Northumbria University
NEWCASTLE

NEURAL ATTENTION MECHANISMS FOR ROBUST AND INTERPRETABLE FEATURE REPRESENTATION LEARNING

DANIEL ORGANISCIAK

PhD

2022

NEURAL ATTENTION MECHANISMS FOR ROBUST AND INTERPRETABLE FEATURE REPRESENTATION LEARNING

DANIEL ORGANISCIAK

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Faculty of Engineering and Environment

September 2022

Abstract

The field of artificial intelligence has exploded since improved hardware capabilities have allowed highly complex algorithms to processed within a reasonable timescale. This has resulted in the burgeoning study of deep neural networks, which has allowed models to learn powerful feature representations of data instances that they are trained to evaluate. However, deep neural networks come with significant ethical dilemmas because they lack transparency. Deep learning algorithms often contain millions of parameters; they are too complex to be interpreted by humans. This creates an important trust issue that must be overcome, particularly on domains where data is sensitive, such as medicine and security.

Deep learning has shown a strong ability to discover an association between input data instances and output labels, even when generalising to unseen data from the same distribution. However, deep learning can often exhibit a lack of robustness when applied to new scenarios, such as generalising to data from a different source from which the model is trained.

The use of attention mechanisms has recently begun to proliferate, particularly within language models for natural language processing. Motivated by the fact that the human visual system also makes use of attention, this thesis explores the application of attention mechanisms on tabular and image data. In particular, the potential of attention mechanisms to jointly solve the interpretability and robustness problems is evaluated, facilitating the uptake of deep learning within the real world. Attention mechanisms can be directly inspected to view a saliency map, to interpret regions that the deep learning model deems to be important. Furthermore, because attention can minimise the influence of spurious features, deep learning models become more robust and can better adjust to tougher scenarios.

Extensive testing demonstrates attention mechanisms maintain these capabilities within a variety of deep models including deep neural networks, convolutional neural networks, and generative adversarial networks. Attention mechanisms are found to improve robustness against data attacks, poor-quality images, out-of-distribution generalisation, and extreme scale variations. Overall, this

thesis demonstrates that attention mechanisms are an essential inclusion to deep learning models for real-world applications.

List of Publications

[1] **Organisciak, D.**, Sakkos, D., Ho, E. S. L., Aslam, N. and Shum, H. P. H., "Unifying person and vehicle te-identification," *IEEE Access*, vol. 8, pp. 115 673–115 684, 2020.

[2] **Organisciak, D.**, Riachy, C., Aslam, N. and Shum, H. P. H.,"Triplet loss with channel attention for person re-identification," *Journal of WSCG*, vol. 27, no. 2, pp. 161–169, 2019.

[3] **Organisciak, D.**, Shum, H. P. H., Nwoye, E., and Woo, W. L., RobIn: A robust, interpretable deep network for schizophrenia diagnosis. *Expert Systems with Applications 201 (2022): 117158.*.

[4] **Organisciak, D.**, Ho, E. S. L., and Shum, H. P. H., "Makeup style transfer on low-quality images with weighted multi-scale attention.," In 2020 *25th International Conference on Pattern Recognition* (ICPR) (pp. 6011-6018). IEEE.

[5] **Organisciak, D.**, Isaac-Medina, B.K., Poyser, M., Hu, S., Breckon, T.P. and Shum, H.P.H, 2021. UAV-ReID: A benchmark on unmanned aerial vehicle re-identification. *Proceedings of the 2022 International Conference on Computer Vision Theory and Applications (VISAPP)*

[6] Isaac-Medina, B.K., Poyser, M., **Organisciak, D.**, Willcocks, C.G., Breckon, T.P. and Shum, H.P., 2021. Unmanned aerial vehicle visual detection and tracking using deep neural networks: A performance benchmark. In 2021 *International Conference of Computer Vision (ICCV), 2nd Anti-UAV Workshop and Challenge*

[7] Long, Y., Tan, Y., **Organisciak, D.**, Yang, L. and Shao, L., 2018. Towards lightweight annotations: Fuzzy interpolative reasoning for zero-shot image classification. In *29th British Machine Vision Conference* (BMVC 2018). Newcastle University.

[8] Riachy, C., Al-Maadeed, N., **Organisciak, D.**, Khelifi, F., and Bouridane, A. "3D Gaussian descriptor for video-based person re-identification," International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), 2019.

[9] Zuo, Z., **Organisciak, D.**, Shum, H. P. H., and Yang, L. "Saliency-informed spatio-temporal vector of locally aggregated descriptors and fisher vector for visual action recognition," *BMVC Workshop on Image Analysis for Human Facial and Activity Recognition*, 2018.

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
AuC	Area under the Curve
CE	Cross-Entropy
CLS	Classification
CNN	Convolutional Neural Network
CUHK	Chinese University of Hong Kong
DINO	Self-distillation with no labels
DMT	Disentangled Makeup Transfer
DNN	Deep Neural Network
DSM	Diagnostic and Statistical Manual of Mental Disorders
DWE	Dynamically Weighted Euclidean
EEG	Electroencephalogram
EHR	Electronic Health Records
fMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
FPA	Face Parsing Algorithm
GAN	Generative Adversarial Network

GPU	Graphics Processing Unit
HACNN	Harmonious Attention Network
i.i.d.	Independent and identically distributed
ICD	International Classification of Diseases
ID	Identification
mAP	Mean Average Precision
ML	Machine Learning
MLFN	Multi-level Factorisation Network
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
OOD	Out-of-distribution
OSNet	Omni-scale Network
PCB	Parts-based Convolutional Baseline
PVUD	Person and Vehicle Unified Data Set
RDoC	Research Domain Criteria
re-ID	Re-identification
ReLU	Rectified Linear Unit
RLL	Ranked List Loss
RobIn	Robust Interpretable Network
ROC	Receiver Operating Characteristic
SANN	Self-Attention Neural Network

SE	Squeeze and Excitation
SENN	Squeeze and Excitation Neural Network
SHAP	Shapley Additive exPlanations
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UAV	Unmanned Aerial Vehicle
UT	Unification Terms
VAE	Variational Auto-Encoder
VeRi	Vehicle Re-identification
VIPeR	Viewpoint Invariant Pedestrian Recognition
ViT	Vision Transformer

Acknowledgements

I did not want to finish writing this thesis because I enjoyed my PhD so much. I want to acknowledge everyone that has helped me develop, everyone that I've had insightful technical discussions with, and everyone I've shared a beer with for making this journey so much fun.

First, I want to thank my family, for offering consistent, unconditional support. Without you, none of this would be possible.

My supervision team have been incredible. I have a deep gratitude to Hubert Shum for teaching me so much, and selflessly taking every opportunity to develop me as an academic, and as a person. Edmond Ho has been so helpful at every stage, I cannot thank you enough for the support you have offered. Shanfeng Hu has been an amazing supervisor; I've thoroughly enjoyed the deep technical discussions we have had, the projects that we've collaborated on, and the Chinese food we've shared! You were an excellent mentor, and then became an even better supervisor. I hope the future consists of many fruitful collaborations.

Yang Long, Bingzhang Hu, Junyan Zhu - I cannot put into words the amount of support that you have offered me. You have served as inspirations and role models, and I am eternally indebted to you.

The PGR society has been an amazing support through the last 3 years and an excellent platform to vent frustrations. The committee, Steven Thirkle, Megan Doherty, Pete Kruithof, Daria Onitiu and Neera Jeyamohan, have been amazing to work with, and to drink with. I'm also deeply thankful to my other friends in Newcastle, in particular Kevin McCay, Shirine Riachy, Dimitrios Sakkos, Baqar Rizvi, Rebecca Oswald, Omer Ogutcen, Zheming Zuo, who have made this process so enjoyable. Lastly, I want to thank my close pre-PhD friends: Jack Roberts, Dan Boddice, and Amy Johnson for keeping me sane (somewhat) throughout this chapter of my life.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the *Faculty Ethics Committee* on *03/02/2019*, Submission Ref: 14194.

I declare that the Word Count of this thesis is 35877 words.

