate camera poses and Crese [2] as our solver to optimize the Θ. As the nodes from raw OSM data are not equally distributed, we interpolate the nodes every 5 meters in QGIS [6] to get a dense distribution of the map prior.

4 Experimental Results

To validate our approach, we have used 896 GSV images (112 panoramas split into 8 images each) collected in Dublin city centre. The object of interest corresponds to a traffic light. We fine-tuned the input camera poses using the SfM (cf. Fig. 5). Our method corrected the bearing and position information on average by 4.36 degrees and 0.71 meters respectively. Moreover, the use of the OSM prior results in an average refinement of the prediction by 0.17 meters.

#Actual	#Detected	TP	Precision↑	Recall↑	F-measure↑	Geo-localization error↓	Geo-localization error↓(with OSM)
no correction [15]							
76	94	58	0.61	0.76	0.68	2.71	2.64
correction on τ only							
76	89	57	0.64	0.75	0.69	2.79	2.74
correction on R and τ							
76	92	54	0.57	0.72	0.64	2.53	2.48

Table 1: We evaluate the impact of metadata correction by a comparison with results that do not use any pose correction. By correcting the full camera pose (R and τ), the geo-location accuracy reaches error of around 2.5 meters to a reference point. It outperforms the result with no correction by 18cm, and 16 cm after applying the OSM prior. We reach the highest F-measure if only the τ is corrected.

By using our SFM module we can check the impact of the following correction of the metadata: correction on τ only (i.e. GPS location of the camera), correction on R and τ (i.e. correction of both GPS location and bearing of the camera). To validate our approach, we use the original metadata as our baseline for comparisons. Table 1 shows the testing results in terms of geo-localization error and precision and recall detection metrics. We consider traffic lights to be recovered accurately (true positive) if they are located within 6 meters from the reference position, otherwise it is viewed as a false positive. The geo-localization error measures the average Haversine distance between the prediction and its reference target in meters. A small distance indicates accurate position prediction.

[5] https://github.com/mapillary/OpenSfM
[6] https://www.qgis.org/en/site/

We compare our results with related public asset geo-location approaches in Table 2. The proposed technique reaches the smallest positional error, however, the results are not directly comparable due to the different complexity of the scene and detected objects.

Comparison with other methods			
Method	Dataset	F-measure↑	Geo-localization error↓
Siamese CNN [21]	Pasadena [28]	0.51	3.13
Siamese CNN	Mapillary [22]	0.72	4.36
GNN-Geo [19]	Pasadena	0.64	2.75
GNN-Geo	Mapillary	0.87	4.21
Ours	**DTL** [15]	**0.64**	**2.48**

Table 2: In comparison with other approaches, our method achieves the smallest geo-localization error, although the other datasets might be more challenging for object detection.

5 Conclusion

We have shown that by denoising metadata associated with street view imagery using SfM, and by using context information such as road and building shapes extracted from OSM, assets of interest can be geolocalized with higher accuracy. Currently, our pipeline is geotagging one class of objects at a given time, and future work will investigate multiple static object class tagging with additional priors associated with their relative positioning in the scene, to improve further geolocation accuracy.

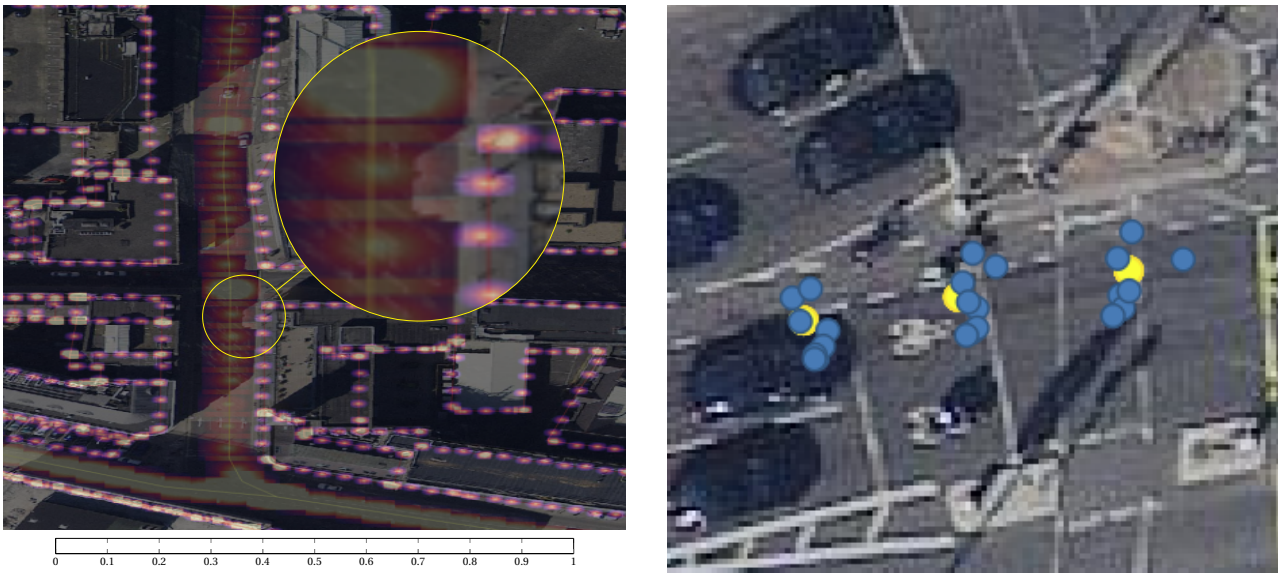


Figure 5: The figure (**left**) shows the prior map information overlaid on an aerial image. The normal kernel is applied to each node that is imported from OSM. The heatmap outlines improbable object locations that will have a smaller contribution towards the weighted sum. On the (**right**) yellow dots are the positions taken from image metadata and the blue dots represent their corrected versions with SfM.

Acknowledgments

This research was supported by the ADAPT Centre for Digital Content Technology funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs.

References

[1] Pratik Agarwal, Wolfram Burgard, and Luciano Spinello. Metric localization using google street view. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3111–3118. IEEE, 2015.

[2] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *European conference on computer vision*, pages 29–42. Springer, 2010.

[3] Alex M Andrew. Multiple view geometry in computer vision. *Kybernetes*, 2001.

[4] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[5] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6861–6871, 2019.

[6] Guillaume Bresson, Li Yu, Cyril Joly, and Fabien Moutarde. Urban localization with street views using a convolutional neural network for end-to-end camera pose regression. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1199–1204. IEEE, 2019.

[7] Mark Cummins. Highly scalable appearance-only slam-fab-map 2.0. *Proc. Robotics: Sciences and Systems (RSS), 2009*, 2009.

[8] Rozenn Dahyot. *Analyse d'images séquentielles de scènes routières par modèles d'apparence pour la gestion du réseau routier (Appearance based road scene video analysis for the management of the road network)*. PhD thesis, University of Strasbourg I, France, November 2001. (published in French).

[9] Ramya Hebbalaguppe, Gaurav Garg, Ehtesham Hassan, Hiranmay Ghosh, and Ankit Verma. Telecom inventory management via object recognition and localisation on google street view images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 725–733. IEEE, 2017.

[10] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3287–3295, 2015.

[11] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] Bryan Klingner, David Martin, and James Roseborough. Street view motion-from-structure-from-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 953–960, 2013.

[14] Vladimir A. Krylov and Rozenn Dahyot. Object geolocation using mrf based multi-sensor fusion. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2745–2749, 2018.

[15] Vladimir A. Krylov, Eamonn Kenny, and Rozenn Dahyot. Automatic discovery and geotagging of objects from street view imagery. *Remote Sensing*, 10(5), 2018.

[16] Taehee Lee. Robust 3d street-view reconstruction using sky motion estimation. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1840–1847. IEEE, 2009.

[17] Maxime Lhuillier. Fusion of gps and structure-from-motion using constrained bundle adjustments. In *CVPR 2011*, pages 3025–3032. IEEE, 2011.

[18] Marius Muja and David G Lowe. Fast matching of binary features. In *2012 Ninth conference on computer and robot vision*, pages 404–410. IEEE, 2012.

[19] Ahmed Samy Nassar, Stefano D'Aronco, Sébastien Lefèvre, and Jan D Wegner. Geograph: Graph-based multi-view object detection with geometric cues end-to-end. In *European Conference on Computer Vision*, pages 488–504. Springer, 2020.

[20] Ahmed Samy Nassar, Nico Lang, Sébastien Lefèvre, and Jan D Wegner. Learning geometric soft constraints for multi-view instance matching across street-level panoramas. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE, 2019.

[21] Ahmed Samy Nassar, Sébastien Lefèvre, and Jan Dirk Wegner. Multi-view instance matching with learned geometric soft-constraints. *ISPRS International Journal of Geo-Information*, 9(11):687, 2020.

[22] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[23] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019.

[24] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Xiaozhi Qu, Bahman Soheilian, and Nicolas Paparoditis. Vehicle localization using mono-camera and geo-referenced traffic signs. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 605–610. IEEE, 2015.

[26] Bahman Soheilian, Olivier Tournaire, Nicolas Paparoditis, Bruno Vallet, and Jean-Pierre Papelard. Generation of an integrated 3d city model with visual landmarks for autonomous navigation in dense urban areas. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 304–309. IEEE, 2013.

[27] Akihiko Torii, Michal Havlena, and Tomáš Pajdla. From google street view to 3d city models. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 2188–2195. IEEE, 2009.

[28] Jan D Wegner, Steven Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023, 2016.

[29] Li Yu, Cyril Joly, Guillaume Bresson, and Fabien Moutarde. Monocular urban localization using street view. In *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 1–6. IEEE, 2016.

CLADA: Contrastive Learning for Adversarial Domain Adaptation

Richard Greene and Kevin McGuinness

School of Electronic Engineering, Dublin City University

Abstract

This paper focuses on the challenging problem of unsupervised domain adaptation of synthetic data for the semantic segmentation task of autonomous driving scenes. It is motivated by the generative adversarial methods that apply image-to-image translation by learning a mapping between the source and target domains. Fully supervised training of deep models for semantic segmentation do not generalize well to unseen target data. By applying domain adaptation, a model can be fit that generalizes to the target domain. Previous work has shown that combining generative adversarial networks with cycle consistency is effective for mapping images between domains, which can then be used to train a domain invariant semantic segmentation model. However, this requires additional networks to implement the cycle-consistency constraint. This paper proposes replacing this with a more efficient contrastive objective for the semantic segmentation task. By reducing the training time and computational resources, more complex end-to-end domain adaptation architectures may be used.

Keywords: Deep Learning, Generative Adversarial Network, Domain Adaptation, Contrastive Learning

1 Introduction

Deep convolutional neural networks have produced impressive results in many computer vision tasks, such as image classification, segmentation, object detection, and image generation. Semantic segmentation, in particular, has been substantially improved in recent years and has several important applications, including autonomous driving systems. This research focuses on the semantic segmentation of dashcam images captured by a vehicle to assign a semantic label to each pixel, e.g. road, vehicle, building, pedestrian, etc.

Supervised learning is the most common approach to fitting a semantic segmentation model. Using a large labelled dataset, a model can be trained to classify each pixel of the input based on the labels provided. Generating the large datasets required for autonomous driving perception tasks is, however, time consuming and expensive, due to the time and cost associated with manually annotating these datasets with dense pixel-level labels. This supervised approach also assumes both training data and unseen test data are drawn from the same distribution. If this assumption is violated, the model trained on the source data will fail to generalize to unseen test data due to the differences between the two distributions. This is commonly referred to as domain shift.

Domain adaptation is a type of transfer learning that attempts to reduce this domain shift, with the aim of transferring knowledge learnt from labelled data in a sourced domain to another target domain, where labelled data may be unavailable. By leveraging 3D graphics engine technology, commonly used for game development, large amounts of synthetic training images and corresponding labels can be generated in a fraction of the time and cost compared to collecting and hand labelling real world dashcam imagery. A semantic segmentation model can be fit using this synthetic dataset, and domain adaptation techniques applied to reduce the domain shift. Domain adaptation attempts to ensure the models' performance does not drop in the target domain when trained only with the synthetic source domain data. Successful domain adaptation eliminates the need for labelled data in the target domain allowing models to be trained using more cost effective and larger scale synthetic datasets.

This research focuses on the adversarial image-to-image translation approach to unsupervised domain adaptation of synthetic dashcam images, where ground truth labels are only available in the source domain. The goal is to learn a task model that performs well in an unseen target domain. A new contrastive learning based objective

function is proposed as a more efficient alternative to the cycle consistency loss commonly used. We refer to our full network architecture and approach as CLADA, which includes both pixel-, and feature-level domain adaptation to train a target semantic segmentation task model (See Figure 1). Our full approach requires less computational resources, leading to reduced training time.

2 Related Work

Several approaches have been taken to solve the domain adaptation challenge and deep learning methods have shown great progress by discovering domain invariant feature representations or by mapping images between the source and target domains [Shrivastava et al., 2017, Bousmalis et al., 2017, Li et al., 2019, Hoffman et al., 2018]. Earlier domain adaptation approaches focused on alignment within the feature space using some distance metric between the first- or second-order statistics of the source and target domains. By aligning the feature space representations of both domains, such that the feature embeddings follow the same distribution, a domain invariant model that generalizes better to the target domain can be learned [Sun and Saenko, 2016, Tzeng et al., 2014, Long et al., 2015]. Domain adversarial objectives have also been applied to feature space alignment, where a domain classifier is trained to distinguish between the source and target representations [Ganin et al., 2016, Tzeng et al., 2017, Tzeng et al., 2015, Ganin and Lempitsky, 2015].

More recently, further improvements have been made by approaching domain adaptation as a pixel-level image-to-image translation problem, leveraging the progress made by generative adversarial networks (GAN) [Goodfellow et al., 2014] in the image synthesis and style transfer domains. Earlier GAN based approaches to image-to-image translation required paired image samples [Isola et al., 2017], which would not be practical for the autonomous driving perception task.

Shrivastava et al. [Shrivastava et al., 2017] proposed SimGAN to translate unpaired images from synthetic source images to a target domain with the introduction of an additional self-regularizing function. This approach is successful in domains where there is a limited domain shift in pixel space. The addition of this L1 reconstruction loss for the generator during training preserves the annotations of the source data by penalizing large changes to the global structure during translation. Preserving this structural content is essential in pixel-level domain mapping, otherwise the source annotations would not accurately represent the new translated data used for supervised learning of the semantic segmentation model.

Zhu et al. [Zhu et al., 2017] introduced CycleGAN, which proposed a learned mapping applied to an input x in both directions should be cycle consistent. That is, mapping a sample x from the source domain X to a target domain Y using a learned mapping function $G_{s\rightarrow t}$, and then mapping back to the source domain using a learned mapping function $G_{t\rightarrow s}$, the result should be consistent with the original input x. CycleGAN uses the L1 distance to measure the reconstruction error, which they call cycle consistency loss. The forward and backward consistency constraint is used to train the generator model along with the original discriminator GAN loss. Zhu et al. [Zhu et al., 2017] reported better results for several image translation experiments when compared to SimGAN. CycleGAN, however, requires additional generator and discriminator models to implement cycle consistency, which results in higher computational requirements and time during training.

Hoffman et al. [Hoffman et al., 2018] proposed CyCADA, a combined approach to domain adaptation for the semantic segmentation task by applying both feature space domain invariant feature learning and pixel space domain mapping. CyCADA separates domain adaptation into two sequential steps. First, performing image translation from the source to target domain with a CycleGAN, and then further decreasing the domain gap by adding a domain adversary to the features of the semantic segmentation model. The advantage of pixel space adaptation is that it is more human interpretable, which allows visualizing the progress of the model as it is trained. This approach allows for interpretability at the pixel level, while also regularizing the feature level.