Name: Daniel Organisciak

Signature:

Date: 30 September 2022

Chapter 1

Introduction

Deep learning [121] has shown extremely powerful predictive capabilities on a wide variety of tasks. The deep learning paradigm shift was initiated when Krizhevsky *et al.* [113] comprehensively outperformed the competition on the ImageNet challenge. Since then, deep neural networks (DNNs) have demonstrated their efficacy across different sectors, becoming the state of the art in natural language processing (NLP) [20], computer vision (CV) [19, 273], and graphical data [235]. DNNs have demonstrated their ability in medicine [184, 212]; security applications such as detection [140], tracking [39], and re-identification [269]; cybersecurity [14]; data generation [73, 86, 110]; and finance [185]. However, there has been resistance to embracing DNNs for applications in the real world. DNNs cannot be understood. Any decisions made come without context or reasoning. These decisions are therefore unable to be trusted, especially if they oppose the opinion of human experts. Furthermore, in academia, many DNNs are trained on clean, high-quality data. These models often struggle with noisy, unstructured data that is more commonly found in the real world. To enable society to fully embrace the potential of DNNs, and to reap the positive benefits that DNNs can offer, the issues of interpretability and robustness must be overcome. This thesis explores the role that attention mechanisms [234, 89] can play to solve these problems, thereby enabling the widespread adoption of DNNs.

1.1 Problem

Despite the clear performance advantages of DNNs [113], there are a number of drawbacks that limit the ability of DNNs to be applied in real-world scenarios. There are human issues to consider when evaluating whether a DNN can be applied to make a decision. For examples, the trustworthiness of the model and the moral implications of using AI to make life or death decisions. DNNs also have weaknesses from a technical perspective. Although DNNs can generalise to unseen data from the same distribution, they often struggle to generalise outside of the distribution of data on which the DNN is trained [45]. These issues pertain to the interpretability and robustness of DNNs, two major challenges that need to be addressed before DNNs are embraced in the real world.

Interpretability [159] describes the necessity of being able to look inside the model, in order to have an understanding of how the model reaches a decision. This is particularly important to verify that the model is learning from the salient regions of the data. It allows us to see whether the model is looking at appropriate features, or whether it has found a shortcut [62] which allows it to cheat. For example, a convolutional neural network (CNN) trained to classify between cows and camels will learn that cows are usually captured on grassy plains and camels on sandy desert, which causes a misclassification when either animal is photographed outside of its natural habitat, such as a cow on the beach [12]. Interpretability aims to illuminate the areas the model identifies as important, to allow us to ensure this cheating does not occur.

Robustness evaluates the ability of DNNs to perform effectively in new conditions. This may be due to a distribution shift of the data that it is analysing, or due to an attack on data from the learnt distribution. Although DNNs have received much acclaim in academic circles for their generalisation ability, there is still resistance to utilise them for many real-world problems, particularly where data is sensitive. The lack of robustness to real-world data plays a large part in this. Academic data sets on which DNNs thrive are typically high quality and carefully constructed to ensure optimal performance. However, data in the real-world is often noisy, low-quality, highly variable, and unstructured.

Security and medicine are two domains that can potentially benefit the most from the widespread adoption of deep learning in the real world. Detection, face verification, and re-identification are

among the most popular tasks that use DNNs in academia [21]. DNNs have also demonstrated the ability to outperform the ability of human experts to diagnose a wide variety of diseases [184]. DNNs have the further advantage of not being affected by emotions or fatigue, so decision-making remains consistent. However, medical and security applications are highly sensitive, with critical decisions needing to be made. DNNs have yet to be fully embraced for real-world applications because they have not yet proven to be trustworthy for such crucial decisions.

1.2 Motivation

This thesis specifically explores the ability of attention mechanisms to jointly contribute to improving interpretability and robustness of DNNs. Attention is an important part of the human visual system. This is most famously demonstrated by the example of *the invisible gorilla* [27, 54], where study participants are asked to count the number of passes made in a basketball video. With the participants focusing their attention on the basketball, a gorilla can walk across the court entirely undetected. Being able to selectively attend to certain visual regions has assisted humans to excel at individual tasks (such as counting basketball passes). Guided by the example given to us by nature, it is therefore intuitive to incorporate attention into CNNs to improve their ability to perform on an individual task.

The attention mechanisms we study in this thesis have similar properties. They become a *dependent* part of the neural network and their associated weights are learnt simultaneously as the rest of the network. By analysing these weights, the features that the network deems important can be interpreted. Additionally, due to the network giving certain features more attention and other features less, important features gain more influence whereas the influence of superfluous (potentially confounding) features is minimised. This helps to promote robustness in the model. Security and medicine are two of the most sensitive areas where DNNs are commonly applied. Therefore, these are the areas that will be explored.

The response of the AI community to the Covid-19 pandemic exemplifies the issues that are currently faced. Roberts *et al.* [197] conduct a survey of AI models used to diagnose coronavirus. From 1 January to 30 October 2020, 2212 such studies were identified, but none of these models were deemed suitable for clinical use. In their analysis, they mention that many models do not

demonstrate the necessary robustness. Further, they "*stress the importance of not only reporting results but also demonstrating model interpretability with methods such as saliency maps, which is a necessary consideration for adoption into clinical practice*". Attention mechanisms have the potential to solve these two problems simultaneously, as they provide a saliency map which is easy to inspect, and help to improve robustness by minimising the impact of spurious features.

1.2.1 Interpretability

One of the significant drawbacks of deep learning is that it is not immediately possible to understand the process that the algorithm has used to arrive at a decision. DNNs often contain billions of parameters, and therefore too much information for a human to follow. This results in the *black-box* problem [26], which characterises the algorithm as an unintuitive, impenetrable object that produces an answer but does not explain the reason for that answer. This inhibits the ability of DNNs to be used in real-world scenarios, *despite* their demonstrable efficacy, because there is a possibility that a DNN has somehow cheated, or used a shortcut method of coming to a conclusion.

This drastically reduces the likelihood of deep learning-based methods being taken up in the real world. Without verification that the DNN has returned a result in an appropriate way, the result itself cannot be trusted. If a DNN has learned a shortcut on the data distribution used in training that does not exist outside of that distribution, its efficacy will be greatly reduced. In medicine, this scenario could result in a missed diagnosis, which can worsen patient health outcomes and even potentially cause a premature death. However, if a DNN can highlight the important information that it has observed to make its decision, this reasoning can be analysed by the human decision-maker. This significantly improves the trust in the model, aiding the adoption of DNNs in the real world.

For the purpose of this thesis, interpretability is defined as follows:

Definition 1.2.1. *Interpretability* is the design of machine learning models so that a human can understand the inner mechanics that lead to a decision being made.

This definition is inspired by Gilpin *et al.* [68], who state "*A third approach to interpretability is to create explanation-producing systems with architectures that are designed to simplify interpre-*

tation of their own behavior. Such architectures can be designed to make either their processing, representations, or other aspects of their operation easier for people to understand."

Note that there is a subtle difference between interpretability and *explainability*. Explainability refers to algorithms that are separate from the model itself that try to explain a decision made by the model. Interpretability, as used in this thesis, instead focuses on mechanisms that are part of the model that makes the decision, that can also be understood by humans. Explainability tries to explain a black box, interpretability requires adjusting the black box so it becomes more transparent. Explainability and interpretability are often used interchangeably, incorrectly. Definition 1.2.1 is focused only on the interpretability, in line with the discussion in Rudin *et al.* [200], who introduce fundamental principles of interpretable deep learning. Of particular importance is Principle 5, which states "*For high stakes decisions, interpretable models should be used if possible, rather than "explained" black box models*".

1.2.2 Robustness

As deep learning becomes more commonplace in society, it is essential to study its ability to function effectively in real-world scenarios. Most studies demonstrate that DNNs perform well on lab-created data sets, which tend to be high-quality and lack noise. These idealistic tasks are simplistic compared to most scenarios where DNNs would need to be deployed in the real world.

To deploy a machine learning model in a real-world setting, it must be robust when generalising to data from a new, unseen distribution. A typical machine learning experiment will split a data set into a train and a test set. However, despite the test set being unseen during training, its distribution is often independent and identically distributed (i.i.d) to the training set. DNNs will typically perform excellently on the unseen data from this identical distribution. However, when DNNs are tested on new distributions that are not i.i.d. to the trained distribution, they tend to struggle much more.

Another potential problem is the issue of shortcut learning [62]. If a model can identify an easy way to learn a relationship on a specific distribution, it will likely prioritise this relationship over a more complex relationship that is more representative of the real underlying structure of data.

For the purpose of this thesis, robustness is defined as follows:

Definition 1.2.2. *Robustness* is the ability of a machine learning model to lose minimal performance when applied to data likely to be discovered in the real world compared to i.i.d. data.

Concretely, "data likely to be discovered in the real world" refers to data that is non-i.i.d. to the training data, or data that is noisy, low-quality, or presents significant challenges like outliers and mislabelling. This data is commonly found in the real-world but popular academic data sets have usually been through many phases of cleaning, which means they are easier to work with. For example, two models may have equal performance on a standard train/test split of a data set, but the most robust model would perform better when transferred to real-world data sets for the same task.

1.3 Proposal

Attention modules are mechanisms that can be incorporated into DNNs. They are designed so that the information flows into a subsidiary branch, undergoes a transformation and is fed back into the main branch via a multiplicative operator. The attention module therefore learns, via backpropagation, the salient regions of the input data, which then receive a larger weight following the multiplicative operation.

We explore the use of attention modules to improve the interpretability *and* robustness of DNNs. Because the flow of information is controlled in such a way that the DNN has to learn to assign a weight to each input feature, these weights can be examined after a decision has been made to understand what the DNN considers important. For images, this leads to a heatmap indicating which image regions are determined to be important, which should align with human understanding of the task at hand.

Furthermore, because the DNN assigns a greater weight to features that are most effective at minimising the loss from a backpropagation perspective, this has the additional benefit of improving robustness: noisy features that contribute less to the classification performance will usually be assigned a smaller weight, so the model can focus on what is really important to perform well in the task. This is important to make the model robust because certain features may help to minimise the training loss but for the wrong reasons. In the earlier example of a model being

confused when classifying a cow on a beach, if the model is trained with an attention module to focus on the foreground mammal, the confounding background information has much lower contribution to the output classification, which allows the model to remain robust despite the change of scenery.

Lastly, even if we do not consider interpretability and robustness, the pure ability of attention mechanisms to improve model performance cannot be understated. Transformers [234], based on self-attention and cross-attention mechanisms, have completely revolutionised natural language processing. Vision transformers [53] are also showing performance competitive with CNNs in the computer vision domain. Although other methods exist that can improve model interpretability or robustness in isolation, attention mechanisms are explored in this thesis for their ability to simultaneously improve interpretability, robustness, and performance.

The goal of this thesis is to demonstrate that attention mechanisms improve robustness and interpretability of DNNs, with a particular focus on domains in medicine and security where these traits are most important. This brings major challenges. Firstly, demonstrating that attention mechanisms improve robustness and are interpretable over a narrow set of tasks would not be sufficient. For example, focusing purely on discriminative vision models in security would go some way to demonstrating robustness, but would not give any justification of the ability of attention mechanisms on tabular data in mechanism. Therefore, this thesis will attempt to select applications that are dissimilar from each other. Verifying interpretability and robustness properties at different extremes provides more convincing evidence of these properties across the entire spectrum. Secondly, sourcing appropriate data sets is a significant challenge, particularly in these sensitive domains. To evaluate model robustness, real-world data sets are required, otherwise it is not possible to recommend incorporating attention mechanisms for real-world applications. However, in medicine and security, there are strong ethical considerations at play regarding safeguarding personal information of data samples. As data is difficult to source, this places a natural limit on the number of tasks that can be presented, and optimal applications may have to be looked over.

1.4 Contributions

This thesis demonstrates the flexibility and efficacy of attention mechanisms to enhance the robustness and interpretability of deep neural networks. More robust deep neural networks can be applied to a wider range of practical challenges. In particular, this thesis focuses on the impact that attention mechanisms can have to handle data distribution shift and low-quality data. This strongly facilitates the application of deep neural networks in the real world, where data is noisier, messier, and with less-structure than in data sets commonly seen in academia. Therefore, work of this kind is essential for deep neural networks to have real-world impact. We also focus on inspecting attention mechanisms to be able to interpret the network. Without an understanding of how a model reaches a decision, it is very difficult to trust that decision, especially if there is a significant negative outcome for getting the decision wrong. Attention mechanisms can be inspect to evaluate what the model deems important, allowing us to understand why the model reached the classification that it made. If the models interpretation of the data matches that of the user, the model is more likely to be trusted. As deep neural networks have outperformed humans across a wide range of tasks, the sooner they can be trusted, the more positive real-world impact they can have.

In short, this thesis identifies two major roadblocks that stop deep neural networks being utilised in the real world: lack of interpretability and lack of robustness. This thesis then proceeds to demonstrate, across a wide range of applications, that attention mechanisms help to solve both of these roadblocks, simultaneously.

The chapter-by-chapter contributions of this thesis are as follows:

- Similar to the problems with using AI to diagnose covid-19 discussed earlier, research on using AI to diagnose Schizophrenia has also recently been criticised for not translating to the real world. To handle this, a new data set is constructed which complies with current diagnostic guidelines. A novel fused attention module for Schizophrenia diagnosis with one pathway for robustness and another pathway for interpretability is proposed. Furthermore, we design stress tests which attack the data to examine the robustness of our method compared to benchmarks, and show that our module is both most interpretable and robust to perturbations.

- We propose a multi-scale attention block and demonstrate that it can be used within a DNN system to transfer makeup from one image to another. This multi-scale attention framework outperforms the state-of-the-art makeup style transfer frameworks on low-quality makeup data. Furthermore, the framework does not rely on external modules and is generic enough that it could also be applied for other image-to-image translation tasks. Therefore it is hugely advantageous compared to the current trends in the field.
- The task of re-identification is thoroughly evaluated. Channel attention is demonstrated to improve performance of person re-identification, along side an additional proposed attention function which takes advantage of the variance of each feature. To tackle the robustness of re-id, the task of simultaneously re-identifying pedestrians and vehicles is proposed. New loss function terms are proposed, which attend to positive pairs and negative pairs, respectively. Lastly, on this new task, the challenge of out-of-distribution generalisation is tackled, and the proposed channel attention model outperforms all other tested frameworks.
- We propose the task of UAV re-identification and perform an exhaustive benchmark study on two settings: 'Temporally-Near' and 'Big-to-Small'. These are designed to mimic scenarios in which UAV re-identification would be useful in real life. Temporally-Near is designed because tracking algorithms include a re-id component that matches objects in nearby frames. Big-to-Small is designed to challenge the re-id framework's robustness, to match distant UAVs to known UAVs, to assess whether a detected UAV may be a threat when there is still time to deal with that threat.

1.5 Thesis Outline

Chapter 2 covers a vast amount of prior work to give a mathematical background of relevant frameworks, further motivate the problem, and explore past studies on attention modules and representation learning within the medical and security domains. As we are concerned with improving the real-world adoption of deep learning systems like CNNs and GANs, these systems are introduced mathematically to set up for future chapters. Following this, a literature review of interpretability and robustness is performed, focusing on the need for DNNs to exhibit these properties and a variety of common interpretability and robustness approaches, along with their limitations. Chapter

2.2 introduces channel attention, spatial attention, and self-attention, that will be studied in this thesis, along with a review of how they have been used in the past. Lastly, the literature review covers a broad range of research applying DNNs to medicine and security, the challenges associated with these critical domains explicitly, and the limitations that deny deep models to be utilised in the real world. Relevant works that take advantage of attention mechanisms for each domain are also explored.

Chapter 3 exhibits the benefits of using attention modules within a standard neural network structure, for the task of schizophrenia diagnosis. A dual attention neural network is constructed, taking advantage of channel attention to improve the DNNs robustness, and self-attention to allow clinicians to interpret the model. Stress tests to evaluate the robustness of all benchmarks are also designed, and show the weakness of standard DNNs and traditional ML methods to generalise outside of the seen distribution, while highlighting the ability of attention modules to improve this capability.

From Chapter 4-6, the use of attention modules within convolutional neural networks are explored. Chapter 4 demonstrates the importance of using attention mechanisms on data that is collected in the wild. It is difficult for neural networks to disentangle the human faces and makeup that is applied to them. This can inhibit the ability of face recognition models to function effectively. Makeup style transfer has been explored to generate synthetic data to allow the model to learn how a subject looks with and without makeup. However, these models are all trained on lab-created data sets with high-quality, front-facing images. When images are taken from the real world, often cropped, angled, and low-resolution, the state-of-the-art makeup style transfer models are much less effective. However, a combination of spatial and channel attention can allow for a much more robust model which is effective in the real-world scenario.

Chapter 5 focuses on the challenge of re-identifying both pedestrians and vehicles. This chapter combines two works: the first explores the potential of channel attention when applied to the task of person re-identification; the second proposes the task of unifying person and vehicle re-identification to create a complex challenge which requires a more robust feature representation. Two additional unification loss terms are designed to improve the model's capability of handling this mixed data. Finally, the two works are conjoined by exploring the ability of the attentional model for out-of-distribution generalisation on the complex mixed data task. Overall, attention is

found to be essential to maintain robustness.