Li et al. [Li et al., 2019] introduced a bidirectional learning framework that uses both a CycleGAN-based image translation network and a segmentation adaptation network, similar to CyCADA. However, an end-to-end bi-directional training process was used, requiring more resources to train the closed-loop end-to-end architecture. Park et al. [Park et al., 2020] recently proposed a new image translation approach, introducing an alternative to the cycle consistency loss that does not require the additional generator and discriminator models for the two-way

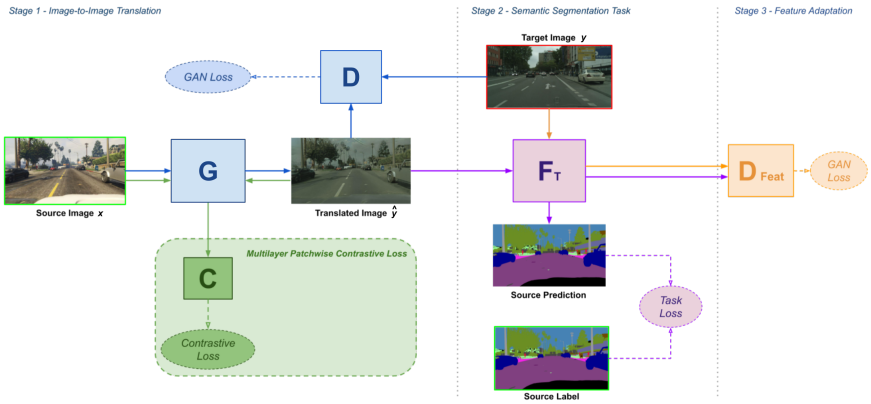


Figure 1: **Proposed CLADA architecture overview:** pixel-level GAN losses are in (blue), the PatchNCE loss in (green), the semantic segmentation task loss in (purple), and the additional feature-level GAN loss in (orange).

cycle consistency calculation. A multilayer patchwise contrastive loss (PatchNCE) is used to learn a one-way unpaired image translation that maintains the content of the input image while allowing the appearance to be adapted. The authors suggest that this alternative method is faster and requires less computational resources than a CycleGAN, which relies on additional auxiliary networks. Park et al. [Park et al., 2020] claim that their full method, including an additional identity loss, is 40% faster and 31% more memory efficient than CycleGAN at training time. The results shown in [Park et al., 2020] strongly suggest that PatchNCE could provide a more efficient alternative to cycle consistency in domain adaptation.

Motivated by [Park et al., 2020] and [Hoffman et al., 2018] this research evaluates the impact of replacing cycle consistency with a PatchNCE loss for unsupervised domain adaptation of synthetic autonomous driving dashcam images, which are then used to perform supervised learning for semantic segmentation. It shows that competitive results can be achieved for the semantic segmentation task using a simplified model architecture and less resources, producing faster training times. To our knowledge, no study has evaluated this approach on unsupervised domain adaptation for semantic segmentation.

3 Approach

Provided with source data X and ground truth labels, and target data Y, with no labels, the aim is to learn a task model F_t that when trained on the source data can correctly predict the semantic labels for the target data. Figure 1 shows the proposed architecture.

Similar to the staged-based approach taken by [Hoffman et al., 2018], we begin by fitting an image translation model G (identical to $G_{s\to t}$ in [Hoffman et al., 2018]) that will apply pixel level domain adaptation to reduce the domain gap between the source and target data. This model G is trained using a generative adversarial approach where it learns to map source images to the target domain, thus fooling an adversarial discriminator D based on the GAN loss:

$$\mathscr{L}_{\text{GAN}}(G,D,X,Y) = \mathbb{E}_{y\sim Y} \log D(y) + \mathbb{E}_{x\sim X} \log(1 - D(G(x))). \quad (1)$$

To preserve the content structure of the source samples x_s, many previous approaches have used the cycle consistency loss. Here we propose to replace this with the contrastive learning based PatchNCE loss proposed by [Park et al., 2020]. This is based on a type of contrastive loss function, the InfoNCE loss [Oord et al., 2018], which aims to learn an encoder that associates corresponding patches with each other. The aim is to match corresponding patches between the input and output images. For example, a patch in the input image showing a traffic light should be associated with the corresponding patch showing a traffic light in the translated image. [Park et al., 2020] propose selecting multiple positive and negative pairs of patches from several layers within the feature stack of

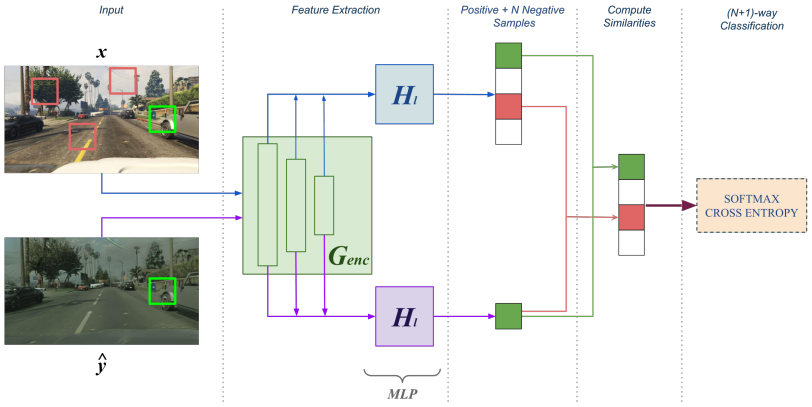


Figure 2: Multilayer patchwise contrastive loss.

the encoder G_{enc} based on the normalized temperature scaled cross-entropy (NT-Xent) loss:

$$\ell(v, v^+, v^-) = -\log\left[\frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(v \cdot v_n^-/\tau)}\right], \tag{2}$$

where v, v^+ and v^- are patches taken from layers and spatial locations within the feature stack of the image translation generator and τ is the temperature. By feeding the feature maps into a small multi-layer perceptron H_l and selecting 1 positive and N negative samples from a number of spatial locations, an $(N+1)$ way classification problem is setup. The contrastive loss (PatchNCE),

$$\mathscr{L}_{\text{PatchNCE}}(G_{\text{enc}}, H, X) = \mathbb{E}_{x \sim X}\left[\sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell(\hat{z}_l^s, z_l^s, z_l^{S/s})\right], \tag{3}$$

can then be calculated and fed back to the generator during the training cycle. L represents the layers within the generator G_{enc} passed to H_l, S_l represents the spatial locations within each layer, and $\hat{z}_l^s$, z_l^s and $z_l^{S/s}$ represent the query, positive and negative patches respectively. This bypasses the need for a predefined similarity function. Figure 2 shows an overview of this multilayer patchwise contrastive loss architecture.

Park et al. [Park et al., 2020] introduces two variants of their architecture, which they refer to as CUT and FastCUT. Their CUT model includes an additional identity loss to impose a content structure constraint and selects $N = 256$ patches from $L = 5$ layers of the encoder, where FastCUT omits the identity loss and selects only $N = 16$ patches from within each layer. These proposed L and N values were also chosen for our CUT and FastCUT model variants in our experiments. FastCUT also applies a weight, λ, to the PatchNCE loss to constrain the content structure in the absence of the identity loss. The resulting loss function is

$$\mathscr{L}_{\text{GAN}}(G, D, X, Y) + \lambda_X \mathscr{L}_{\text{PatchNCE}}(G_{\text{enc}}, H, X) + \lambda_Y \mathscr{L}_{\text{PatchNCE}}(G_{\text{enc}}, H, Y). \tag{4}$$

For the CUT model, λ_X and λ_Y are set to 1.0 to jointly train with the identity loss. [Park et al., 2020] proposes using $\lambda_Y = 0.0$ for FastCUT to omit the identity loss and $\lambda_X = 10.0$ to compensate for its absence. We found $\lambda_X = 10.0$ too high in our experiments: qualitative results showed the model failed to translate the Cityscapes style, resulting in images more similar to the source GTA5 data. Reducing λ_X to 5.0 improved the image translation results. We refer to this model variant as FastCUTL5 in the remainder of the paper.

Once an image-to-image translation model for producing translated images that are similar to images in the target domain has been fit, the learned model G is used to generate a new translated dataset. This new translated data, along with the corresponding source labels, is used as training data for the next stage, where a fully supervised learning approach is taken to train a target semantic segmentation model F_t.

Lastly, additional domain adaptation is applied within the feature embedding space of the task model F_t using a domain adversarial approach. By introducing a feature level GAN loss, we fit a discriminator D_{feat} that can

(a) Source (b) CycleGAN (c) CUT (d) FastCUT L5

Figure 3: **Image translation:** GTA5 source input and translation models output examples. We observed all models successfully transferred the Cityscapes saturation levels and textures, such as the smoother road surface.

distinguish between the feature embeddings of inputs from the translated source and target datasets, when passed through the task model F_t, and feed that back to the task model during a round of fine tuning.

4 Experiments

Model	FID	Time
Source	68.6	–
CycleGAN	28.6	426ms
CUT	29.4	321ms
FastCUT L5	33.2	170ms

Table 1: Fréchet Inception Distance (FID) and training time for a single iteration.

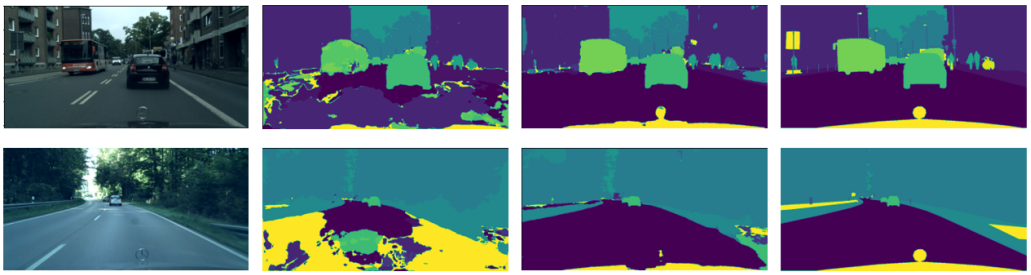
We evaluate our approach on the challenging unsupervised adaptation of the GTA5 [Richter et al., 2016] to Cityscapes [Cordts et al., 2016] datasets for the semantic segmentation task. Given that the ground truth labels are only available for the source GTA5 dataset, a task model is fit and its performance is evaluated in the target Cityscapes domain, where labelled data is available in a validation set. GTA5 is made up of 24,966 synthetic images extracted from the GTA5 computer game, with corresponding semantic labels at 1914×1052 resolution. Cityscapes provides real world dashcam images captured in Germany, and is split into train, validation, train extra and test sets at a resolution of 2048×1024. The train split has 2,975 images with dense ground truth labels, the validation split has 500 images with dense ground truth labels, and the train extra has 19,998 images without labels. Given that no labels are available for the test split, we use the validation split to evaluate our models performance.

Similar to CyCADA, a staged based approach is taken in which the image translation models are trained first. This allows us to interpret the progress of the pixel-level domain adaptation stage, and qualitatively evaluate the impact of replacing cycle consistency loss (CCL) with PatchNCE before proceeding to subsequent stages that include training the task model F_t and further adaptation in feature space.

We trained a CycleGAN using the network architecture and training procedure from CyCADA. The images were resized to a width of 1024, maintaining the aspect ratio, from which random 400×400 crops were taken as input. The model was trained with a batch size of 1 for 400k iterations with a learning rate of 2×10^{-4}. After 200k iterations, the learning rate was linearly decayed to 0. The same procedure was used to train additional CUT and FastCUT models, where CCL was replaced with PatchNCE loss. As discussed $\lambda_X = 10$ for the PatchNCE loss was found to be too high for the GTA5-Cityscapes task; reducing it to $\lambda_X = 5$ achieved better results.

Following initial domain adaptation at the pixel-level, all models produced good quality images when translated to the Cityscapes appearance (Fig 3 illustrates an example). In particular, we noted that our CUT based models learned to adapt similar characteristics of the target Cityscapes domain to those adapted by CycleGAN, such as the image saturation levels, contrast and texture. All models, for example, learned that the road surface is much smoother in the target Cityscapes domain. We also noted that both the CycleGAN and full CUT model attempt to transfer the hood ornament, but FastCUT was not as prone to this. This is likely due to the lack of the additional identity loss, which may cause such features to be transferred.

Once all image translation models were trained and producing good qualitative image translation results, the



(a) Test Image (b) Source Only (c) CLADA (d) Ground Truth

Figure 4: **Semantic segmentation:** a test image (a) along with the corresponding source only model (b) predictions; our CLADA model (c) predictions and the ground truth masks (d).

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	**mIoU**	**fwIoU**	**Pixel acc.**
Source Only	31.3	14.0	54.2	10.9	9.8	21.8	21.4	4.9	76.5	19.5	66.2	41.9	2.2	53.5	13.8	5.6	0.0	2.8	0.0	23.7	44.7	56.4
CycleGAN	78.9	30.4	75.6	20.5	1.3	**29.5**	24.3	0.0	80.3	32.2	69.2	48.0	0.0	78.9	24.6	31.0	0.0	0.0	0.0	32.9	71.0	79.2
CUT	77.4	27.2	75.1	18.4	17.9	27.9	0.0	0.0	**80.4**	29.8	72.9	47.6	0.0	79.6	28.3	28.6	0.0	0.0	0.0	32.2	70.4	78.4
FastCUT L5	78.6	**32.1**	**76.8**	24.1	**19.6**	24.9	11.4	**13.1**	79.4	31.2	**73.2**	47.0	5.7	80.0	22.6	27.3	0.0	0.3	0.0	33.4	71.4	79.0
CLADA (FastCUT L5 + FeatAda)	**79.9**	31.0	**76.8**	**24.5**	18.7	28.0	**24.4**	12.7	79.6	**31.6**	72.2	**51.0**	**11.7**	**81.5**	**29.9**	**33.9**	**3.4**	**7.6**	0.0	**36.8**	**72.2**	**80.3**
Oracle	92.1	68.0	84.6	41.2	41.9	44.2	32.7	51.4	87.9	48.2	87.5	67.6	41.3	89.7	50.8	59.3	42.5	1.2	61.8	57.6	84.8	89.5

Table 2: GTA5-Cityscapes semantic segmentation task model evaluation results showing IoU for individual classes and mean IoU, frequency weighted IoU and pixel accuracy.

full GTA5 dataset was used to generate new translated datasets using each trained image translation model. A qualitative and quantitative evaluation was then performed. Each of these new translated datasets were compared to the Cityscapes data and the Fréchet Inception Distance (FID) [Heusel et al., 2017] calculated. Table 1 shows the results, and illustrates that, based on the statistical comparison between the translated datasets and the target Cityscapes data, the adapted images are more similar to Cityscapes than the original GTA5 images. Training time was measured and shown to be reduced when using the contrastive approach, in particular for the FastCUT variant that omits the identity loss. We found FastCUT is 47% faster than CUT and 60% faster than CycleGAN during training, where CycleGAN takes 426ms per iteration and FastCUT only takes 170ms (see Table 1).

The new translated datasets are then used in the next stage where we trained our task semantic segmentation models. The goal is to train a task model F_t that performs well when evaluated on the Cityscapes validation split. Each task model is evaluated using three metrics: mean intersection-over-union (mIoU), frequency weighted intersection-over-union (fwIoU) and pixel accuracy (pixel acc.):

$$\text{mIoU} = \frac{1}{N} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \qquad \text{fwIoU} = \frac{1}{\sum_k t_k} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \qquad \text{pixel acc.} = \frac{\sum_i n_{ii}}{\sum_i t_i}, \tag{5}$$

where N is the total number of classes, n_{ij} is the number of pixels of class i predicted as class j and $t_i = \sum_j n_{ij}$ is the total number of pixels of class i.

For the semantic segmentation task, we use an EfficientNetB3 [Tan and Le, 2019] model, pretrained on ImageNet [Deng et al., 2009], as an encoder within a U-Net [Ronneberger et al., 2015] architecture. Each task model was trained with a batch size of 8 for 120k iterations with an initial learning of 2×10^{-4}, which was stepped down to 10^{-5} for the final 40k iterations. The same training procedure was used for all model variants, with the encoder frozen for the first 40k iterations. The *Source Only* and *Oracle* models were used to set lower and upper bounds on the achievable accuracies. The *Source Only* model was trained using the original GTA5 data and labels and the *Oracle* model was trained using the Cityscapes data and dense ground truth labels provided in the train split. We then trained our semantic segmentation task models using each of the translated datasets created using our CycleGAN, CUT, and FastCUTL5 image translation models. Finally, each of these trained task models was evaluated on the Cityscapes validation split using our task evaluation metrics (see Table 2).

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	**mIoU**	**fwIoU**	**Pixel acc.**
Source Only	60.8	54.0	30.5	30.3	32.1	22.4	11.3	46.5	11.4	28.7	21.3	25.7	39.1	36.2	37.0	53.7	42.5	-1.6	61.8	33.9	40.2	33.1
Ours (CLADA)	12.2	37.0	7.8	16.7	23.2	16.2	8.3	38.7	8.3	16.6	15.3	16.6	29.6	8.2	20.9	25.4	39.1	-6.4	61.8	20.8	12.6	9.2
% Performance Gain	79.9	31.5	74.4	44.8	27.8	27.8	26.7	16.7	27.4	42.2	28.1	35.4	24.3	77.4	43.6	52.7	8.0	-	0.0	38.7	68.6	72.2

Table 3: Performance gap for the *source only* model and our full CLADA model, when compared to the oracle performance. The % performance gain is also shown for our full CLADA model versus the source only model. *Note: for the motorcycle class, our model outperforms the oracle*

The final end task results achieved using PathNCE during the image translation stage are comparable to those achieved with a CCL based CycleGAN. All models perform well on common classes. The performance of all models is poor for the *train* and *bicycle* classes due to these classes being under represented in the dataset. Semantic mask predictions (Fig 4) show qualitative results that correlate with the quantitative results. At this stage, FastCUTL5 has slightly better results, and has closed the performance gap for mIoU by approx 29%, which is competitive to CycleGAN and suggests that the PatchNCE contrastive objective is a viable replacement for CCL if faster training times is required, which may also lead to reduced training costs.

To further close the performance gap with the upper bound, as per [Hoffman et al., 2018], we performed further domain adaptation in feature space using a domain adversarial approach where we fine-tuned the FastCUTL5 model using a domain discriminator to classify the feature embeddings of the intermediate layer of the task model when source and target data are used as input. The final model, which we call CLADA, was evaluated using the same procedure and metrics and the results show the mIoU gap is closed by a further 10% (see Table 2). Overall, CLADA recovers approx, 39% mIoU lost to domain shift for the target segmentation task. In some cases, for well represented classes such as road, building, and car, it recovered >70% of the IoU performance lost (Table 3).

5 Conclusion

Our experiments show that by using a contrastive learning based objective function, PatchNCE, similar results to cycle consistency loss can be achieved for the challenging GTA5-Cityscapes semantic segmentation task, with faster training times and using a simplified model architecture. For the full approach, CLADA, the image translation stage is 60% faster during training and recovers approximately 39% of the mIoU performance lost to domain shift for the target semantic segmentation. By reducing the resources, costs, and time required to train generative adversarial domain adaptation models, our findings support further research into more complex end-to-end image translation approaches to domain adaptation of synthetic data for semantic segmentation.

References

[Bousmalis et al., 2017] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[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 *Computer Vision and Pattern Recognition (CVPR)*.

[Ganin and Lempitsky, 2015] Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, volume 37.

[Ganin et al., 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27.

[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. In *International Conference on Neural Information Processing Systems (NIPS)*.

[Hoffman et al., 2018] Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. (2018). CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *International Conference on Machine Learning (ICML)*.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *Computer Vision and Pattern Recognition (CVPR)*.

[Li et al., 2019] Li, Y., Yuan, L., and Vasconcelos, N. (2019). Bidirectional learning for domain adaptation of semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

[Long et al., 2015] Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*.

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[Park et al., 2020] Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive Learning for Conditional Image Synthesis. In *European Conference on Computer Vision (ECCV)*.

[Richter et al., 2016] Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, pages 102–118.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *LNCS*, volume 9351, pages 234–241.

[Shrivastava et al., 2017] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *Computer Vision and Pattern Recognition (CVPR)*.

[Sun and Saenko, 2016] Sun, B. and Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV)*.

[Tan and Le, 2019] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114.

[Tzeng et al., 2015] Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*.

[Tzeng et al., 2017] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176.

[Tzeng et al., 2014] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

[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 *International Conference on Computer Vision (ICCV)*.

Algorithm architecture comparison for mammogram anomaly classification

Jonathan Armstrong, Paul Miller and Jesus Martinez del Rincon

Queen's University Belfast

Abstract

Early detection of cancer is crucial to patient recovery, thus screening programs have been developed to spot signs of cancer before it becomes detrimental to health and life expectancy. These programs cause additional burden on already overworked clinicians so computational solutions were developed to reduce the load. Deep learning has paved the way for high accuracy techniques that aid the clinical diagnosis though most require large amounts of highly curated data which sometimes can be insurmountable to new stakeholders entering the field. We seek to investigate image input hyperparameters while comparing state-of-the-art algorithms for mammogram detection to tailor a patch classifier to show that with an insight in the data high accuracy can be achieved with a small dataset.

Keywords: Imaging, Image Processing, Machine Vision, Breast Cancer, Mammograms

1 Introduction

The World Health Organisation (WHO) states that cancer is the leading cause of death in the in the world accounting for an estimated 9.6 million deaths in 2018 ["WHO cancer key facts" n.d.]. Breast and Colorectal, being the second and third most common respectively after lung ["WHO cancer key facts" n.d.]. WHO also estimates that 30-50% of cancer deaths can be avoided by following current prevention strategies. Early detection is key to avoiding the more serious cases ["WHO cancer key facts" n.d.].

With the goal of early detection, the National Health Service (NHS) screening programs have been introduced to proactively diagnose cancer and thus avoid some of the serious cases. In the UK two of the largest cancer centric programs are the Breast Cancer Screening Program [Cancer 2006] and the Bowel Cancer Screening Program [Logan *et al.* 2012] involving the areas of mammography and digital pathology. The availability of images in these fields opens the opportunity of easing the burden on the medical professionals by turning to computational analysis methods. Specifically, we will be focusing on mammograms in this study.

The aim of this paper is to investigate the image input hyperparameters, define the best state-of-the-art algorithm for mammogram detection and, with an understand of the outputs, explain and tailor a patch classifier to show that with an insight in the data high accuracy can be achieved with a small dataset. This will be done through an extensive study of the misclassified patches and a comparison of algorithm architectures with augmentation. For the mammograms analysis we used the open dataset MIAS [Lee *et al.* 2017] containing 322 annotated images.

2 State of the Art

There are many techniques available to physicians to examine the breast, which cover a plethora of different technologies such as Conventional radiography, Mammography, Computed Tomography, Ultrasound, Magnetic Resonance Imaging [WHO 2020] among others. Despite this, mammograms are the most used method as they are highly effective at detecting breast cancer before it becomes clinically palpable.

Historic studies have shown that mammogram reading and diagnosis is susceptible to false positive and false negative reading [R E Bird, T W Wallace 1992][Kerlikowske *et al.* 2000]. To mitigate these issues the possibility of double reading was theorised [E L Thurfjell, K A Lernevall 1994], but this subsequently creates a greater load

on the consultant as now the images have to be seen twice. Mammograms over the years have changed from Screen Film Mammography to Digital Mammography allowing higher quality images and greater availability of data. With this greater availability, computational techniques have been introduced as part of the diagnostic workflow. These systems are classified as Computer Aided Detection or Diagnosis (CAD) systems [Markey & Bovik n.d.). Initial CAD systems centred around basic image enhancement and clustering techniques to flag the relevant cases by varying the image properties. This allows to better visualize the micro calcifications(MC) present in 30-50% of mammograms, which are indicators of potential cancer [Linguraru *et al.* 2006; Nishikawa *et al.* 1995; Tang *et al.* 2009].

More recently with the advent of high-powered computing, accessibility to data and large amounts of storage, researchers have turned to more advanced so-called Machine Learning (ML) techniques, with special emphasis in Deep Learning (DL) in the later years. Sampaio *et al.* [Borges Sampaio *et al.* 2011] theorized the possibility of using a simple convolutional neural network (CNN) in conjunction with a support vector machine (SVM) to initially improve the image and them extract certain objects outside the breast for classification.[Jadoon *et al.* 2017; Shen 2017]. For CNNs one of the most widely used methods is classification according to the BI-RADS scale, which ranges from 0 (needing more imagery), to 6 (known biopsy-proven malignancy) [He *et al.* 2017; Huang *et al.* 2019]. Other combinations of state-of-the-art architectures have been shown to be highly efficacious when used with mammogram images. Chin *et al.* [Chin & Liu 2019] used a combination of residual layers [He *et al.* 2016] and Inception V2 layers [Szegedy *et al.* 2015] to predict and safely monitor patients with pure atypical ductal hyperplasia on a dataset of 298 images and reached an AUC score of 0.86 concluding that a larger dataset would improve on these results further.

Kooi *et al.* [Kooi *et al.* 2017] also saw the need of more data for mammography analysis and collected a 44000 image dataset in the Netherlads to evaluate a CNN against the common CAD systems. They adapted the VGG architecture using a scaled down version and also did up to 16x augmentation on the image patches extracted from the mammograms meaning that around 1.3 Million images were used in training. This achieved a AUC score of 0.929 which was better when compared to 3 radiologist who had a mean AUC of 0.911. Finally and international evaluation was published in 2021 using over 25000 mammography images from the UK and the US as training data and it generated state of the best results so far in the field, outperforming the radiologists themselves [McKinney *et al.* 2020]. Since large amounts of highly curated data can be insurmountable to new stakeholders entering the field, we seek to show that with an insight in the data high accuracy can be achieved with a small dataset.

3 Method

In order to ensure the size of the CNNs is manageable and reduce the computational burden, we approached the problem of classifying anomalies in mammograms as a patch classifyer, similarly to Shen *et al*. [20]. The original dataset was divided into smaller patches for input into the algorithm. These patches are separated in two categories, anomalous and normal. The definition of an anomalous patch is if its centre of mass is within the area of anomaly set out in the original ground truth, the ground truth had 6 classes which we joined into one anomaly class. Negative patches are chosen at random in equal numbers as the anomalous patches to ensure a 50/50 split.

Since experiments are carried in a patch classifier format, we investigate and optimise the best data ingestion hyperparameters (patch size and step between patches) to ensure the best outcome in the algorithm comparison. As the scaled VGG16 had already shown promising results in mammogram anomaly detection [Kooi *et al.* 2017], this architecture was used initially as baseline to test three different step sizes and four different patch sizes. These experiments were done using 5-fold cross validation – 80/20 split for training and validation- with an additional 10% saved for the final testing of the best architectures. All algorithms were run on 100 epochs, the specific breakdown of hyperparameters and train, validation and test sets can be found in section 5.2.

After the initial CNN architecture was developed and refined using,, three state-of-art achitectures were compared: a breast tissue focused algorithm CancerNet [Rosebrock 2019] (a combination of Mobilenet[Howard *et al.* 2017] and the Xception[Chollet 2017] architecture),a scaled VGG16[Simonyan & Zisserman 2015] similar to Kooi *et al.* [Kooi *et al.* 2017] and ResNet50[He *et al.* 2016]. Once the best algorithm was defined for this particular

problem augmentation was also tested to see if there was any improvement.

4 Scaled VGG architecture

A scaled VGG model was used due to a reduced computational burden and the need or large dataset. Given that the mammograms analysed are greyscale, similar scaled algorithm methods have been shown to produce good results in mammogram analysis [Ribli *et al.* 2018; Tardy *et al.* 2019].

The basis to the VGG format is the convolutional layer that with a stride of one and padding can be described as follows:

$$\widehat{M}_{h,i,j} = \sum_{k,l,m} \widehat{K}_{k,l,m,j}\, \mathrm{C}_{h+k-1,i+l-1,m} \tag{1}$$

where $\widehat{K}$ is the convolutional kernel of size $N_K \times N_K \times T \times D$, N being the spatial dimension of the convolution K, T the number of input channels and D is the depth of the output channels. The standard convolution kernel K is multiplied by the input feature map C generating the feature map $\widehat{M}$.

The network used is composed of three blocks of two convolutional layers each with number of filters 64/64 - 128/128 - 256/256, and two blocks of three convolutional layers each with number of filters 512/512- 512/512. The kernels of all convolutional filters are set to 3 × 3, each convolutional layer has a ReLU activation and there is max pooling at the end of each block. Finally, after the convolutional network, there are two dense layers of 4096 with ReLU activation. For the classification the result of dense layers has been fed to a 2-classes prediction layer with sigmoid activation.

4.1 Dataset

The dataset used was the MIAS dataset [Lee *et al.* 2017], comprising of 322 mammogram images all of size 1024x1024 pixels. Each image has ground truth information by two radiologists where the background tissue, the abnormality, the severity of the abnormality and its location are described. The abnormalities are relatively evenly distributed in the dataset among:

- Calcification
- Well-defined/circumscribed masses
- Spiculated masses
- Other, ill-defined masses
- Architectural distortion
- Asymmetry
- Normal

with calcifications being the most represented and Ill-defined masses the least. The 6 abnormal classes where joined in to one class as from a clinician standpoint removing their need of seeing normal reduces workflow.

In accordance with the location of the anomaly on the ground truth, the images are then divided into smaller patches for ingestion into the algorithm. Patches with large amounts of background were removed through a simple averaging technique. These patches are divided into two categories, anomalous and normal. The definition of an anomalous patch is if the centre of mass is within the area of anomaly set out in the ground truth and the negative patches are chosen at random in equal numbers as the anomalous patches to ensure a 50/50 split.

4.2 Data augmentation

Data augmentation was applied at random as the patches were read into the model. Data augmentation techniques included flips, rotation, zoom, width and height shift, brightness, shear and contrast (see sample augmentation on one patch in Figure 1). This resulted in a 16x augmentation of the dataset.

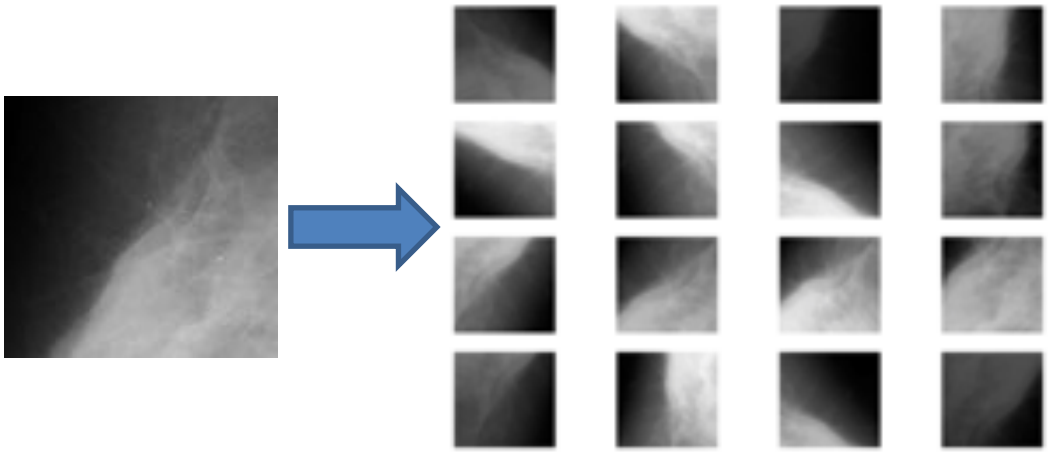


Figure 1 Random 16x augmentation of a mammogram patch.

5 Experimental section

In the field of cancer diagnosis, the overarching aim is to eliminate false negatives, in other words, miss no malignant anomalies. Sometimes the cost of eliminating false negatives can lead to an unreasonable amount of false positives which in turn causes additional worry and uncomfortable procedures on the patient's side, thus an analysis has to be done to ascertain the lowest values possible for both these parameters.

This study uses a patch classifier approach to ascertain the effects of varying the initial input hyperparameters and then compare the current state-of-the-art algorithms with and without augmentation.

For every test a 5-fold validation technique was used. First10% from the complete dataset was taken randomly for a final test dataset. With the remaining 90% divided up into five different groups, the algorithm was trained on 80% and validated on 20% five different times and then the results averaged over the 5 folds. The fold wise validation results were used to determine the hyperparameters and the testing set was used in the final algorithm comparison.

Using the scaled VGG16 architecture, an analysis of hyperparameters was carried out.The two main input parameters, Patch Size and the Step across the image, were varied independently.. The step was varied between 4, 8 and 16, then the patch size between 28x28, 56x56, 112x112 and 224x224. Other hyperparameters used in these experiments were fixed and using the following values: Batch size: 64, Optimizer: SGD, Loss: Binary Crossentropy, Learning rate: 0.001 w/ decay, Dropout: no, Augmentation: no.

Table 1 Step Hyperparameter variation results VGG16

Step	*Accuracy*	*F1 Score*	*AUC*
4	0.9930	0.9930	0.9931
8	0.9853	0.9853	0.9660
16	0.8278	0.8261	0.8314

Table 1 shows the results for a patch size of 224x224 and varying step sizes. As we augment the step size and thus reduce the number of training patches the performance of the algorithm decreases, varying from over 137000 with a step of 4 to 8675 with a step of 16. This means a lot more training data is available for the smaller step input and can lead to a more robust algorithm.

Table 2 shows the results for a step size of 4 and varying patch sizes. The 28x28 patch is a lot worse as it will not have as much contextual information. The other patch sizes seem to be largely equivalent but going forward we maintain the original intended patch size for the VGG algorithm of 224x224.

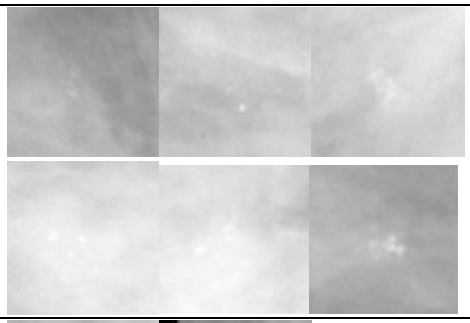
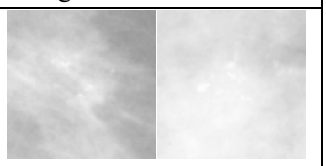
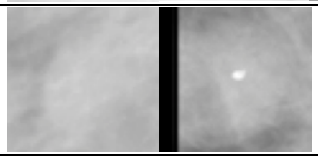
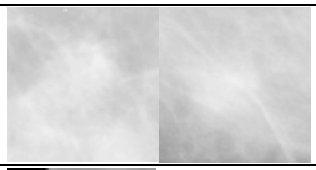
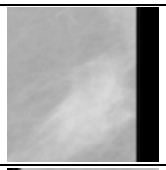
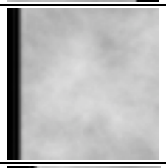
Table 2 Patch Hyperparameter variation results VGG16

Patch Size	*Accuracy*	*F1 Score*	*AUC*
28x28	0.8953	0.8612	0.8803
56x56	0.9935	0.9935	0.9968
112x112	0.9934	0.9934	0.9971
224x224	0.9930	0.9930	0.9931

5.1 Patch classification analysis

As already discussed in the dataset section, the anomalies are divided into 6 classes according to the dataset ground truth and are represented relatively evenly in the training dataset, except the calcification class which is around 7% above average. Upon analysis of the most common false negative patches, calcifications appeared over 2.5x more than each of the other anomalies despite a higher representation in the training data. Visual examination reveals that the calcifications can be very faint and occluded under other brighter tissue.