Chapter 6 introduces the challenge of UAV re-identification, and performs an exhaustive benchmarking of popular re-identification techniques. Two settings are created: a 'Temporally-Near' setting which models the challenge of re-identification modules within tracking frameworks, and a 'Big-to-Small' setting that models security systems ability to match small and distant UAVs with large UAVs stored in a local database. In particular, the self-attention mechanism inherent in vision transformers, makes them much more robust to changes in object size compared with traditional computer vision models with no attention.

Chapter 2

Literature Review

DNNs have become the most popular machine learning tool in recent years as they automatically learn effective features from the data, rather than rely on human input to design hand-crafted features. Even with expert knowledge, these algorithms were sometimes ineffective, or had unwittingly incorporated the human experts' bias. Therefore, DNNs typically attain stronger performance and generalisation ability compared with traditional machine learning algorithms [5]. This is particularly true of CNNs when applied to images, as they are able to take advantage of implicit bias within the composition of the image to extract information.

This section will give an overview of DNNs and CNNs, and explore the strategies that have been used to interpret them, and make them more robust. The applications of DNNs and CNNs within the medical and security domains will also be covered. Because this thesis is focused on interpretability and robustness within deep learning, methods of learning representations with traditional machine learning will not be covered in this section.

2.1 Representation Learning

One of the key roles of any machine learning algorithm is to represent data instances (that may be high-dimensional) in such a way that the initial instances may be easily classified. This learnt representation is often in the form a feature vector. The first section of this chapter will begin by exploring how deep neural networks obtain a feature vector from high dimensional data and explore prior work to improve the quality of this representation.

2.1.1 Deep Neural Networks

Deep Neural Networks (DNN) [121] are biologically inspired structures composed of layers of neurons. Every instance from the input data is sequentially passed forward through each DNN layer. The output of the DNN is a feature representation of the initial input, with a significantly lower dimensionality [85].

Let $f^* : D \longrightarrow \{0,1\}^n$ be the function which correctly maps input data, $\mathbf{x}_i \in X$, to the true label, y_i, for all $i = 1, \ldots, n$; i.e.

$$y = f^*(\mathbf{x}). \tag{2.1}$$

Here, n is the total number of data samples, $X \subset \mathbb{R}^{k \times n}$ is collected data with k features of n samples and $\mathbf{x}_i$ refers to the data instance i.

A DNN learns the function $f : X \longrightarrow \{0,1\}^n$ to get as close to f^* as possible. For $i = 1, \ldots, n$, we obtain

$$\hat{y_i} = f(\mathbf{x}_i; \theta), \tag{2.2}$$

where $\hat{y_i}$ is the predicted label of instance i, $\mathbf{x}_i \in X$ is the observed data of instance i, and θ are the weights and biases that the DNN learns. The goal is to minimise deviation from the predicted labels $\hat{y}$ to the true labels y.

The DNN f is comprised of multiple layers $f_1, \ldots, f_L$ given by:

$$\mathbf{h}_1 = f_1(\mathbf{x}; \mathbf{W}_l, \mathbf{b}_l) \quad = \phi(\mathbf{W}_0{}^\top \mathbf{x} + \mathbf{b}_0), \tag{2.3}$$

$$\mathbf{h}_l = f_l(\mathbf{h}_{l-1}; \mathbf{W}_l, \mathbf{b}_l) = \phi(\mathbf{W}_l{}^\top \mathbf{h}_{l-1} + \mathbf{b}_l), \tag{2.4}$$

where ϕ is the non-linear activation function, $\mathbf{x}$ is the input data, $\mathbf{h}_l$ is the output of the l^{th} hidden layer, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the neuron weights and biases of hidden layer l respectively.

The quality of the representation will be poor in the initial stages of training. To optimise the representation, a loss function is used to evaluate the error between the estimated labels and real labels. Backpropagation of errors [137, 202] is then used to identify the direction that the weights in the neural network need to be adjusted to reduce that error. Over the course of training, the error of an appropriate model with sufficiently descriptive data will converge towards zero.

One common technique to improve the ability of DNNs to train and avoid overfitting is dropout [220]. A proportion, p, of the nodes within each layer of the neural network are randomly eliminated at each forward pass of training. This stops layers of hidden neurons being overly reliant on a small number of nodes, which can often happen when the data set is small and can be easily estimated.

Another technique which has been widely adopted to improve performance of DNNs is batch normalisation [94]. Data is often normalised before training starts, but the operation of passing data through a layer of a neural network can result de-normalise the data between layers. Batch normalisation re-normalises the data after each layer, resulting in a more stable and smooth training procedure, with gradients that are more predictive and well-behaved [203].

2.1.2 Convolutional Neural Networks

Convolutional neural networks (CNN) are an extension of neural networks that are designed to be more suitable for analysis on image data. At each layer, a set of filters (usually with $\mathrm{height} = \mathrm{width} = 3$) act as a sliding window to scan the feature maps calculated at the previous layer. These filters contain learnt parameters and are commonly referred to as the features learnt by the network. Intuitively, CNNs are designed to focus on local receptive fields, which introduces an inductive bias whereby relationships between nearby pixels are deemed as important, whilst relationships between distant pixels are not considered. *The process of convolution itself can be thought of as a hand-crafted attention mechanism.* However, this can result in weaker performance in scenarios where the relationship between distant pixels is important, such as image segmentation.

Convolution can be represented mathematically as follows: at the l^{th} layer, a CNN takes a stack of feature maps $f : \mathbb{Z}^2 \longrightarrow \mathbb{R}^{K_l}$ and convolves it with filters, ψ, from the $(l+1)^{\mathrm{th}}$ layer

$$[f \star \psi](x) = \sum_{z \in \mathbb{Z}^2} \sum_{k=1}^{K_{l+1}} f_k(z) \psi_k^i(x - z), \tag{2.5}$$

where z is the co-ordinate position, $i \in 1, \ldots, K_{l+1}$, k is the k^{th} filter of ψ at layer l.

CNNs were introduced for recognition on the mnist data set [48] by LeCun *et al.* [122, 123], who introduced LeNet, a simple 5 layer neural network consisting of convolutions, sub-sampling and fully connected layers. However, machines were not powerful enough to handle the large number

of computations for CNNs to be useful for larger data. Nearly fifteen years later, as technology became more advanced, a major breakthrough was made when Krizhevsky *et al.* [113] won the ImageNet [46] competition using their CNN, Alexnet. AlexNet began the trend of going deeper with CNNs, with five convolutional layers and 3 fully connected layers.

Szegedy *et al.* [224] proposed the Inception network, which started to make advancements on both the depth, and the design of CNNs. In particular, they design an inception module, which processes the input at three different convolutional levels: $1 \times 1, 3 \times 3$, and 5×5, as well as performing max-pooling. The outputs of these processes then get concatenated to form the output of the Inception module. This allows their model to be much deeper, with 22 layers.

He *et al.* [81] proposed ResNet, which uses residual layers to allow much deeper neural networks. They observed that as CNNs got deeper, training performance got worse, i.e. additional convolutional layers performed worse than an identity function. To rectify this, they introduced skip connections, where the layer input is added to layer output, i.e combining convolutional layers with an identity mapping. This allows deep networks to expand beyond 100 layers, although the most commonly used network is ResNet-50. Li *et al.* [129] showed that the loss landscape of ResNet is much smoother than the loss landscape of a regular CNN, which allows for more efficient and consistent optimisation.

Off the back of the success of these models, many different designs for CNNs have been proposed. The state-of-the-art fully-convolutional models on ImageNet today include EfficientNet [227] and NFNet[19]. EfficientNet rely on model scaling that uniformly scales depth, width, and resolution using a compound coefficient. This allows it to attain state-of-the-art performance while also being over six times faster than conventional models. NFNet identify that batch normalisation can severely hinder CNNs due to computational cost, a difference between model behaviour during training and testing, and removing independence between training samples. Batch normalisation is therefore replaced with the proposed adaptive gradient clipping, where the gradient clipping parameter λ is chosen adaptively based on the weights of the model.

2.1.3 Loss Functions

To optimise a DNN with backpropagation, an appropriate loss function must be utilised to guide the algorithm towards the global minimum. Common loss functions to be explored in this chapter

include cross-entropy and triplet loss.

Cross-entropy Loss

Cross-entropy is the most commonly used loss function for classification tasks. Let p and q be two distributions over $\mathbb{R}$. The negative likelihood function, $-\ln q(x)$ measures information given by q. The cross-entropy of p with respect to q instead measures the information given by $q(x)$ that is assessed by $p(x)$ [24]. In its general form, it is given by

$$\mathcal{L}_{\mathrm{CE}} = \mathbb{E}_p[-\ln q(x)] = -\int_{\mathbb{R}} p(x) \ln q(x) \, \mathrm{d}x. \tag{2.6}$$

For our purposes, we may simplify this to the discrete cases as we will work with discrete data throughout this thesis.