Table 3 Most common false negative patches across all algorithms divided by ground truth classification

Class of abnormality	Benign	Malignant
CALC – Calcification Around 20px in size		
CIRC – Well-defined/circumscribed masses		
SPIC – Spiculated masses		
MISC – Other ill-defined masses		
ARCH – Architectural distortion		
ASYM – Asymmetry		

5.2 Best model comparison

In this section, we investigate the performance of different state-of-the-art architectures when looking at the MIAS

mammogram dataset and using the testing dataset for final scores. For the purpose of this study 4 different architectures were chosen Cancernet, Scaled VGG16, Scaled VGG16 with augmentation and ResNet50 as already discussed. Each architecture was trained over 100 epochs on 100,800 images, validated on 25,200 and tested on 14,000. All had 50/50 split between the normal and abnormal class.

Table 1: Comparison between state-of-the-art algorithms on the MIAS dataset on testing data

	Hyperparameters	Accuracy	F1 Score	AUC Score
Cancernet	**Patch size:** 224x224 **Image Step:** 4 **Batch size:** 32 **Optimizer:** Adagrad **Loss:** Binary Crossentropy **Learning rate:** 0.01 w/ decay **Dropout:** 0.25	0.9398	0.9397	0.85
VGG16	**Patch size:** 224x224 **Image Step:** 4 **Batch size:** 64 **Optimizer:** SGD **Loss:** Binary Crossentropy **Learning rate:** 0.001 w/ decay **Dropout:** no	0.9930	0.9930	0.9931
ResNet50	**Patch size:** 224x224 **Image Step:** 4 **Batch size:** 128 **Optimizer:** SGD **Loss:** Binary Crossentropy **Learning rate:** 0.001 w/ decay **Dropout:** no	0.9916	0.9903	0.9902

Results are shown in Table 4. As the Cancernet algorithm is one of the simplest it is to be expected that it would generate the lowest accuracy but was a good comparison point as it was developed for this very problem of breast cancer.

An additional test (Table 5) was done with the best architecture to add some 16x data augmentation this was also trained over 100 epochs but with 2.4 million images, and smaller validation and test sets of 37,800 and 21,000 respectively. All had 50/50 split between the normal and abnormal class

Table 2: Comparison between state-of-the-art algorithms on the MIAS dataset on testing data

	Hyperparameters	Accuracy	F1 Score	AUC Score
VGG16 with 16x Augmentation	**Patch size:** 224x224 **Image Step:** 4 **Batch size:** 512 **Optimizer:** SGD **Loss:** Binary Crossentropy **Learning rate:** 0.001 w/ decay **Dropout:** no	0.9935	0.9925	0.9948

Overall, the VGG16 with augmentation seemed to be best at the patch classification but it is such a small improvement in accuracy and AUC are hardly worth the additional computing time of 16x more training data and run the danger of overfitting.

6 Conclusions

In this paper we have proposed and investigation on different CNN architectures and input hyperparameters to perform patch-level mammography anomalous classification. We particularly focus on the most usual case of limited amount of images in training due to privacy concerns. Out of our study, these are the most relevant conclusions:

- The parameter that makes the biggest difference to the quality of the model is the number of training patches.
    - Smaller datasets as in the 16 step model are more susceptible to abrupt colour variations of man-made artifacts.
    - The largest datasets such as the step 4 model are robust enough to identify changes in tissue morphology and ignore that man-made artifacts.
- Dense background tissue causes the algorithm the most problems from two aspects:
    - False positives – around 60% of them happened in dense tissue as it can in general have a more tumour like aspect.
    - False negative microcalcifications – around 60% of false negatives were microcalcifications all in dense tissue due to the fact of the calcifications are very faint the dense tissue can hide them form the model.

References

Borges Sampaio, W., Moraes Diniz, E., Corrêa Silva, A., Cardoso de Paiva, A., & Gattass, M. (2011). Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Computers in Biology and Medicine*, **41**(8), 653–664.

Cancer, C. on B. (2006). Screening for breast cancer in England: Past and future. *Journal of Medical Screening*, **13**(2), 59–61.

Chin, C., & Liu, M. Z. (2019). Accuracy of Distinguishing Atypical Ductal Hyperplasia From Ductal Carcinoma In Situ With Convolutional Neural Network– Based Machine Learning Approach Using Mammographic Image Data, (May), 1166–1171.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, **2017-Janua**, 1800–1807.

E L Thurfjell, K A Lernevall, A. A. T. (1994). Benefit of independent double reading in a population-based mammography screening program., **191**(1).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2016-Decem**, 770–778.

He, T., Puppala, M., Ogunti, R., … Wong, S. T. C. (2017). Deep learning analytics for diagnostic support of breast cancer disease management. *2017 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2017*, 365–368.

Howard, A. G., Zhu, M., Chen, B., … Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*.

Huang, Y., Han, L., Dou, H., … Yin, G. (2019). Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images. *BioMedical Engineering Online*, **18**(1), 1–18.

Jadoon, M. M., Zhang, Q., Haq, I. U., Butt, S., & Jadoon, A. (2017). Three-Class Mammogram Classification Based on Descriptive CNN Features. *BioMed Research International*, **2017**. doi:10.1155/2017/3640901

Kerlikowske, K., Carney, P. A., Geller, B., … Ballard-Barbash, R. (2000). Performance of screening mammography

among women with and without a first-degree relative with breast cancer. *Annals of Internal Medicine*, **133**(11), 855–863.

Kooi, T., Litjens, G., van Ginneken, B., … Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, **35**, 303–312.

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). Data Descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, **4**, 1–9.

Linguraru, M. G., Marias, K., English, R., & Brady, M. (2006). A biologically inspired algorithm for microcalcification cluster detection. *Medical Image Analysis*, **10**(6), 850–862.

Logan, R. F. A., Patnick, J., Nickerson, C., Coleman, L., Rutter, M. D., & Von Wagner, C. (2012). Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut*, **61**(10), 1439–1446.

Markey, M. K., & Bovik, A. C. (n.d.). Computer-Aided Detection and Diagnosis in Mammography. *Handbook of Image and Video Processing*.

McKinney, S. M., Sieniek, M., Godbole, V., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, **577**(7788), 89–94.

Nishikawa, R. M., Giger, M. L., Doi, K., Vyborny, C. J., & Schmidt, R. A. (1995). Computer-aided detection of clustered microcalcifications on digital mammograms. *Medical and Biological Engineering and Computing*, **33**(2), 174–178.

R E Bird, T W Wallace, B. C. Y. (1992). Analysis of cancers missed at screening mammography., **184**(3).

Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Scientific Reports*, **8**(1), 1–7.

Rosebrock, A. (2019). Breast cancer classification with Keras and Deep Learning. Retrieved February 25, 2021, from https://www.pyimagesearch.com/2019/02/18/breast-cancer-classification-with-keras-and-deep-learning/

Shen, L. (2017). End-to-end Training for Whole Image Breast Cancer Diagnosis using An All Convolutional Design. *ArXiv E-Prints*, **3000**, 1–10.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.

Szegedy, C., Liu, W., Jia, Y., … Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **07-12-June**, 1–9.

Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I. E., & Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *IEEE Transactions on Information Technology in Biomedicine*, **13**(2), 236–251.

Tardy, M., Scheffer, B., & Mateus, D. (2019). Breast density quantification using weakly annotated dataset. *Proceedings - International Symposium on Biomedical Imaging*, **2019-April**(Isbi), 1087–1091.

WHO. (2020). Imaging modalities. Retrieved March 11, 2020, from https://www.who.int/diagnostic_imaging/imaging_modalities/en/

WHO cancer key facts. (n.d.). Retrieved March 11, 2020, from https://www.who.int/news-room/fact-sheets/detail/cancer

Finding people in GPS denied environments using an autonomous drone

James O'Donnell & Gerard Lacey

School of Computer Science and Statistics, Trinity College Dublin

Abstract

Drones are used extensively for outdoor search and rescue as they allow teams to survey large areas efficiently without putting themselves at risk. However, complex underground urban, cave and tunnel systems present significant challenges for first responders. These GPS-denied environments are hazardous for personnel and difficult to navigate, especially in time-sensitive search and rescue scenarios. The DARPA Subterranean Challenge aims to explore new approaches to map, navigate and search these complex underground environments using autonomous robots. This paper describes the development of a visually guided autonomous drone with search and rescue capabilities. The drone uses a 3 layer architecture with obstacle avoidance, navigation and people detection. Obstacle avoidance is trained using deep reinforcement learning. Navigation uses ORB-SLAM2 and people detection is achieved using YOLO3. Tests in an indoor environment achieved a mapping accuracy of 0.36m±0.12 and people were detected in 85% of trials.

Keywords: Autonomous drones, robotics, SLAM, people detection, DARPA Subterranean Challenge,

1 Introduction

In search and rescue (SAR), time is often the most critical factor and is coupled with the uncertainty of the location of the missing people. Therefore, SAR services must be able to search large areas in a short space of time. Typically, large powerful drones are used, each capable of carrying heavy payloads such as searchlights, loudspeakers, high-quality sensors, and powerful wireless communications [Balta et al., 2017]. A greater challenge occurs in tightly confined indoor or underground areas such as semi-collapsed buildings, caves, or subways. In this scenario, small lightweight autonomous drones could search these GPS-denied environments more efficiently than a large piloted drone. A multi-robot approach could also be utilised, with a ground-based robot acting as a base station for a swarm of highly manoeuvrable drones.

The DARPA Subterranean Challenge (SubT) [DARPA, 2018] takes place in human-made tunnel systems, urban underground environments, and natural cave networks. The teams demonstrate their robotic solutions focusing on autonomy, networking, perception and mobility, to compete for a $2 million prize. The competition has run over the course of 3 years with the final event due to take place in September 2021. The objectives of this paper are to develop an autonomous drone to find people in GPS-denied environments. We use a low-cost drone with a monocular camera and an internal inertial measurement unit (IMU). We evaluate the accuracy of the SLAM map and the accuracy of object detection and identify improvements needed to address tasks similar to the DARPA Subterranean Challenge.

2 State of the Art

Robotics search and rescue range from specialised drones that then change shape [Falanga et al., 2018] to networks of drones [Balta et al., 2017]. This is a very active research area with research focused on full autonomy [Sandino et al., 2020] or novels sensors such as omni-directional video [Valenti et al., 2018]. The focus of our research is the SAR capability of a single drone with a focus on autonomous navigation and people detection as we intend our drone(s) to be used in partnership with our mobile robot base as in Figure 1. Our approach is inspired by the teams competing in the DARPA Subterranean challenge.

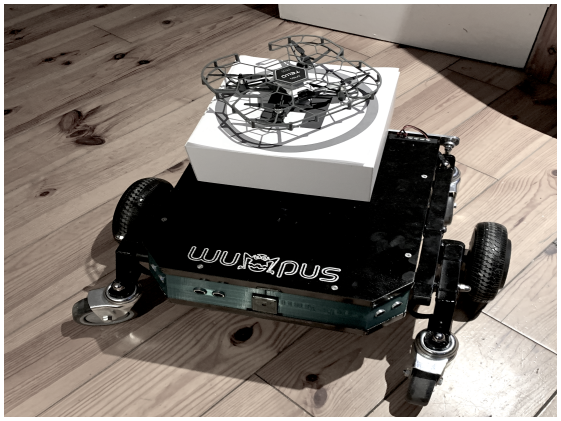


Figure 1: The Tello drone a Wumpus robot base

2.1 DARPA Subterranean Challenge

Each of the 12 teams competing in the DARPA SubT Challenge use a combination of SLAM and visual object recognition. Team 'Explorer' [CMU & OS, 2020], from Carnegie Mellon and Oregon State, use a multi-robot approach with ground-based robots and a drone. They employ multiple sensors, including a depth sensing camera, a thermal imaging camera, an IMU and a microphone array. For communications, they drop small communication nodes out of the back of the ground-based robot as it travels along to establish an ad-hoc mesh network. For mapping and localisation, they use multi-robot SLAM. Team CoSTAR [Agha et al., 2021] from the NASA Jet Propulsion Laboratory and a consortium of universities, came runner up in the first challenge and won the second Urban Circuit challenge. CoSTAR have a much wider range of robots, including legged, flying, wheeled and tracked robots. Team CERBERUS made up of several research institutes including the Oxford Robotics Institute, employs legged-wheeled robots [Bjelonic et al., 2020]. This is a hybrid approach uses wheels or legs depending on the terrain. A few of the other teams have also published papers describing their approaches [Rouček et al., 2019] [Hudson et al., 2021].

2.2 Obstacle Avoidance with Reinforcement Learning

To avoid obstacles the robot must choose its next action based on its current state estimate to avoid a collision. Q-Learning [Watkins & Dayan, 1992] iteratively updates each state-action pair to find the optimal action to take for a given state. A Q-Learning table is used to map states to actions and finds an optimal policy by maximising the expected value of the total reward over all states. A neural network can be used to approximate the Q function mapping states to Q-values. Deep Q-learning Networks (DQN) was introduced by Google's DeepMind [Mnih et al., 2015] to learn to play ATARI games. A weakness of DQN is that experience replay memory is sampled uniformly but in [Schaul et al., 2015] experiences that lead to the greatest change in the network are prioritised. Q-learning can sometimes overestimate the action values in noisy environments, slowing the learning. In [Van Hasselt et al., 2016] Double Q-Learning and DDQN are combined to reduce overestimation and improve performance.

2.3 Simultaneous Localisation and Mapping (SLAM)

SLAM has made significant progress over the last 30 years by enabling large-scale real-world applications from self-driving cars to UAVs. ORB-SLAM [Mur-Artal et al., 2015] and the more recent ORB-SLAM2 [Mur-Artal & Tardos, 2017] are feature-based SLAM implementations for monocular, stereo and RGB-D cameras. ORB-SLAM pre-processes input images and extracts features as keypoints. The input images are then discarded and all operations take place using only these acquired features. ORB-SLAM2 works simultaneously

on tracking, local mapping and loop closure. Tracking involves estimating the pose of the camera and deciding when to insert a new keyframe. Mapping is the process of updating map points and keyframes. Loop closure includes candidate keyframe selection for potential loops. The algorithm uses ORB features [Rublee et al., 2011] to perform feature matching. ORB uses a multi-scale image pyramid for feature detection using the FAST feature detector to detect keypoints at each of these levels. The matched keypoints are converted to a binary feature vector using BRIEF. However, ORB modifies the original implementation of BRIEF to handle rotations.

There are two widely used metrics for calculating the accuracy of SLAM algorithms on these benchmarks: absolute trajectory error (ATE) [Sturm et al., 2012] and relative pose error (RPE) [Geiger et al., 2013]. ATE is suited for assessing the performance of SLAM systems while RPE is suited for measuring the drift of visual odometry systems [Sturm et al., 2012]. The absolute trajectory error measures the difference between the estimated trajectory and the ground truth trajectory.

2.4 Vision-based Object Detection

There are two categories of object detectors: one-stage and two-stage detectors. The two-stage detectors have high localisation and object recognition accuracy while the one-stage detector achieves faster inference speed. Two-stage detectors, such as Faster R-CNN [Ren et al., 2016], introduce a Region Proposal Network (RPN) to generate regions of interest in the first stage. The second stage performs object classification and bounding box regression. In a one-stage detector such as YOLO [Redmon et al., 2016] and SSD [Liu et al., 2016], object detection is treated as a regression problem. It learns the class probabilities and bounding box coordinates directly from the input image. These have lower accuracy but are much faster. YOLO (You Only Look Once) results in lower accuracy than Faster R-CNN but faster, e.g. 45 fps than Faster R-CNN with 7 fps. YOLO has evolved into YOLO V2 [Redmon & Farhadi, 2017] which capable of predicting 9,000 classes of objects, and in 2018, YOLO v3, [Redmon & Farhadi, 2018] with changes to the underlying architecture to improve performance.

The evaluation criteria for object detection consists of frames per second (fps), precision and recall (true positive rate). The most frequently used evaluation metric for object detection for a single object class is average precision (AP) and to compare performance across all object categories, the mean AP (mAP) is adopted. To measure the object localisation accuracy, 'intersection over union' (IoU) is used. This is the overlap ratio between the predicted bounding box and the ground truth box. The ratio is typically greater than a predefined threshold of 0.5 [Liu et al., 2019].

3 System Design

The robot system consisted of a low-cost Tello drone and a MacBook Air with a 1.6 GHz Intel Core i5 processor and an Intel HD Graphics 6000 graphics card. The Tello performed three key tasks: obstacle avoidance, mapping and object detection. The architecture is based on three layers: Functional, Executive and Planner. The Functional layer provides autonomous obstacle avoidance and typically interacts with the hardware. Obstacle avoidance is trained using the DQN algorithm in a PEDRA simulator and then fine-tunes this in the real-world environment. The Executive layer uses ORB-SLAM2 to build a map and localise the drone in real-time. Finally, the Planning layer uses YOLO3 to recognise objects of interest and mark the location on the map.

The system is developed using ROS Melodic on Ubuntu 18.04. Each of the ROS nodes can be developed and tested independently in ROS and the dataflow is managed using the publisher/subscriber model. To connect to the Tello drone, the "Flock driver" is used. The flight data e.g. the battery state, pose (pitch, roll, yaw, x y z), flight mode, equipment status and temperature.

3.0.1 Obstacle Avoidance

Obstacle Avoidance is trained using Deep Reinforcement Learning[Anwar & Raychowdhury, 2019] using a Sim-to-Real approach. The PEDRA simulator is used to train the drone using the DQN algorithm. Our work

builds upon [Li, 2020] which showed that a drone trained in the PEDRA simulator yielded much better results than a drone trained in the Gazebo simulator. This was due to the realistic visual features available in PEDRA environments.

The field of view of the drone's camera is divided into $n \times n$ windows. Each of these windows corresponds to an action in the action space. The drone implements the next action by selecting the action's pitch and yaw and moves forward 0.5m to the next state. A uniform noise is added to the pitch and yaw angles to make these actions probabilistic instead of deterministic. This results in a maximum possible distance of 0.2m from the target position to the actual position. The reward for taking an action from the current state is based upon the depth estimation in that direction. The drone, therefore, chooses the action which takes it towards the window, in its field of view, for which there is the greatest estimated depth in the environment. For the real-world system the action space is reduced 3 actions: advance by 0.5m, rotate clockwise by 45 degrees, and rotate clockwise by 45 degrees. The action space does not include any altitude adjustments.

The deep reinforcement algorithm DQN is used to train the drone to detect obstacles and navigate around them. The mission of the drone is to fly as far as possible in the environment without crashing. The drone learns the state-action pairs and converges to a policy that maximises the expected cumulative reward. The velocity is set to zero after each step to allow the drone time to observe the state. The agent gets a positive reward for navigating to a state with no obstacles in front of it and a negative reward for getting too close to an obstacle or crashing. All of the layers of the deep neural network are trained in the simulator the last two layers are then trained in the real world while all the other layers are frozen. This fine-tuning helps to adapt the network to the real-world environment.

3.1 SLAM

The ORB-SLAM2 algorithm [Mur-Artal & Tardos, 2017] is chosen as ORB feature extraction and matching mean that accurate mapping can be performed in real-time on standard CPUs achieving state-of-the-art accuracy across the KITTI, EuRoC and TUM benchmarks. ORB-SLAM2 is implemented using the Tello_ROS_ORB-SLAM framework [Autonomous Drones Lab, 2019]. To provide a real-world scale to the ORB-SLAM2 algorithm, the drone must perform a calibration at the start of each flight. Upon take-off and successful initialisation of the SLAM, the drone is instructed to fly vertically up a distance of 0.5m metres and down again. The readings from the Tello's internal altimeter are sampled and the height published by the SLAM algorithm.

3.2 People Detection

YOLOv3 was chosen for people detection as it achieves the best speed-accuracy balance on the MS COCO dataset with a mAP of 57.9% in 51ms. The YOLO implementation for ROS by ETH Zurich was used [at ETH Zürich, 2018]. YOLOv3 was configured by adjusting the confidence threshold to 80%. The detection classes were limited to objects that could be present in the DARPA SubT Challenge such as "person", "backpack" and "cell phone". When an object is detected in the current frame, its location is then imposed as a marker on the 2D point cloud map. The object must be reported with an accuracy of 5 metres. When the object is detected in front of the drone it is first predicted to be positioned next to the identified features in the current frame. This prediction is updated as the drone moves closer to the object.

4 Performance Evaluation

4.1 Obstacle Avoidance Performance

The DQN model is trained end-to-end for 150,000 iterations, making up ~5000 episodes in the virtual environment of the PEDRA simulator. The safe flight distance is calculated as the distance the drone can travel in the environment without crashing. The last 2 layers of the model are trained for 20 episodes in a real-world indoor environment using the Tello drone. Following the training, the drone is tested in both the simulator and in the real world. The mean safe flight (MSF) distance is calculated over 20 flights in each environment.

After training the drone achieves an MSF distance of 45.3 m in the simulator. After fine-tuning these weights with the Tello in the real world, the MSF distance achieved is only 15.7m. On approximately 80% of all of these real-world trials, the trial is terminated because the drone's battery is depleted and the drone is forced to land. It could be expected that the drone would perform worse in the real-world environment as there are several other factors such as noise and different lighting conditions which may upset the model's calculations. The impediments introduced by the very short battery life make it difficult to evaluate the drone's real-world autonomous navigation performance fully.

4.2 SLAM Accuracy

To assess the accuracy of the scaled SLAM map, the SLAM estimation of trajectory travelled is compared to the ground truth. The Absolute Trajectory Error (ATE) RMSE is the metric calculated for trails along 10 different routes. The RMSE is the Root Mean Square Error. For each route, the drone starts at a different point within the indoor environment. Some of these routes contain loops in which the drone is expected to perform loop closure. The average ATE RMSE accuracy was 0.36m±0.12.

4.3 Object Detection Accuracy

The drone is tasked with detecting a person and accurately marking the position on the generated map in real-time. For this experiment, there are 3 routes. The drone travels each route 20 times with the person positioned in a different location along the route each time. A successful detection occurs when the system places a marker in real-time on the generated map at the person's position to within 5 metres to count as a successful detection. An average detection rate of 85% was achieved over 20 trials.

5 Discussion

5.1 Computational Load

It was very computationally expensive to perform obstacle avoidance as the system uses two deep neural networks, one for creating a depth map from the RGB image and one for the DQN algorithm. Furthermore, obstacle avoidance was run parallel with SLAM and object detection. A possible solution to this would be to assign each of the ROS nodes processes to different CPUs to allow them all to run in parallel in real-time. There is no functionality within ROS to support this but [Wei et al., 2016] presents a ROS architecture that allows different ROS nodes to be run on different processor cores also tools such as "taskset" for Linux allows the user to pin tasks to CPUs.

5.2 SLAM Accuracy

The most accurate result is achieved on Route 6, with an ATE RMSE value of 0.11m while the least accurate result is achieved on Route 4 with an ATE RMSE value of 0.53m. The accuracy from both these results is good enough to enable the system to detect an object of interest within 5m subsequently. Over long flights, however, if loop closure is not detected, this inaccuracy becomes worse. These results are poorer than the original ORB-SLAM paper with a monocular camera [Mur-Artal et al., 2015] which reports an ATE RMSE on the TUM RGB-D dataset ranging from 0.0124m to 0.0345m. This discrepancy is largely because the real-world scale is known, whereas it has to be calculated by our drone. Moreover, conditions such as blur, missed frames and strong rotation also occur on the test routes. Monocular SLAM algorithms including ORB-SLAM, do not perform well under pure rotations of the camera. During many of the tests, the SLAM algorithm loses track of the features and the mapping fails. On some blank walls, there is a lack of features for ORB-SLAM2 to track. In this case, the object detection evaluation test has to be restarted. To prevent this from occurring, a black and white checkerboard pattern, printed on A4 paper, is stuck on different walls to create a more feature-rich environment.

5.3 Object Detection Accuracy

For each of the three routes in the indoor environment, the experiment begins after SLAM initialises and the real-world scale has been calibrated. Route 2 has the highest rate of successful object detection with the person being detected in 95% of the tests. This is likely due to slightly better lighting along this route and the person being in positions relative to the drone that were easier to detect. A successful detection is only counted when the person's position is estimated to be within 5 metres of the person's actual position. This accuracy is one of the requirements in the SubT Challenge. Over a longer flight duration, however, the scale drift becomes more of an issue and would cause the person's position to be less accurate. Flights with long duration were not assessed because of limited battery life.

5.4 Limitations of the System

Further improvements needed for the system to be successful in SubT type challenges and are examined from the point of view of the technological goals of the SubT Challenge: mobility, autonomy, perception, and networking. The flight time of the Tello is 13 minutes. Moreover, about 40 seconds are taken up by calibration after take-off. The drone rises an extra 0.5m from its take-off height of 1m to perform calibration. It is worth also noting that the drone would be taking off from its mounted position on top of the ground-based robot. These time and height constraints severely impact the drone's mobility. Our system is currently not fully autonomous and requires the user to watch the drone's performance closely and the difficulties encountered with the computational load prevented full autonomous testing. ORB-SLAM2 with the Tello's monocular camera provides sufficient accuracy for the SubT Challenge, it is evident from the experiments that the algorithm fails often, in terms of initialisation and feature tracking. In the SubT Challenge, the drone may get only one chance to launch and perform its mission. Relying on a single low-cost sensor for perception limits the robustness of the system. Finally, the Tello only uses the 2.4GHz band WiFi with a 100m range. In practice, the range is considerably less, especially when the line of sight is not maintained. The Tello, could not travel far out of the line of sight of the ground-based robot without an improved networking solution.

6 Conclusion

This research aimed to develop and evaluate a SAR system using a low-cost drone with a monocular camera in GPS-denied environments. The drone would act as an auxiliary robot for a ground-based robot in the DARPA Subterranean Challenge. The drone system that was developed completes the mapping and victim detection tasks to sufficient accuracy for the SubT Challenge. The drone was successfully trained for autonomous navigation and obstacle avoidance with a simulation-to-real approach using an optimised deep reinforcement learning. The fully autonomous performance could not be assessed due to battery life limitations and the computational load of running all the processes on the laptop.

The benefits of using a monocular camera and an IMU to operate in GPS-denied environments are low weight and power consumption. Light-weight drones are more agile and enable longer flight times. The challenges of scale ambiguity, scale drift and robustness were also recognised when performing SLAM with a monocular camera. The results from the experiments showed that the SLAM mapping was accurate to 0.36m$\pm$0.12, this accuracy was sufficient for enabling the subsequent task of people detection, but that this error could accumulate over longer flights. The difficulties encountered during experiments, associated with SLAM initialisation and tracking, demonstrated the lack of robustness of the SLAM algorithm with a monocular camera. The ORB-SLAM2 would achieve better performance with either a stereo or RGB-D camera and provide depth measurement for obstacle avoidance.

The object detection algorithm, YOLOv3, performed very well. The mean detection rate of the victim in the environment across the trials was 85%. This demonstrates that the algorithm's balance between speed and accuracy makes it a good choice for this real-time application. It is clear why many of the teams currently taking part in the DARPA Subterranean Challenge use this object detection algorithm.

The are many exciting improvements that can be made to the system and several tests to consider in the areas of mobility, autonomy, perception and networking. Building hardware and software systems is an iterative process. The future work for this project should be undertaken with careful consideration of the complete multi-robot SAR system with a holistic view of how the drone operates within an overall multi-robot system.

References

[Agha et al., 2021] Agha, A., Otsu, K., Morrell, B., Fan, D. D., Thakker, R., Santamaria-Navarro, A., Kim, S.-K., Bouman, A., Lei, X., Edlund, J., et al. (2021). Nebula: Quest for robotic autonomy in challenging environments; team costar at the darpa subterranean challenge. *arXiv preprint arXiv:2103.11470*.

[Anwar & Raychowdhury, 2019] Anwar, A. & Raychowdhury, A. (2019). Autonomous navigation via deep reinforcement learning for resource constraint edge nodes using transfer learning.

[at ETH Zürich, 2018] at ETH Zürich, R. S. L. L. R. (2018). Yolo ros. `https://github.com/leggedrobotics/darknet_ros`. Accessed: 14-06-2021.

[Autonomous Drones Lab, 2019] Autonomous Drones Lab, T. A. U. (2019). Tello_ros_orbslam. `https://github.com/tau-adl/Tello_ROS_ORBSLAM`. Accessed: 14-06-2021.

[Balta et al., 2017] Balta, H., Bedkowski, J., Govindaraj, S., Majek, K., Musialik, P., Serrano, D., Alexis, K., Siegwart, R., & Cubber, G. (2017). Integrated data management for a fleet of search-and-rescue robots. *J. Field Robot.*, 34(3), 539–582.

[Bjelonic et al., 2020] Bjelonic, M., Sankar, P. K., Bellicoso, C. D., Vallery, H., & Hutter, M. (2020). Rolling in the deep–hybrid locomotion for wheeled-legged robots using online trajectory optimization. *IEEE Robotics and Automation Letters*, 5(2), 3626–3633.

[CMU & OS, 2020] CMU & OS (2020). Team explorer, fall 2020 update. `https://www.subt-explorer.com/blog`. Accessed: 14-06-2021.

[DARPA, 2018] DARPA (2018). Darpa subterranean challenge. `https://www.subtchallenge.com`. Accessed: 14-06-2021.

[Falanga et al., 2018] Falanga, D., Kleber, K., Mintchev, S., Floreano, D., & Scaramuzza, D. (2018). The foldable drone: A morphing quadrotor that can squeeze and fly. *IEEE Robotics and Automation Letters*, 4(2), 209–216.

[Geiger et al., 2013] Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11), 1231–1237.

[Hudson et al., 2021] Hudson, N., Talbot, F., Cox, M., Williams, J., Hines, T., Pitt, A., Wood, B., Frousheger, D., Surdo, K. L., Molnar, T., et al. (2021). Heterogeneous ground and air platforms, homogeneous sensing: Team csiro data61's approach to the darpa subterranean challenge. *arXiv preprint arXiv:2104.09053*.

[Li, 2020] Li, H. (2020). Simulation-to-real learning to control a visually guided autonomous drone. Msc dissertation, School of Computer Science and Statistics, Trinity College Dublin.

[Liu et al., 2019] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2019). Deep learning for generic object detection: A survey.

[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. *Lecture Notes in Computer Science*, (pp. 21–37).

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540), 529–533.

[Mur-Artal et al., 2015] Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5), 1147–1163.

[Mur-Artal & Tardos, 2017] Mur-Artal, R. & Tardos, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection.

[Redmon & Farhadi, 2017] Redmon, J. & Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263–7271).

[Redmon & Farhadi, 2018] Redmon, J. & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[Ren et al., 2016] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks.

[Rouček et al., 2019] Rouček, T., Pecka, M., Čížek, P., Petříček, T., Bayer, J., Šalanskỳ, V., Heřt, D., Petrlík, M., Báča, T., Spurnỳ, V., et al. (2019). Darpa subterranean challenge: Multi-robotic exploration of underground environments. In *International Conference on Modelling and Simulation for Autonomous Systesm* (pp. 274–290).: Springer.

[Rublee et al., 2011] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). Orb: an efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2564–2571).

[Sandino et al., 2020] Sandino, J., Vanegas, F., Gonzalez, F., & Maire, F. (2020). Autonomous uav navigation for active perception of targets in uncertain and cluttered environments. In *2020 IEEE Aerospace Conference* (pp. 1–12).

[Schaul et al., 2015] Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

[Sturm et al., 2012] Sturm, J., Engelhard, N., Endres, F., Burgard, W., & Cremers, D. (2012). A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*.

[Valenti et al., 2018] Valenti, F., Giaquinto, D., Musto, L., Zinelli, A., Bertozzi, M., & Broggi, A. (2018). Enabling computer vision-based autonomous navigation for unmanned aerial vehicles in cluttered gps-denied environments. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 3886–3891).

[Van Hasselt et al., 2016] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

[Watkins & Dayan, 1992] Watkins, C. J. & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279–292.

[Wei et al., 2016] Wei, H., Shao, Z., Huang, Z., Chen, R., Guan, Y., Tan, J., & Shao, Z. (2016). Rt-ros: A real-time ros architecture on multi-core processors. *Future Generation Computer Systems*, 56, 171–178.

Comparing the automatic evaluation of CPR compression rates using a smartwatch vs a smartphone.

Senan d'Art and Kenneth Dawson-Howe

School of Computer Science and Statistics,
Trinity College Dublin, Ireland

Abstract

This paper presents a novel algorithm for automatically evaluating the compression rate for Cardio Pulmonary Resuscitation (CPR) from the sensor data from a smartwatch. The results are compared with manually derived ground truth and with the results of an automatic system based on the analysis of video from a smartphone.