Binary cross-entropy is used for binary decision making. This is most commonly seen in medical applications where a patient either does or does not have a disease. For seen data X, the binary cross-entropy loss function is given by

$$\mathcal{L}_{\mathrm{BCE}} = -\sum_{x \in X} (y \ln(f(x;\theta)) + (1-y) \ln(1 - f(x;\theta))), \tag{2.7}$$

where y is the class and f is the output of the DNN f with parameters θ that an observation x belongs to the class $y = 1$.

Categorical cross-entropy can be seen as an extension of binary cross-entropy where there are more than two output classes. The categorical cross-entropy is given by

$$\mathcal{L}_{\mathrm{CCE}} = -\sum_{x \in X} y(x) \ln\left(f(x;\theta)\right), \tag{2.8}$$

where $y(x)$ is the true class label of data instance x. From this point on, we abuse notation and drop the usage of parameters θ when describing a DNN $f(x)$.

Triplet loss

We denote a triplet, $t = (x, x^+, x^-)$, where x is the query image, x^+ is a positive image, and x^- is a negative image. The triplet loss function is formulated as follows:

$$\mathcal{L}_{trp} = \sum_{t \in \mathcal{T}} \max((||f(x) - f(x^+)||_2 - ||f(x) - f(x^-)||_2 + \alpha), 0), \tag{2.9}$$

where the feature vector of an image x obtained from the convolutional neural network is denoted as $f(x)$, $\mathcal{T}$ is the set of mined triplets and $|| \cdot ||_2$ denotes the Euclidean distance. This loss will force negative images to be a distance of at least α away from the positive pair.

Let p be the identity of the image $x_{p,i}$ in the batch B, where $f(x_{p,i})$ is its feature vector, $p = 1, \ldots, P$ and $i = 1, \ldots, 4$. Each query image $x_{p,i}$ is paired with its hardest positive image x^+ and hardest negative image x^-, which are found via the equations:

$$x^+ = \max_{x_{q,j} \in B} (||(f(x_{p,i}) - f(x_{q,j})||_2), \text{ where } p = q, \tag{2.10}$$

$$x^- = \min_{x_{q,j} \in B} (||(f(x_{p,i}) - f(x_{q,j})||_2), \text{ where } p \neq q. \tag{2.11}$$

From equations (6.2) - (2.11), we see that obtaining the feature representation $f(x)$, and computing the distance between feature representations of any two images, $||(f(x_1) - f(x_2)||$, are essential components of the triplet loss. We focus our research on improving these two aspects.

2.1.4 Generative Adversarial Networks

Generative Adversarial Networks (GAN) [73] are one of the most well-studied deep learning architectures. Deep learning was becoming hugely popular following the success of Krizhevsky *et al.* [113], but was only successful for discriminative tasks like classification. DNNs had not made much of an impact when it came to generative modelling. GANs were an attempt to take advantage of the major successes in discriminative modelling to drive generative modelling.

Original Architecture

GANs are formulated as a two-player minimax game between a generator and a discriminator. The generator aims to generate a realistic image, whilst the discriminator attempts to distinguish

between real and generated images. The original formulation is a game-theoretic one, with generator and discriminator both starting the game with zero knowledge of the problem (i.e. randomly initialised). The generator and discriminator are then shown to reach a Nash equilibrium [176] when the discriminator is indifferent between guessing that an image is real or fake. Note that the discriminator is the only player to observe any real data, whilst the generator can only learn via successfully fooling the discriminator. This makes GANs a combination of supervised and unsupervised learning.

Game theory often assumes that both players play optimally. DNNs are selected to achieve as close to optimal play as possible. To ensure this, the loss function needs to be formalised in a manner whereby the generator and discriminator are actively competing against each other. The proposed loss function is a two-player, continuous extension of the binary cross-entropy loss function in Equation 2.7. The loss is presented as:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}} \left[\log D(x)\right] + \mathbb{E}_{z\sim p_z} \left[\log(1 - D(G(z)))\right], \tag{2.12}$$

where D is the discriminator, G is the generator, p_{data} is the real data probability distribution, and p_z is the n-dimensional distribution randomly sampled from as input to the generator.

Improvements

Training Stability:

GANs are notoriously difficult to train. If either the discriminator or the generator become too strong relative to the other player, training will collapse as e.g. the generator will always fool the discriminator, so can't get any useful feedback to further improve.

Arjovsky *et al.* [8] show that the formulation in Equation 2.14 essentially attempts to minimise the Kullback-Leibler (KL) divergence [114] between p_g, the generated data, and p_{data}, the real data. Instead, they prove that minimising the Wasserstein distance between real data and generated images. In doing so, they develop an algorithm that allows for alternated training of generator and discriminator, so they are not so dependant on each other during training, which further helps to improve stability. Furthermore, most GANs at that time struggled with mode collapse, where the generator discovers one particular image it can generate to fool the discriminator, and only

generates that image. An additional advantage of their proposed Wasserstein GAN is that it did not suffer from mode collapse.

Wasserstein GANs also obey a property called Lipschitz continuity [57]. Miyato *et al.* [158] searched for a discriminator from the set of K-Lipschitz continuous functions:

$$\max_{||f||_{\text{Lip}} \leq K} \mathcal{L}(G, D), \tag{2.13}$$

where $\mathcal{L}(G, D)$ is a standard GAN loss formulation and $||f||_{\text{Lip}}$ is the smallest M where $||f(x_i) - f(x_j)|| ||x - x'|| \leq M$ for all x_i, x_j. They propose spectral normalisation to constrain each layer of the discriminator to be precisely 1-Lipschitz. This has resulted in significantly more stable training procedures when following the original loss formulation, and has thus resulted in major performance improvements in training GANs.

Architecture Improvements:

As well as new formulations to improve training stability, a large number of model architectures have been designed to train more effective generators and discriminators. The original GAN in [73] used standard fully-connected neural networks. Radford *et al.* [189] proposed Deep Convolutional GAN by upgrading these models to CNNs, as CNNs had been shown to perform better on images than the traditional neural networks. SAGAN [275] went a step further, and proposed Self-Attention GAN, which made use of the self-attention mechanism from [234, 244] to enable the generator and discriminator to discover long-term dependencies between distant pixels. However, GANs are already computationally expensive due to needing to train two models simultaneously. Performing self-attention across pixels drastically adds to that burden.

An alternative research direction explored starting training with very small images. Karras *et al.* [105] propose ProgGAN, which progressively grows the size of images that both generator and discriminator see as the models get further into training, starting with 4×4 images, moving up to 8×8 etc. Importantly, the model for image size $h_t \times w_t$, where h is height, $wiswidth$ and t is a given timestep, will still pass through the layers trained for timesteps $t' < t$, and all layers will continue training. As well as improved image quality, this method hugely benefitted the stability of training, as generating smaller images is a much simpler problem than generating larger ones, so the model benefits from learning in this manner. This model was extended into StyleGAN

[106], which first maps latent vectors z into an intermediate latent space, which can impact the generator via adaptive instance normalisation [92] at each layer for every image size. Although the generative power GANs has now been surpassed by other methods such as diffusion networks [86], the StyleGAN family of networks remain state-of-the-art for GANs.

Conditional Generation

Conditional GANs [157] were proposed for conditional image synthesis, following a very similar formulation to the original GAN paper [73]. They propose a small change to Equation 2.14 with a conditional constraint:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}} \left[\log D(x|c)\right] + \mathbb{E}_{z\sim p_z} \left[\log(1 - D(G(z|c)))\right], \tag{2.14}$$

where c is the data class to be generated, and all other terms are identical to those in Equation 2.14. In practice, this is achieved by concatenating the one-hot encoding of the class c to the input latent vector z, and the first hidden representation of the data sample x.

CycleGAN [296] was a revolutionary framework which changed the entire landscape of conditional image synthesis. In CycleGAN, two GANs work in tandem with each other. The idea of CycleGAN is similar to that of an auto-encoder [201]. In an auto-encoder, an image is encoded by an encoder network M_E, before a decoder M_D attempts to reconstruct the image. With a CycleGAN, the image goes through two separate GANs, that train on different domains. The first GAN will convert the image from one domain to the other, while the second GAN will try to reconstruct the image.

Formally, given a source domain, X, and a target domain, Y, a CycleGAN will try to convert an image $x \in X$ into a fake image $\tilde{y} \in Y$. This CycleGAN consists of two GANs, (G_X, D_X) and (G_Y, D_Y) with respective losses:

$$\mathcal{L}^X_{\text{GAN}} = \min_{G_X} \max_{D_X} \; \mathbb{E}_{x\sim p_{\text{X}}} \left[\log D_X(x)\right] + \mathbb{E}_{y\sim p_{\text{Y}}} \left[\log(1 - D_X(G_X(y)))\right], \tag{2.15}$$

$$\mathcal{L}^Y_{\text{GAN}} = \min_{G_Y} \max_{D_Y} \; \mathbb{E}_{y\sim p_{\text{Y}}} \left[\log D_Y(y)\right] + \mathbb{E}_{x\sim p_{\text{X}}} \left[\log(1 - D_Y(G_Y(x)))\right], \tag{2.16}$$

Compared to standard the GAN loss 2.14, each generator takes as input a sample from the opposite domain, rather than a randomly sampled latent vector. Also note that each discriminator only sees

data from their respective domain; i.e. if D_X is trained effectively, the expected output of (without loss of generality) $D_X(y)$ is FAKE. In the loss function, $D_X(G_X(y))$ forces G_X to convert y into an image believably from domain X.

On top of these objectives is a cycle consistency loss, which performs a similar role to the reconstruction loss as described earlier:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x\sim X}\left[||G_X(G_Y(x)) - x||\right] + \mathbb{E}_{y\sim Y}\left[||G_Y(G_X(y)) - y||\right]. \tag{2.17}$$

Note that this loss enables gradients to flow between generators via backpropagation.

The full CycleGAN loss function then becomes:

$$\mathcal{L}_{\text{CycleGAN}} = \mathcal{L}^X_{\text{GAN}} + \mathcal{L}^Y_{\text{GAN}} + \lambda\mathcal{L}_{\text{cyc}}, \tag{2.18}$$

where lambda is a tuneable parameter.

In this formulation, four models are trained simultaneously: two generators and two discriminators. As with only training one GAN, it can be difficult to find a strong balance between models to ensure that they converge, but the techniques discussed above to improve training stability for regular GANs translate well to CycleGAN.

2.1.5 Quality of Representation

Interpretability

For completeness, we will focus on interpretable and explainable models in this subsection. We will refer to these as internal and external explanations, respectively, where internal explanations refer to methods where the model, training or inference regime itself provides understanding of the decision making process, whilst external explanations keeps everything frozen and tries to uncover that process.

Internal Explanations:

To allow real world practitioners to trust a DNN and the representation that it generates, it is important that the neural network is interpretable. A number of different methods have been

proposed to expose the inner-workings of the black box [159], to enable practitioners to visualise the key factors that a DNN has considered when arriving at a decision.

A simple method to attempt to interpret neural networks is to attempt to understand the features that are learnt. This is most common in early works on CNNs [168] because the learned features are images themselves which are assigned a weight. It is well-known that features in early layers often represent building blocks of an object, such as straight lines and corners, whereas later layers have much more class-specific representations [72]. However, although it may be possible to track weights and features for small CNNs trained on simple tasks like MNIST [48], for complicated tasks and networks scaled up to billions of parameters, it is still asking too much of humans to understand how a model reaches a decision based on this primitive explanation. Feature visualisation can be extremely effective if the task is appropriate. Goh *et al.* [71] visualise multimodal neurons within CLIP [188] to understand the visual features associated with a textual input. Although this allows us to get a much better understanding of the model itself, it is less useful during inference when trying to understand exactly how a caption is generated from an image.

A common method to improve model interpretability is the addition of an extra regularisation function, often termed the interpretability loss [281]. The principle is to train a decision tree on a set of data, and a deep model, M, with a cumulative set of weights W. When the decision tree is trained, the average path length $\Omega(W)$ is calculated. Ideally, the interpretable decision tree would be able to explain the deep model. However, decision trees are not differentiable so no further training can be completed. A simple multi-layer-perceptron (MLP), $\hat{\Omega}(W)$, is therefore trained to approximate the decision tree, minimised with a squared-error loss. $\hat{\Omega}(W)$ is described as a surrogate loss function. Importantly, $\hat{\Omega}(W)$ can be trained simultaneously with the deep network, M, by continually computing the cumulative features, W, at specific points during training. This method does have limitations. The additional loss function skews the performance of the model during training [15]. All internal explanations will modify the performance of the model in some way. To adopt models into the real world, it is important that too much performance is not sacrificed for interpretability. Ideally, model performance would actually increase due to internal explanations. It is also slow to train three models simultaneously in order to obtain an explanation.

Recently, there has been a lot of work within natural language processing (NLP) architectures on retrieval-based mechanisms. This is particularly common for question-answering systems, with

REALM [78], RAG [127], and FiD [98] all being proposed recently. Given a data bank, such as wikipedia, each page is passed through an encoder network to obtain a feature representation. When a question is asked, it is also encoded (usually but not always with a different encoder, trained to align questions with answers of the data encoder), and a search mechanism such as cosine similarity is employed over the entirety of the data bank encodings. The question-answering model the makes a decision based on the original input and the passages that have been retrieved. This encourages the model to make use of retrieved information, instead of memorising all the information in the original data bank, which makes it very easy to understand how the model has reached a decision, simply by looking through the retrieved passages. This can be taken further by ranking the retrieved passages via attentional weight addition, as proposed in [97]. Borgeaud *et al.* [17] took this a step further by designing a fully generative language model, which obeys standard principles of auto-regression. Through using a retrieval mechanism rather than memorising the entirety of the data, they are able to compete with state-of-the-art models several sizes larger than theirs, whilst also being able to interpret the outputs by inspecting the retrieved samples. Furthermore, retrieval mechanisms are now being incorporated into image-based frameworks in a similar way. Blattman *et al.* [16] propose to use retrieval to improve the capability of diffusion-based generative models.

External Explanations:

Although the interpretability discussed in this thesis focuses on model-based interpretability, there is a great deal of work exploring the A very common way to explain how neural networks arrive at a decision is to generate and inspect saliency maps based on the neural network. An early work by Simonyan *et al.* [216] began a trend of using gradients of the neural network to generate these saliency maps. Specifically, an image, x, is passed through a CNN to obtain a class activation function, S_c, for class c. The gradient of the score S_c with respect to the pixels of x is computed. Intuitively, this pixel attribution map shows the change in confidence of the DNN that image x is in class c if each pixel undergoes a small change.

Grad-CAM [206] is an extension of this idea; however, the gradient is only backpropogated to the final convolutional layer, rather than all the way back to the image. Smilkov *et al.* [217] take this further by generating n versions of the image by adding noise, then averaging the pixel attribution maps for each image. Although they often look promising, gradient-based methods have some

large problems. They are often not robust to small image perturbations, as these often lead to drastic changes in the pixel attribution maps [66]. Furthermore, Adebayo *et al.* [2] found that these gradient-based methods are often insensitive to model and data.

SHapley Additive exPlanations (SHAP) [148] is an alternative line of work towards model explanation, based on Shapley values from game theory. Here, each feature is a player of the game, and the object is the output of the model. Shapley values assign importance to each feature based on the average marginal contribution of all feature permutations containing that feature. Many works attempted to utilise Shapley values [196, 221, 214] in order to explain machine learning black boxes. SHAP is a unification of these works, exploiting a linear additive framework. Although SHAP generally performs well, it is less frequently used for image models, because it requires the calculation of image superpixels as the input features. Even then, because Shapley values require calculation based on the the power set of input features, they are extremely slow to compute, and they scale especially poorly as feature sets grow larger.

A recent method for DNN explanation uses image generation techniques to generate counterfactuals. Counterfactual explanations are a subset of causal learning [182] whereby a model tries to identify 'if X had not occurred, Y would not have occurred' [159]. Ganalyze [70] uses GANs to generate a manifold of images that can be traversed to explain image aesthetics such as memorability. Lang *et al.* attach a DNN to a StyleGAN, and explore 'style space' to identify features important for classification and generate images on either side of the class border. However, these models really on generative networks that can be unstable to train and often require huge computational power, so are not currently available for common usage.

The attention mechanisms discussed in this thesis, presented in Section 2.2, can be thought of as combining traits of internal and external explanations. Similar to external explanations like Grad-CAM, they produce a saliency map to understand important parts of the decision-making process, but they do so as an internal mechanism, so they are actively part of making that decision.