Keywords: Image Processing, CPR, Motion Analysis

1 Introduction

For Out of Hospital Cardiac Arrest (OHCA), the likelihood of survival is quite slim. A 2017 study looking at 28 European countries showed an average survival rate of 8% [Gräsner, et al., 2020] and the percentage of cases where a bystander performed CPR ranged from 13% to 82%. The application of CPR early after the onset of cardiac arrest is crucial. The percentage of patients who survive 1 month after experiencing cardiac arrest is 2.2% in cases where no CPR is performed, but increases when CPR is performed by a layperson (4.9%) or a healthcare professional (9.2%) [Herlitz et al., 2005].

The use of real-time feedback during training substantially increases the quality of CPR being performed [Baldi et al., 2017]. There are several existing solutions for providing feedback on CPR quality during training but these are typically expensive and as a result inaccessible to most organisations. The focus of the research here is the provision of feedback using commonly available technologies (such as smartphones and smartwatches). If a smartwatch or smart phone can provide reliable feedback to a rescuer performing CPR, whether in training or also in OHCA, it could result in increased survival rates.

2 State of the Art

Arguably the most important aspects of CPR are the Chest Compression Rate (CCR) which should be 100-120 compressions per minute and the Chest Compression Depth (CCD) which should be 5- 6cm. Many defibrillators are capable of measuring the CCR and CCD but as these will not, at least initially, be available other technologies have been investigated. The CCR can be computed reliably from a smartphone camera facing the person performing CPR [Corkery and Dawson-Howe, 2019] (See Figure 1) or using a view looking upwards at the person applying CPR [Meinich-Bache et al., 2017], where the smart-phone is lying flat on the ground. Other work has demonstrated the use of smartphone accelerometers (which also exist in smart

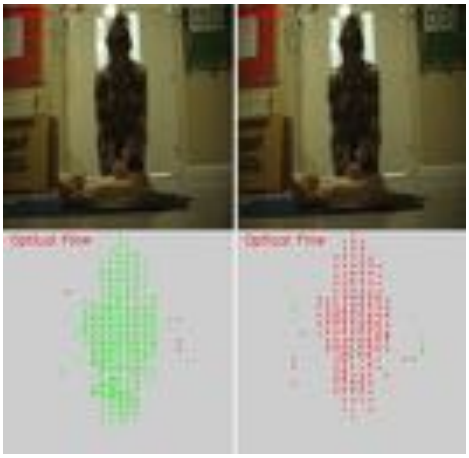


Figure 1: Analysis of CPR using optical flow [Corkery & Dawson-Howe, 2019]

watches) to evaluate the CCD [Song et al., 2015] with a resulting inaccuracy of only ~3mm. Ahn et al. have also demonstrated the use of smartwatches for providing feedback for externally calculated CCR & CCD [Ahn et al., 2017].

3 Detecting compressions using a smartwatch

The watch used in this project was an LG G Watch W100 which provides information on acceleration (including the force of gravity) as well as the force of gravity itself, rotational information and magnetic orientation information. The sensors yielded acceleration information in 3 axes: X, Y and Z. If the watch is worn on the outside of the left arm, the positive X-axis points from the elbow to the fingertips, the positive Y-axis points across from the thumb to the little finger and the positive Z-axis points directly outwards from the watch face. For the detection of compressions the principle axis is along the arm wearing the watch of the person giving the compressions which corresponds to the X axis. Typically that arm will be almost vertical as the compressions are done downwards. The forces recorded in the Y and Z axes are primarily caused by the watch shaking on the rescuer's wrist. As such, the only axis that provides reliable, consistent data was the X axis.

The raw data from the smartwatch is very noisy and therefore a mean smoothing filter was applied. This smoothing filter was experimentally chosen to cover a range of 200ms centred around the observation being considered. This removed the vast majority of spurious peaks (which were relatively high frequency) and troughs, without damaging the compression data (which was around 2Hz).

Compressions result in sharp peaks and troughs in the X axis acceleration data. We designate the trough as the compression although any point on this regular cycle could be chosen. The method used to locate the peaks and troughs involved taking each data point and determining if it was the lowest or highest point within the previous 125ms and the next 125ms. This range allows for smaller local minima and maxima to be ignored. Examples are shown in Figure 2.

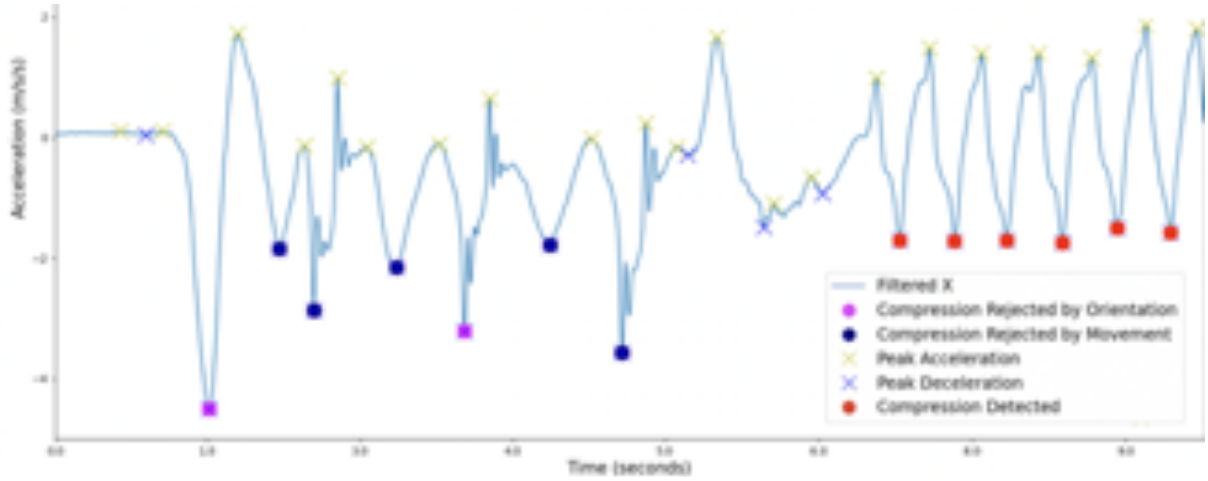


Figure 2: Filtered acceleration data from the X axis of the watch from the clapping at the start of the recording (left) followed by 6 compressions (right). The peak deceleration points are all possible compression locations but the invalid have been rejected due to movement on other axes, incorrect orientation of the watch or by an insufficient change between the highest and lowest points.

Having found the peaks and troughs we label each valley as a compression if:

1. The trough follows a peak and is followed by another peak where the absolute differences in acceleration between both peaks and the trough are at least 1 m/s^2.

2. The force in the X-axis is at least 70% of the total magnitude of the force of gravity which allows for a maximum deviation of approximately 45 degrees from vertical. This ensures that the arm of the person is more-or-less vertical.
3. At least 30% of the movement was in the X direction (i.e. there was limited movement orthogonal to the movement of the arm).

The generous allowance of deviation from vertical tolerates cases where the watch may be pressing on the rescuer's wrist or the rescuer's arms may not be positioned completely vertically, both of which result in the orientation of the watch deviating slightly from vertical.

4 Results

Five separate recordings of a person administering CPR to a dummy were made where each recording comprised both video (from the camera phone) and sensor data (from the smartwatch). Overall in these videos 440 compressions were recorded. The ground truth was created by manually stepping through each frame of the video and marking the frames that were the top or bottom of a compression. Note that the video assessment of compressions was performed using the techniques described by Corkery and Dawson-Howe [Corkery and Dawson-Howe, 2019].

In order to be able to assess the compressions detected from smartwatch sensor data it was necessary to first align the sensor data and the ground truth. This was achieved by requiring the rescuer to clap their hands three times at the start and end of each recording. This clapping was clearly visible in both the recorded video and in the data from the smartwatch, which were then manually aligned using unfiltered X axis acceleration data.

The results of evaluating the output of the two approaches and comparing them to the ground truth can be seen in Table 1. This shows a very high level of overall accuracy for both algorithms although the computer vision based algorithm resulted in more misidentified compressions than the smartwatch based algorithm.

	TP	FP	FN	Precision	Recall	Accuracy
Smartwatch (this paper)	438	0	2	1.00	0.99	0.99
Video [Corkery and Dawson-Howe, 2019]	440	12	0	0.97	1.00	0.97

Table 1: This table shows the compressions correctly detected (TP), missed (FN), and incorrectly detected (FP) by both algorithms summarised across all 5 test videos. Precision = TP/(TP+FP), Recall = TP/(TP+FN), Accuracy = TP/(TP+FP+FN)

The detected compressions are used to compute the CCR and the difference in the CCR for the Smartwatch and the Video measurements as compared to the Ground Truth are shown in Figure 3.

5 Conclusions

Overall analysis of the smartwatch data provided CCR data which was as good as, if not better than, that obtained from video analysis using smartphone video. It appears that a smartwatch can be used to detect CCR and CCD to high precision, while a smartphone can detect CCR to high precision. It is unclear if a smartphone can be used to reliably measure the CCD as there are issues converting pixel measurements to physical distances.

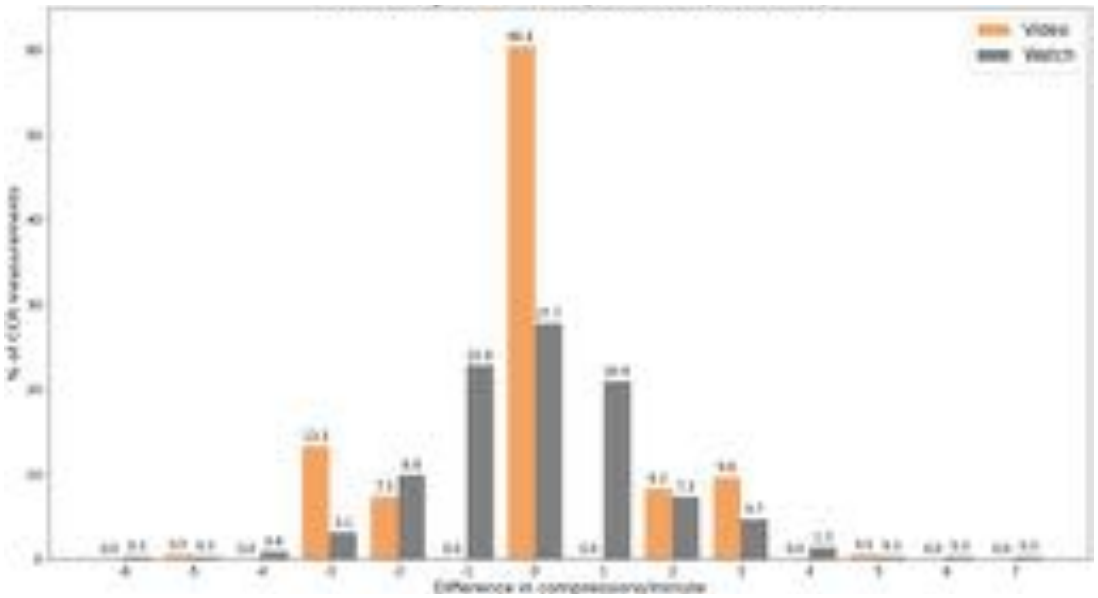


Figure 3: The difference in frames between the calculated CCR for both algorithms and the ground truth as a fraction of all calculated CCRs. CCR calculated using 4 compressions.

References

[Ahn et al., 2017] Ahn, C., + 7 othersLee, J., Oh, J., Song, Y., Chee, Y., Lim, T. H., Kang, H., and Shin, H. (2017). *Effectiveness of feed- back with a smartwatch for high-quality chest com- pressions during adult cardiac arrest: A randomized controlled simulation study*. PloS one, 12:e0169046.

[Baldi et al., 2017] Baldi E. et al. (12 authors) (2019) *Real-time visual feedback during training improves laypersons' CPR quality: a randomized controlled manikin study*. In: CJEM 19.6 (2017), pp. 480–487.

[Corkery and Dawson-Howe, 2019] Corkery G. and Dawson-Howe K. (2019) *A Smartphone Tool for Evaluating Cardiopulmonary Resuscitation (CPR) Delivery*. In: VISIGRAPP (4: VISAPP). 2019, pp. 489–496.

[Gräsner, et al., 2020] Gräsner, J.T. et al. (37 authors) (2020). *Survival after out-of-hospital cardiac arrest in Europe - Results of the EuReCa TWO study*. In: Resuscitation 148, 218–226.

[Herlitz et al., 2005] Herlitz J., Svensson L., Holmberg S., Änquist K.A., and Young, M., (2005) *Efficacy of bystander CPR: intervention by lay people and by health care professionals*. Resuscitation 66: 291-295.

[Meinich-Bache et al., 2017] Meinich-Bache, Ø., Engan, K., Eftestøl, T., and Austvoll, I. (2017). *Detecting chest compression depth using a smartphone camera and motion segmentation*. In Lecture Notes in Computer Science, volume 10270, pages 53–64.

[Song et al., 2015] Song Y., Oh J.,, and Chee Y (2015). *A new chest compression depth feedback algorithm for high-quality CPR based on smartphone*. In: Telemedicine and e-Health 21.1 pp. 36–41.

Semi-supervised Learning of Cardiac MRI using Image Registration

Carles Garcia-Cabrera[1], Kathleen M. Curran[2], Noel E. O'Connor[1], and Kevin McGuinness[1]

[1]*Dublin City University*

[2]*University College Dublin*

Abstract

In this work, we propose a method to aid the 2-D segmentation of short-axis cardiac MRI. In particular, the deformation fields obtained during the registration are used to propagate the labels to all time frames, resulting in a weakly supervised segmentation approach that benefits from the features in unlabelled volumes along with the annotated data. Experimental results over the M&Ms datasets show that the addition of the synthetically obtained labels to the original dataset yields promising results in the performance and improves the capability of the network to generalise to scanners from different vendors.

Keywords: Cardiac MRI, Image Segmentation, Semi-Supervised Learning, Image Registration, Medical Imaging.

1 Introduction

Cardiac image segmentation is an important first step for many approaches to quantitative analysis for cardiac diagnostic assessment. This process requires partitioning the image into a number of clinically meaningful regions such as left ventricle (LV), right ventricle (RV), or myocardium (MYO). Acquiring this information allows clinicians understand important features such as the ejection fraction and the volume that the heart is managing at different times. Those features are later used to determine if there is any possible pathology and how bad it is [Kawel-Boehm et al., 2015].

Data is a key challenge when trying to use off-the-shelf algorithms in this area, specifically the limited amount of annotated data available, but also its quality. Many researchers report struggling to achieve improved results with existing annotated data, especially when working with open datasets [Chen et al., 2020]. Difficulties range from low availability of data, to domain shift using data from a certain scanner vendors, to images from patients with rare conditions.

There is growing interest in the community in understanding how to transfer models that work well for specific scanners to unseen ones. In this work we address these challenges using the M&Ms challenge dataset, which we believe to be the representative dataset for this issue [Campello et al., 2020].

Convolutional neural networks (CNN) are the most common type of deep neural networks for image analysis and have advanced the state-of-the-art in many object segmentation tasks, including in the medical imaging domain. In particular, U-Net [Ronneberger et al., 2015] is the architecture with the best results over most of the challenges that the cardiac MRI currently faces [Chen et al., 2020]. Hence, we have selected it to run our experiments.

Recently, [Zhang et al., 2021] proposed a method where labels can be propagated using image registration in an unsupervised manner and those labels are used to enhance the process where a 3D U-Net learns features. Even though results ranked second in the M&Ms challenge [Campello et al., 2020], the study did not test those propositions in 2D networks, a key novelty of our work, nor explored different tools for the registration part.

Contributions: this work proposes a method to enhance the segmentation of short-axis cardiac MRI by synthetically labelling volumes without annotations. The proposed system works by registering the labelled volumes of the end-systolic (ES) and the end-diastolic (ED) time-frames and using the warping field over their ground truth. This helps the network to learn robust features from those volumes, which are important towards increasing the performance over different scanners. Our study extends recent works where the propagation of labels is used in a 3D U-Net and using a different

registration tool. In contrast to that work, we investigate whether this technique is valid for 2D U-Net applications that can be used in less computationally powerful machines and expect to have better performances in anisotropic datasets.

2 Related Work

The clinical interest in ventricle segmentation has pushed the community towards improving the performance for this task. With the advent of deep learning, a succession of different approaches were investigated. One of the first, was the usage of a fully convolutional network [Tran, 2016]. From there, many works have improved the networks, increasing the learning capacity for segmentation [Khened et al., 2019]. In that regard, the low through-plane resolution and the motion artifacts limited the applicability of 3D networks [Baumgartner et al., 2017].