Robustness

General Robustness Improvements: In order to make models more robust, a standard method is perturb the data distribution that the model is training on. This is commonly described as data augmentation [213]. The intuition is that each data sample is a discrete point in an N-dimensional

space, where N is the input dimensionality. By augmenting data, the data set size artificially increases, but more importantly, each discrete 'real' data point is now surrounded by multiple augmented data points, which has the impact of 'fuzzifying' the distribution. Basic data augmentation techniques such as flipping, cropping, rotation, translation and changing the colour space can already offer significant improvements. Other common techniques like noise injection [160], random erasing [287], and Sobel filtering [104]. Commonly, multiple of these augmentations will be sampled from and performed in random order, to force the model to learn useful representations even when the augmented image looks very different from the original.

Described by Yann LeCun as 'the dark matter of intelligence', self-supervised learning is a hugely popular technique to improve model robustness. Most popular in language models such as the BERT [50] and GPT [190, 20] families, self-supervised learning aims to best make use of the enormous amounts of data that can be collected, but is too large to be fully labelled. Self-supervision is usually performed at the pre-training stage, and followed by task-specific fine-tuning. BERT performs self-supervised learning with a masked-language modelling training scheme. Given a sampled sentence from the data set, 10% of tokens are randomly masked, and BERT has to learn to correctly fill in these tokens from a pre-computed vocabulary. GPT models' self-supervised pre-training scheme is similar but performed in an auto-regressive fashion; that is, it tries to completely reconstruct sentences by filling them in from left-to-right. GPT in particular has shown remarkable robustness, able to easily adapt to new tasks in few-shot, one-shot, and zero-shot settings with extremely impressive performance [20]. Due to the amount of success self-supervised learning has had on text, and the amount of available image data that is too large to label, self-supervision has also started to become popular in computer vision. Chen *et al.* [32] perform data augmentation on input images, then train a model to match the augmented image with the original, while also presented with a large number of negative samples. Bao *et al.* [11] propose BEiT, a vision transformer trained with *masked image modelling*, analagous to the masked sentence modelling of BERT. Images are split into patches, and each patch is assigned a token from a pre-trained image tokeniser, with a vocabulary size of 8192. As with BERT, a fraction of patches are masked out, and the model has to classify across the vocabulary which image token has been masked out. At time of publication BEiT attained state-of-the-art performance on ImageNet [46].

Data Quality:

Many deep learning models are trained on carefully controlled, high quality, lab-created data sets that have been constructed to ensure that information can be easily extracted. However, data encountered in the real world is rarely so idealistic, particularly in the domain of security. Images from video feeds can often be low resolution, or heavily affected by changing weather conditions, which greatly affects the performance of DNNs [161, 251]. However, little work has been done to analyse the ability of DNNs when applied to this more realistic scenario.

Kwan *et al.* compare tracking models on low quality videos [117] and find that none of the trackers perform well in all conditions. Yang *et al.* [267] perform an extensive analysis of detection methods in low-visibility environments, reporting significantly weaker results all round compared to standard detectors. Geirhos *et al.* [63] show that DNNs outperform humans at detection on non-distorted images, but when the images are perturbed with random noise, their performance drops drastically compared to humans. As a large proportion of real-world data is noisy, it is essential that models are trained to be robust to imperfect data.

Out-of-distribution Generalisation:

It is well known that DNNs are extremely proficient at generalising to unseen data. However, this is only true within the distribution of the data that the DNN was trained on. When the distribution changes, the DNN is no longer able to apply what it has learnt in an effective manner [210]. An intuitive example of this is voice recognition models that have been trained on audio samples with clear elocution often struggle to understand strong Scottish or Geordie accents.

An unsupervised method to handle out-of-distribution generalisation is towards training *disentangled* representations, where distinct (usually human-understandable) image features are cleanly separated in the feature representation, i.e. features should be independent. Higgins *et al.* [84] present an extension of a Variational Auto-Encoder (VAE) termed β-VAE, which contains an additional parameter β that controls the trade-off between latent channel capacity and independence constraints. FactorVAE [109] encourages independence across dimensions by training the distribution of representations to be factorial.

Recently the field of causal learning [182] has started to receive more focus, in particular for the task of supervised out-of-distribution classification [204]. Any given system is represented by a causal graph, explaining the causal direction between components, and a structural equations

model. However, causal inference and causal structural learning are very difficult to verify, with the ground-truth almost impossible to obtain because it is very rare that every relationship in a system can be modelled. Most works settle for providing a causal explanation [60, 199].

Arjovsky *et al.* propose Invariant Risk Minimisation [7] as an alternative to the commonly used Empirical Risk Minimisation [233]. Invariant risk minimisation considers data sets under multiple training environments and aims to minimise the risk on invariant correlations, rather than on spurious correlations. By focusing on invariant correlations, more useful relationships between data and labels are learnt, which allows for consistent learning across environments.

Stress Tests: Jacovi *et al.* [99] formalise the idea of trust in deep learning, framing the problem in terms of a contract. D'Amour *et al.* [45] build on this framework and present a detailed report on the issues of underspecification of DNNs. Underspecification is where multiple different possible configurations of weights can achieve equally strong results on the source data. Finding the single configuration to best solve the general problem, not just on that distribution, is a major challenge for deep neural networks. Applying stress tests to the input data is a simple, yet effective, way to identify the most appropriate configuration, because they "*probe a broader set of contracts*. Three types of stress-tests are presented: stratified performance evaluations, shifted performance evaluations and contrastive evaluations.

Stratified evaluations aim to discover whether a DNN has encoded inductive biases, the most famous example of which is presented by Buolamwini and Gebru [22]. A gender classification model is shown to perform much weaker when applied to darker females. Stratification is also a major consideration in healthcare: Oakden-Rayner *et al.* [164] show how hidden stratification can cause medical ML models to fail

Shifted performance evaluations focus on distribution shift. Koh *et al.* [111] present a benchmark of in-the-wild distribution shifts. Ten data sets are included, with each including a sub-population shift, domain generalisation, or both. Zhang *et al.* [274] present a causal view of neural network robustness and design a causal manipulation augmented model to model possible manipulations. A causal generative model is used to perturb input data via unseen causes, i.e. in a way that is consistent with the causal structure of the system being evaluated.

Contrastive evaluations are also common in the fairness literature, often using causal inference to

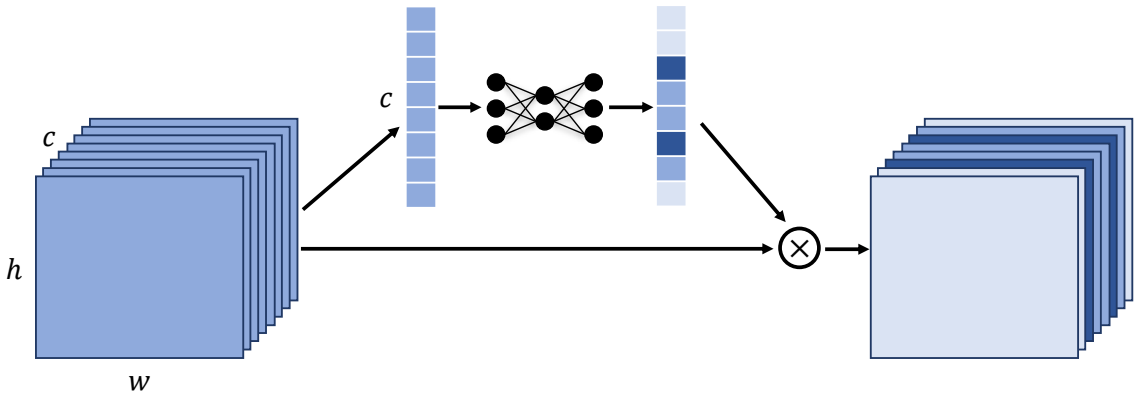


Figure 1: Channel Attention in Convolutional Neural Networks

check counterfactual fairness [116]. Fundamentally, counterfactual fairness checks that a decision that a DNN makes about an individual remains the same decision if that individual is placed into a different demographic. Contrastive evaluations are therefore focused on individual data samples rather than the distribution as whole.

2.2 Attention Mechanisms

Deep learning has high requirements before it can be used in real-world medical applications due to the absolute necessities of trust [1] and interpretability [236]. *Attention* is commonly used to handle these issues [37, 93]. The attention modules highlight important structures within data to assist classification performance. By inspecting these modules, we can better interpret how the neural network comes to a classification decision. There are various types of attention module, including self-attention [234], channel attention [89] and spatial attention [239].

2.2.1 Channel Attention

Channel attention is a mechanism that performs feature recalibration within the framework utilising it. By doing so, it selects features that are the most informative to the framework and accentuates them, while diminishing the importance of less useful features. These informative features then allow the framework to create a more informative feature representation which is more effective at separating classes.