Annotated data is the foundation of fully supervised approaches but at the same time it is scarce and costly to obtain. Moreover, annotating medical images requires significant expertise and manual effort and even then can lead to noisy labels and/or biases. Additionally, there is an imbalance between the amount of available data regarding different source scanner, target pathology, or scanning parameters. All the above-mentioned challenges point strongly towards the need to design methods where the accuracy of the networks is maintained compared to fully supervised approach, even when there is significantly less data available.

A number of works in the literature proposed unsupervised or semi-supervised techniques to overcome the scarcity of labels. One approach was to use a scribble annotation, that consists of a set of quickly drawn labels, and recursively re-training the network using the output segmentation, including a conditional random field and an uncertainty estimator [Can et al., 2018]. The advances in the calculation of optical flow inspired the usage of these algorithms for motion estimation and this has been useful in multi-task approaches where a Siamese network performs the motion estimation and the segmentation simultaneously, exploiting the information contained in unlabelled data [Qin et al., 2018].

Voxelmorph [Balakrishnan et al., 2018] is an image registration tool for alignment and registration that can also model deformations. It has been used for atlas based registration [Dalca et al., 2019] and for probabilistic diffeomorphic registration [Dalca et al., 2018]. Given its proven reliability, we use it for the registration part of our study.

3 Method and Experiment

This section describes the proposed method, detailing the different steps and implementations of all the tools involved in the process. Distinguishing two first steps corresponding to the registration and label propagation, and two last parts corresponding to training and testing the two different models i.e. the original dataset and the one with the addition of the synthetic labels.

3.1 Data

For our experiments we used the M&Ms dataset, released during the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Image Segmentation Challenge. This dataset consists of 345 patients with hypertrophic and dilated cardiomyopathies and healthy subjects and it also permits the opportunity to measure performance over different scanners, and to evaluate the resilience of our approach towards different data sources. In particular the training set includes 150 cases (75 from vendor A and 75 from vendor B). The test set includes cases from two more vendors (C and D).

The CMR images have been segmented by experienced clinicians from the respective host institutions, including contours for the LV and RV blood pools, as well as for the left ventricular MYO.

3.2 Label Propagation

Our selected image registration technique was Voxelmorph [Balakrishnan et al., 2019], and given that the pre-trained models were trained with images from other tissues, we decided to train our own with our available data. We also

included the ground truth segmentation labels in the training process to produce a model that later we used for label propagation. For this part of the process we used the standard configuration of Voxelmorph (Tensorflow version)[1].

After training the model we computed all the warping fields for the time-frame between ED and ES, always in an intra-subject way. With the warping fields we modified the ground truth for the ED time-frame, obtaining the synthetic labels for out target time-frames.

Lastly, we join in a single dataset all the volumes with annotations, including the the original and the synthetic.

3.3 Segmentation

In our approach, we used a 2D U-Net to segment the end-diastolic and end-systolic volumes in the target dataset. In particular, our U-Net model used 32 feature channels with kernel size equal to three in the first level of the convolutional layers, where batch normalization was also applied and the activation function was ReLU, followed by a 2×2 max-pooling layer. The final 1×1 convolution is set with four channels that match the four target regions: background, right ventricle, left ventricle, and myocardium.

To train the model, we used an Adam optimizer (learning rate = 10^{-3}) with a plateau learning rate scheduler, and as a loss function we calculated the cross-entropy loss. In addition to this, we computed the Dice score for all the regions in each epoch for the validation set.

4 Results

Table 1 shows the results obtained for both datasets, including regions and vendors (M&Ms). The proposed method performs 2.6% better than the baseline, where the region that benefits the most from this change is the corresponding to the myocardium. When we compare the difference in the performance of both models in the different scanners, we see how the performance of unseen scanners C and D improves in the model resulting from the proposed method while it sees its performance almost untouched for the scanners present in the training set (A and B).

	Regions			**Vendors**				**Total**
	LV	**Myo**	**RV**	**A**	**B**	**C**	**D**	**Dice**
Baseline (M&Ms)	0.475	0.386	0.394	0.553	0.583	0.419	0.277	0.418
Proposed (M&Ms)	0.504	0.430	0.398	0.548	0.573	0.443	0.338	0.444

Table 1: Results (dice scores) on the M&Ms datasets. Higher is better.

5 Conclusions

Experimental results using synthetic labels generated from VoxelMorph showed an improvement over the baseline, demonstrating the potential for using diffeomorphic image registration as a label propagation technique. Future work will investigate using this technique with a stronger baseline. In particular, we plan to adopt nnUNet [Isensee et al., 2021], an automatic segmentation framework for medical images based on the U-net architectures, which has been demonstrated to give state-of-the-art results. Finally, we believe the results can be improved by leveraging research on unreliable or noisy labels (e.g. [Arazo et al., 2019]), weighting them differently to reflect the level of trust in the annotations, or post-processing the synthetic annotations. Future work will investigate this.

Acknowledgments

This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

[1] VoxelMorph `https://github.com/voxelmorph/voxelmorph`.

References

[Arazo et al., 2019] Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. (2019). Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321. PMLR.

[Balakrishnan et al., 2018] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Dalca, A. V., and Guttag, J. (2018). An unsupervised learning model for deformable medical image registration. *Computer Vision and Pattern Recognition (CVPR)*.

[Balakrishnan et al., 2019] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800.

[Baumgartner et al., 2017] Baumgartner, C. F., Koch, L. M., Pollefeys, M., and Konukoglu, E. (2017). An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation.

[Campello et al., 2020] Campello, V. M., Palomares, J. F. R., Guala, A., Marakas, M., Friedrich, M., and Lekadir, K. (2020). Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge.

[Can et al., 2018] Can, Y. B., Chaitanya, K., Mustafa, B., Koch, L. M., Konukoglu, E., and Baumgartner, C. F. (2018). Learning to segment medical images with scribble-supervision alone.

[Chen et al., 2020] Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., and Rueckert, D. (2020). Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7.

[Dalca et al., 2018] Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R. (2018). Unsupervised learning for fast probabilistic diffeomorphic registration. *Lecture Notes in Computer Science*, page 729–738.

[Dalca et al., 2019] Dalca, A. V., Rakic, M., Guttag, J., and Sabuncu, M. R. (2019). Learning conditional deformable templates with convolutional networks.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211.

[Kawel-Boehm et al., 2015] Kawel-Boehm, N., Maceira, A., Valsangiacomo-Buechel, E. R., Vogel-Claussen, J., Turkbey, E. B., Williams, R., Plein, S., Tee, M., Eng, J., and Bluemke, D. A. (2015). Normal values for cardiovascular magnetic resonance in adults and children. *Journal of Cardiovascular Magnetic Resonance*, 17(1):29.

[Khened et al., 2019] Khened, M., Kollerathu, V. A., and Krishnamurthi, G. (2019). Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*, 51:21 – 45.

[Qin et al., 2018] Qin, C., Bai, W., Schlemper, J., Petersen, S. E., Piechnik, S. K., Neubauer, S., and Rueckert, D. (2018). Joint learning of motion estimation and segmentation for cardiac mr image sequences.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation.

[Tran, 2016] Tran, P. V. (2016). A fully convolutional neural network for cardiac segmentation in short-axis MRI.

[Zhang et al., 2021] Zhang, Y., Yang, J., Hou, F., Liu, Y., Wang, Y., Tian, J., Zhong, C., Zhang, Y., and He, Z. (2021). Semi-supervised cardiac image segmentation via label propagation and style transfer. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 219–227. Springer.

Strictly Ballroom - Analysing Dance Skills with Temporal Segment Networks

He Liu, Gerard Lacey

School of Computer Science and Statistics, Trinity College Dublin, Ireland

Abstract

Ballroom Dancing contains fast complex motions that require significant training and are difficult to analyse and score. Automating the analysis of complex physical skills is useful for sports coaching, medical skills training and rehabilitation. Our paper explores the use of Siamese Temporal Segment Network (TSN) and salient region identification using BlazePose to rank ballroom dancers automatically. The system produces a Spearman Rank correlation with dancer experience of 96%. Our experiments indicate that salient snippet identification using Blazepose had a modest impact on the results. The limitations of our system are the small dataset and the need to manually tune motion segmentation thresholds based on the NP Smoothness metric.

Keywords: Video Action Recognition, Physical Skill Ranking, Temporal Segment Networks

1 Introduction

This paper describes a real-time whole-body physical skills analysis evaluation system using video. The analysis of complex human motion patterns is difficult and has applications in skills training, entertainment and rehabilitation. Wearable motion sensors are often used to capture physical movement and vary from specialised wearable to mobile phone-based sensors. Camera-based motion-sensing varies from a professional grade for movies and games requiring specialised suits, to games console and mobile phone cameras that use motion as an input to applications. These applications range from motion-triggered games such as Just Dance to fitness and sports technique training. In competitive domains such as Dance, Gymnastics and Platform Diving the coordination and control of posture is key. As can be seen in Figure 1, the difference in the body pose is easy to detect. When the body is in motion these differences can be less obvious.

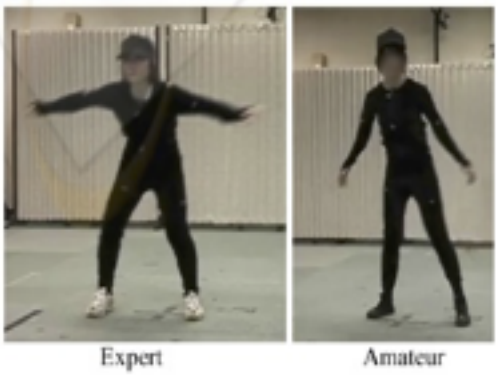


Figure 1: Expert vs Amateur dancer from [Tada & Naemura, 2006]

2 State of the Art

We examine issues related to the analysis of dance skills, Deep Learning approaches to motion analysis specifically the Siamese network and its extension Temporal Siamese Networks (TSN) for action recognition. Finally, we look at pairwise ranking to allow a comparison of skill levels between the action videos.

2.1 Complex motion analysis

Many researchers such as [Tada & Naemura, 2006] analyzed dance using the auto-correlation and mutual-correlation functions with the beat-to-beat interval of music, but they ignored the postural aspects of the dance. Wearable optical motion capture devices have been used[Chan et al., 2010] and were adequate to compare captured motions with templates in the database. The relationship between music rhythm and dance movements was also examined. In[Kishore et al., 2018] a convolutional neural network (CNN) to extract and recognize dance movements from video. A 3D neural network combined with support vector regression [Parmar & Tran Morris, 2017] was used to training to score Olympic Diving, Vault and figure skating. In [Nekoui et al., 2020] pose tracking was combined with an appearance-based approach to score Olympic diving.

2.2 Siamese networks for action recognition

One challenge of complex action recognition is that very large data sets that show the ideal actions are rarely available. The Siamese network, first proposed in [Bromley et al., 1993], allows us to compare if two images are similar and give is the distance between them. The network composed of two identical sub-networks with shared weights and outputs are compared in a loss layer. The original Siamese network was applied to fingerprint and face recognition.

The Siamese network can solve ranking problems as it takes two data sources and generates distance measures between the samples. For a video ranking problem, the Siamese network learns to rank videos from experts and novices and uses a video from the user as evaluation data. The Siamese network can output the semantic similarity which can be used as a proxy for skill level.

Temporal Segment Networks (TSN)[Wang et al., 2016] provided an improved framework for human action recognition based on Two-Stream Convolutional Neural Networks (2S-CNN). It uses two convolutional networks, one each for temporal and spatial video features. The outputs are then fused to generate the score for the action. The TSN sparsely samples the video by first splitting the video into K segments and then randomly sampling snippets of video from each segment. A consensus is achieved across the K spatial and temporal segments prior to being combined to determine the score. To further improve the results a data augmentation strategy of corner cropping and scale jittering was used.

2.3 Pairwise Deep Ranking

Pairwise Ranking is a statistical tool to give priorities to multiple options. [Doughty et al., 2018] combined this concept with deep learning architecture as Pairwise Deep Ranking to rank skills in video recording. For any two videos doing the same task, they aim to determine which one is better, and across a set of videos which one is the best. The deep representation of skills from the videos are first obtained with TSN, then the skill levels differences are labelled for any video pairs. Finally, the weights of TSN are updated with the labelled video pairs data through Siamese Network. They proposed a novel ranking loss function that considers both pairwise ranking and pairwise similarities and achieved from 70.2% to 83.2% precision on tasks as diverse as surgery, dough-rolling, drawing and chopstick-use.

3 Methodology

Our system uses the TSN architecture followed by a pairwise ranking from [Doughty et al., 2018]. In that work pairs of videos were divided into three segments of equal length. The videos had a constant rate of task progression but in ballroom dancing the dancer sometimes moves rapidly and at other times is relatively still. Our aim was to investigate if splitting the videos based on these motion patterns would have an impact on the results.

Figure 2: Tracking Result

3.1 Motion Analysis using Blazepose

For our study we used the Ballroom Dance Dataset (BDD) [Matsuyama et al., 2019] consists of 7 dancers, each dancer performs the same rumba dance steps 20 times. We use the dancer's years of experience as a proxy for relative skill level. The distribution of years of experience is: 17, 5, 5, 4, 3, 1, 1.

There are no annotations for the dance steps within BDD so we used Blazepose [Bazarevsky et al., 2020] to track dancer's limbs and the NP Smoothness metric [Balasubramanian et al., 2015] to split videos when there was a change of direction by the dancer. The NP smoothness was calculated by tracking the forearms and shins of the dancer. When NP smoothness fell below a threshold the video segments were marked with "break" as in Figure 2. This approach over-segmented the video into separate "dance movements". The dance movement segments were then manually merged to create the "dance steps" of the rumba. We evaluated the impact of splitting videos using both the fully automatic dance movement method and manual dance step method.

3.2 The TSN and pairwise ranking architecture

The TSN consisted of two independent networks, a 2S-CNN Spatial Network and a 2S-CNN Temporal Network as shown in Figure 3. Each video is divided into N segments based on the 3 strategies: Equal length, Dance Steps and Dance Movements. From each segment three randomly selected RGB and optical flow images are fed to the Spatial and Temporal networks respectively. The optical flow images are created from a sequence of 5 images and produce a dense estimate of optical flow.

The three input images to Spatial and Temporal networks produce three estimates for the skill score for each segment. These are combined to produce a consensus score represented by Score S(patial) and Score T(emporal). These scores are then combined using a weighted sum: Fusion = a*spatial+(1-a)*temporal. Finally, the skill score for the entire video is determined by summing across all the segments.

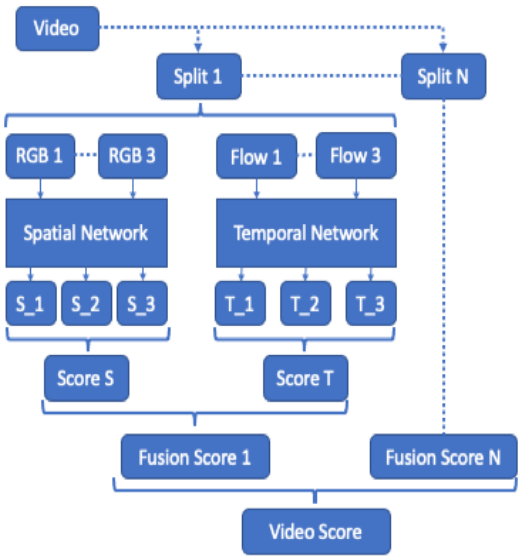


Figure 3: Video Split for Evaluation

4 Results

Ballroom Dance Dataset (BDD) consists of 7 dancers, each dancer performs the same rumba dance steps 20 times. The model was trained with the batch size 128, momentum 0.9 and dropout 0.5. The data was split 80% training and 20% test. The training was stopped after 980 iterations. The system was tested with Intel i7 4790k and NVIDIA GTX 1060. The average calculation time for one video split is 6.4168s.

We calculated Spearman's rank correlation on the test data to measures the strength and direction of association between ranked videos. The greater the value the higher the correlation between the prediction and ground truth. The results are presented in Table 1

5 Conclusions

The segmentation of the BDD videos into discrete dance steps produced better results when we consider the spatial only model or the temporal only model. However, the results of equal segmentation using a fusion model with 60% Spatial and 40% Temporal produce the best results. The results obtained are relatively high and this may be due to the lack of diversity in the data and a relatively small dataset. In addition we have used

Table 1: Correlation results for the different video split methods

	Spearman Rank Correlation		
Model Score Method	Equal Splits	Dance Steps	Dance Movements
Spatial Only	0.943	0.954	0.942
Temporal Only	0.810	0.834	0.805
Fusion: 60% spatial + 40% temporal	0.968	0.952	0.947

years of experience as a proxy for skill level and there was a good spread of experience level leading to clear separation between most skill levels.

References

[Balasubramanian et al., 2015] Balasubramanian, S., Melendez-Calderon, A., Roby-Brami, A., & Burdet, E. (2015). On the analysis of movement smoothness. *Journal of neuroengineering and rehabilitation*, 12(1), 1–11.

[Bazarevsky et al., 2020] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

[Bromley et al., 1993] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., & Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04), 669–688.

[Chan et al., 2010] Chan, J. C., Leung, H., Tang, J. K., & Komura, T. (2010). A virtual reality dance training system using motion capture technology. *IEEE transactions on learning technologies*, 4(2), 187–195.

[Doughty et al., 2018] Doughty, H., Damen, D., & Mayol-Cuevas, W. (2018). Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6057–6066).

[Kishore et al., 2018] Kishore, P., Kumar, K., Kiran Kumar, E., Sastry, A., Teja Kiran, M., Anil Kumar, D., & Prasad, M. (2018). Indian classical dance action identification and classification with convolutional neural networks. *Advances in Multimedia*, 2018.

[Matsuyama et al., 2019] Matsuyama, H., Hiroi, K., Kaji, K., Yonezawa, T., & Kawaguchi, N. (2019). Ballroom dance step type recognition by random forest using video and wearable sensor. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (pp. 774–780).

[Nekoui et al., 2020] Nekoui, M., Cruz, F. O. T., & Cheng, L. (2020). Falcons: Fast learner-grader for contorted poses in sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[Parmar & Tran Morris, 2017] Parmar, P. & Tran Morris, B. (2017). Learning to score olympic events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 20–28).

[Tada & Naemura, 2006] Tada, M. & Naemura, M. (2006). Dance evaluation system based on motion analysis. In *GRAPP*.

[Wang et al., 2016] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2016). Temporal segment networks: Towards good practices for deep action recognition.

An Experimental Comparison of Knowledge Transfer Algorithms in Deep Neural Networks

Seán Quinn, Kevin McGuinness, Alessandra Mileo

Insight Centre for Data Analytics
Dublin City University

Abstract

Neural knowledge transfer methods aim to constrain the hidden representation of one neural network to be similar, or have similar properties, to another by applying specially designed loss functions between the two networks hidden layers. In this way the intangible knowledge encoded by the network's weights is transferred without having to replicate exact weight structures or alter the knowledge representation from its natural highly distributed form. Motivated by the need to enable greater transparency in evaluating such methods by bridging the gap between different experimental setups in the existing literature, the need to cast a wider net in comparing each method to a greater number of its peers and a desire to explore novel combinations of existing methods we conduct an experimental comparison of eight contemporary neural knowledge transfer algorithms and further explore the performance of some combinations. We conduct our experiments on an image classification task and measure relative performance gains over non-knowledge enhanced baseline neural networks in terms of classification accuracy. We observed (i) some interesting contradictions between our results and those reported in original papers, (ii) a general lack of correlation between any given methods standalone performance vs performance when used in combination with knowledge distillation, (iii) a general trend of older simpler methods outperforming newer ones and (iv) Contrastive Representation Distillation (CRD) achieving best performance.

Keywords: Knowledge Transfer, Knowledge Distillation, Deep Learning, Representation Learning, Machine Learning

1 Introduction

Modern deep neural networks solve difficult tasks by learning incredibly complex decision functions from raw training data and appropriate reward signals such as labels. The fields of artificial intelligence, machine learning and computer vision have benefited hugely from rapid advances in the theory of neural learning in recent years [LeCun et al., 2015]. However, as neural approaches begin to achieve a level of maturity and widespread adoption, attention has increasingly shifted from focusing on how best to exploit their well-established strengths to identifying and addressing their inherent and more fundamental limitations [Lake et al., 2017]. Parallels have been drawn between the somewhat opposing strengths and weaknesses of deep neural networks and "old-school" symbolic AI systems [Sun, 1999] – a key attribute of which is the re-use of sophisticated bodies of knowledge, albeit knowledge represented in a vastly different format. The knowledge learned by a deep neural network is encoded across a network of thousands to millions of neurons arranged in sequential layers, often hundreds deep. Individual neurons and their learned weights, the concrete entities which underpin such knowledge, are uninterpretable and of no utility when considered in isolation of the wider network structure. Therefore the knowledge contained within a neural network is an abstract function encoded in a distributed format across the entirety of the network's weights. In this sense the knowledge we seek to reuse is intangible and the existing knowledge transfer and reuse methodologies of the symbolic AI domain are not readily applicable in the deep learning domain. However, the lessons learned from knowledge representation in symbolic

AI along with the envisioned benefits it could bring to neural learning [Lake et al., 2017], has meant that the question remains – how do we emulate these knowledge transfer and reuse capabilities in the vastly different neural learning paradigm? Distillation based neural knowledge transfer approaches seek to do this by augmenting a neural networks learning process with inputs derived from another neural network (a teacher) where it is believed the teachers weights contain knowledge beneficial to the student network on a chosen task. This involves constraining the student networks' hidden representation to have similar properties to the teachers, thereby transferring knowledge from teacher to student in an abstract manner. This is achieved by applying specially designed loss functions between transformed representations of activations collected from hidden layers across both teacher and student networks for the same sample of data. By encouraging the student to represent individual pieces of data in a similar way to the teacher, the student's weights are also encouraged to encode a function similar (or with similar properties) to the teachers while appreciating that the student cannot replicate exact weight structures. In this paper we will contrast the performance of eight such contemporary neural knowledge transfer methods on a computer vision task. Comparisons such as this one are necessary to bridge the gap between the sometimes vastly different experimental setups that appear alongside original publications and to cast a wider net in comparing each method to a greater number of its contemporary peers. The comparison is intended to enable greater transparency and more reliable conclusions in establishing the state-of-the-art. Further to this, as an exploratory component, we will examine the combination of the Knowledge Distillation (KD) method [Hinton et al., 2015] with each of the other seven methods, some of these being previously unexplored combinations.

2 Experimental Comparison

2.1 Models & Data

We use ResNet models [He et al., 2016] in our experimental setup as they are one of the few widely known models that are viable for use with all of the methods examined here. Our teacher, the ResNet-56, is 2.7 times deeper and has 3.1 times as many parameters as the ResNet-20 we use as student; thus it has a much greater representational capacity and the ability to encode a much more powerful decision function. In selecting the parameters for methods reproduced here, we endeavoured to stay as close as possible to those reported in the original papers. However, many of the original papers do not report essential training parameters. In these cases we make choices based on (i) achieving functional gradient descent (ii) achieving parameter consistency across methods (where possible) and (iii) maximising classification accuracy on the validation dataset. We use the CIFAR-10 and CIFAR-100 image classification datasets [Krizhevsky, 2009] in this experimental evaluation. All details necessary for reproduction of these experiments including data loading parameters and exact training parameters for all models trained are available on the GitHub page at `github.com/squinn95/KD_IMVIP_21`.

2.2 Discussion of Results

The results of our experimental comparison are shown in Table 1. We also report the performance of a non-knowledge enhanced baseline student network as a benchmark with which to measure relative performance gains. We observe that **KD** [Hinton et al., 2015] performs strongly in the evaluation, showing best performance on CIFAR-10 and second best on CIFAR-100, this is broadly in line with the strong results reported for KD in the papers which reproduce it [Passalis and Tefas, 2018, Huang and Wang, 2017, Tian et al., 2020]. In our paper it outperforms all methods except CRD. The original **FitNet** evaluation does not report a baseline student figure [Romero et al., 2014], so we cannot contrast relative performance gains, it does however report that the method outperforms its teacher despite the student being a much lower capacity model. We did not observe performance this strong in our evaluation for FitNet or for any of the other methods or combination methods. Despite this we observed stronger gains over the baseline for FitNet than expected, as it is reproduced in several other papers [Yim et al., 2017, Passalis and Tefas, 2018, Huang and Wang, 2017]. The original **FSP** evaluation [Yim et al., 2017] compares with only one other method, FitNet, which it is shown to outperform.

Table 1: Top-1 classification accuracy (%) on the CIFAR datasets [Krizhevsky, 2009] – knowledge transfer from ResNet56 to ResNet20 [He et al., 2016]. Relative improvement over baseline student shown in brackets.

Model	**CIFAR-10**	**CIFAR-100**
Teacher	93.68	72.46
Baseline Student	91.74	68.3
KD [Hinton et al., 2015]	**92.66** (+0.92)	69.78 (+1.48)
FitNet [Romero et al., 2014]	92.22 (+0.48)	69.26 (+0.96)
FSP [Yim et al., 2017]	92.18 (+0.44)	68.76 (+0.46)
PKT [Passalis and Tefas, 2018]	91.88 (+0.14)	69.26 (+0.96)
MMD-Linear [Huang and Wang, 2017]	91.86 (+0.12)	68.9 (+0.6)
MMD-Polynomial [Huang and Wang, 2017]	92.06 (+0.32)	68.46 (+0.16)
MMD-Gaussian [Huang and Wang, 2017]	92.62 (+0.88)	69.38 (+1.08)
CRD [Tian et al., 2020]	91.88 (+0.14)	**70.66** (+2.36)
FitNet+KD	92.36 (+0.62)	69.98 (+1.68)
FSP+KD	92.34 (+0.6)	70.42 (+2.12)
PKT+KD	92.56 (+0.82)	69.88 (+1.58)
MMD-Linear+KD	92.04 (+0.3)	69.32 (+1.02)
MMD-Polynomial+KD	92.52 (+0.78)	69.9 (+1.6)
MMD-Gaussian+KD	92.76 (+1.02)	69.08 (+0.78)
CRD+KD	**92.96** (+1.22)	**70.9** (+2.6)

We observed the opposite result, with FitNet achieving marginally higher performance on CIFAR-10 and significantly higher performance on CIFAR-100. We examined the previously unexplored combination of FSP+KD which achieved surprisingly good performance, second best overall on CIFAR-100, it is noteworthy that FSP seems to offer much more when combined with KD than as a standalone method. The original **PKT** evaluation [Passalis and Tefas, 2018] reports the methods outperforming FitNet and KD in a content-based retrieval experimental setup, we observe the opposite in our classification setup, with both methods outperforming PKT. The original **MMD** paper [Huang and Wang, 2017] reports MMD-Polynomial as the strongest performing of the three MMD kernel variants where in our experiments we observe MMD-Gaussian to be the stronger. Due to this finding the authors did not include results for the combinations MMD-Linear+KD and MMD-Gaussian+KD. While we did find MMD-Polynomial+KD to be the stronger combination on CIFAR-100, MMD-Gaussian+KD prevailed on CIFAR-10, making the extra combinations a worthwhile inclusion in our analysis. They further report MMD-Polynomial as surpassing both FitNet and KD on CIFAR-10 and FitNet on CIFAR-100 where we observe MMD-Gaussian surpassing FitNet only on both datasets and do not observe either of the other two kernel variants outperforming KD or FitNet. We see near identical performance for the **CRD** method on CIFAR-100 in our evaluation and the original paper [Tian et al., 2020], confirming it as the strongest performing method among those examined. Curiously it did not perform as strongly when used without KD on CIFAR-10, but held a clear advantage over other methods in the other three scenarios examined.

3 Conclusion

We find all methods examined improved performance over a non-knowledge enhanced baseline on both the CIFAR-10 and CIFAR-100 datasets. We achieved results roughly in line with expectations for three of the

methods examined (KD, MMD-Linear, CRD), weaker performance in three methods (MMD-Polynomial, PKT, FSP) and slightly stronger than expected performance in two methods (MMD-Gaussian and FitNet). We suspect that these some of these disparities may be due to a tendency for evaluation setups to be slightly biased towards the method they report. We found little correlation between the performance of methods when evaluated standalone vs when combined with KD – one does not offer any reliable insight into the other. This is illustrated by the fact that two of the three new knowledge transfer scenarios explored here (FSP+KD, MMD-Gaussian+KD) achieved surprisingly strong results. From this we conclude that there is value in exploring method combinations even when their constituent parts do not appear especially promising. We observed a general trend of older simpler methods (KD, FitNet) outperforming some more recent methods (FSP, PKT, MMD) in contradiction to the initial evaluations which accompanied these newer methods. Finally, we confirm CRD to be the best performing standalone method and CRD+KD to be the overall strongest knowledge transfer regime.

Acknowledgments

Funded by the Irish Research Council GOIPG/2018/2501 and partially by Science Foundation Ireland SFI/12/RC/2289_P2. Supported by an Nvidia Corporation research hardware grant.

References

[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.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[Huang and Wang, 2017] Huang, Z. and Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

[Krizhevsky, 2009] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

[Lake et al., 2017] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[Passalis and Tefas, 2018] Passalis, N. and Tefas, A. (2018). Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284.

[Romero et al., 2014] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

[Sun, 1999] Sun, R. (1999). Artificial intelligence: Connectionist and symbolic approaches.

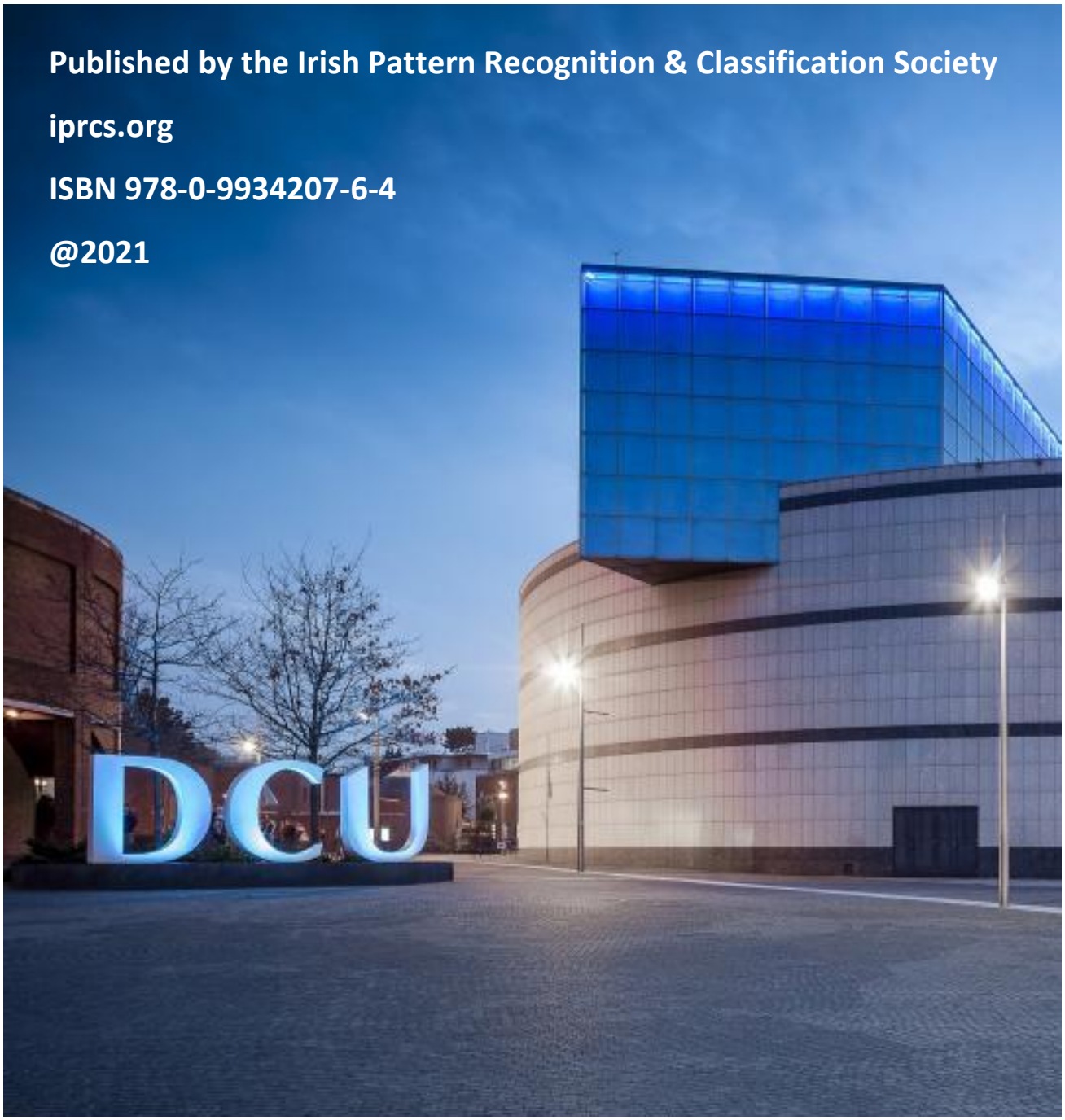
[Tian et al., 2020] Tian, Y., Krishnan, D., and Isola, P. (2020). Contrastive representation distillation. In *International Conference on Learning Representations*.

[Yim et al., 2017] Yim, J., Joo, D., Bae, J., and Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141.

Published by the Irish Pattern Recognition & Classification Society

iprcs.org

ISBN 978-0-9934207-6-4



Irish Pattern Recognition and Classification Society