In this regard, channel attention acts as a process to determine the weight of each channel at each layer of the model. Channel attention performs different roles throughout the network, getting more polarising at deeper layers. As a consequence, unimportant channels are mapped near to 0 in the final block of of a DNN, which has a large effect on the output feature representation of each image.

First, channel-wise spatial information is squeezed into a channel descriptor via Global Average Pooling. Formally, given a channel $u \in \mathbb{R}^{h \times w}$, we squeeze it to obtain its channel descriptor, d, as follows:

$$d = \text{squeeze}(u) := \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} u_{ij}, \tag{2.19}$$

where h and w are the height and width of the channel, respectively. These channel descriptors form a vector $\mathbf{z} = [d_1, \ldots, d_c]$ where c is the total number of channels.

Next, in order to calculate the channel-wise dependencies, this statistic needs to be excited. To achieve this, a simple gating mechanism with a sigmoid activation function is employed similarly to what is used within many spatial attention methods. The vector of squeezed channel descriptors $\mathbf{z}$ is passed through a dimensionality-reduction fully connected layer, a ReLU and then a dimensionality-increasing fully connected layer. This is then processed by a sigmoid activation to obtain the excited channel descriptors.

Formally, this excitation is written as:

$$\mathbf{s} = \text{excite}(\mathbf{z}) := \sigma\left(\mathbf{W_2}\delta(\mathbf{W_1}\mathbf{z})\right), \tag{2.20}$$

where $\mathbf{W_1} \in \mathbb{R}^{\frac{c}{r} \times c}$, $\mathbf{W_2} \in \mathbb{R}^{c \times \frac{c}{r}}$ are the parameters of the dimensionality-reduction and dimensionality-increasing layers respectively, δ is the ReLU function and σ is the sigmoid activation function.

2.2.2 Spatial Attention

Intuitively, spatial attention aims to focus the network on specific image regions, so that salient regions of the image have more influence over the final feature representation. Attention within human visual perception has been studied extensively. Because of this, there have been several

different formulations of spatial attention within CNNs. In this thesis, we refer to spatial attention as being analogous to channel attention but with pooling performed at each pixel location across channels, i.e. channels are averaged to obtain a $h \times w$ feature map representation of the layer in question. This feature map is passed through a convolutional block and aims to identify pixel regions that contribute most to minimising the error. This process is outlined in Figure 2.

Most formulations of spatial attention follow this structure, with incremental additions across the attention path. Many early works used attention for image captioning [262, 31], where attention helped the DNN to identify and classify foreground objects to describe in the caption. Jaderberg *et al.* [100] present Spatial Transformer Networks, which send layers through a localisation network before generating a grid containing foreground objects (note, this is not to be confused with Transformers introduced in the Self-Attention subsection). Wang *et al.* [239] compute attention as part of a residual, with feature maps processed by convolutional layers on the non-attention path.

Spatial attention is often seen in combination with channel attention. Harmonious Attention Networks Harmonious Attention Network (HACNN) [134] combine Spatial Attention with Channel Attention, and add a *hard attention* module, for the task of person re-identification. This combination of attention modules improves performance with negligible change in computational complexity. Harmonious attention provides a heatmap for each data sample, allowing the user to see where the model focuses to categorise the sample. It also improves the robustness of the model by lowering the amount of focus added to the background, so decisions are made based on salient information, such as clothes the pedestrian is carrying and objects they are holding. A more general convolution block attention module [255] has become popular, which sequentially performs channel attention followed by spatial attention.

2.2.3 Self-Attention

Transformers [234] have become the knew paradigm in the field of natural language processing (NLP) [50] due to their superior ability to analyse sequential information, particularly across long distances. They are now also beginning to make advancements in computer vision [192] and graphical data [235].

Self-attention learns three new representations for each input feature: a key, query, and value.

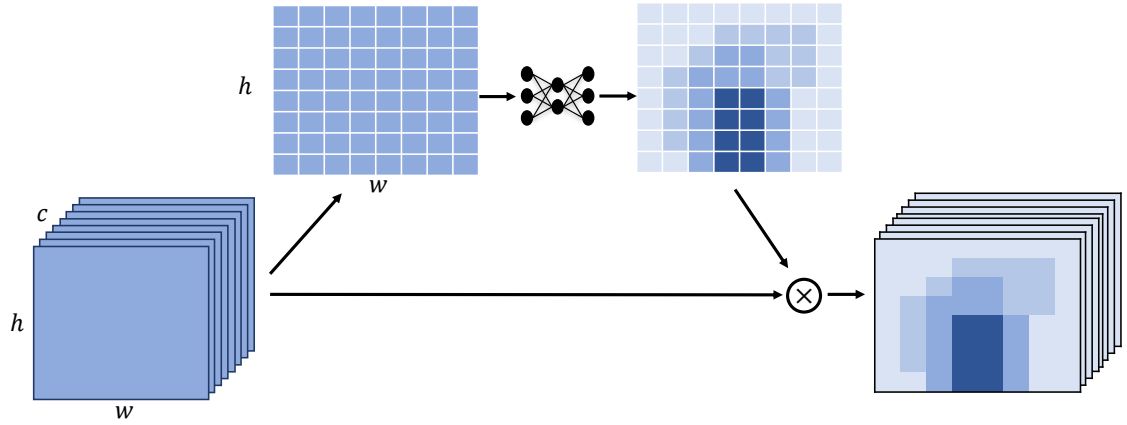


Figure 2: Spatial Attention in Convolutional Neural Networks

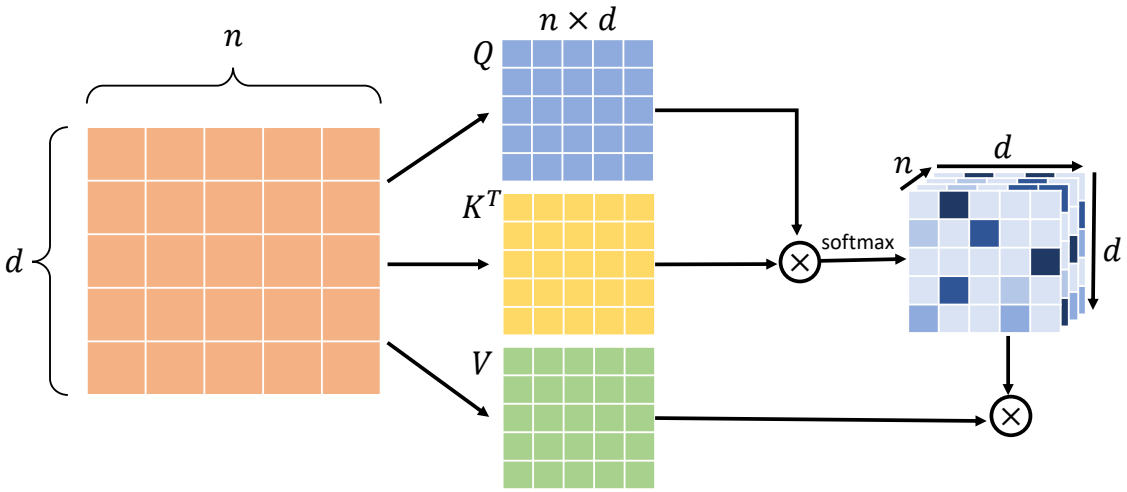


Figure 3: Self-Attention Mechanism

Give an input feature, i: the value vector is a hidden representation of that feature consisting of d attributes; the associated key vector indexes the value vector; and the query vector learns to find keys that are relevant to the value vector.

Given a query, q_i, we calculate attention score a_{ij} of all keys, k_j, using the cosine similarity:

$$a_{ij} = q_i \cdot k_j = ||q_i|| \times ||k_j|| \cos(\theta). \tag{2.21}$$

Noting that the dimensionality of all queries and keys remains constant, the more similar q_i and k_j are, the larger $q_i \cdot k_j$ will be. A softmax function is applied to polarise the attention weights for each query. Finally, the value matrix, V, is then multiplied by these attention scores. The entire

process can be neatly summarised as

$$\text{Attention}(K, Q, V) = \text{softmax}(K^\top Q)V, \tag{2.22}$$

with $K = \mathbf{W}_k x$, $Q = \mathbf{W}_q x$ and $V = \mathbf{W}_v x$, where $\mathbf{W}_k$,$\mathbf{W}_q$, and $\mathbf{W}_v$ are learned weights matrices for keys, queries and values, respectively. This process is shown visually in Figure 3.

Self-attention, although mostly used in NLP, has also seen use in computer vision. Wang *et al.* [244] propose non-local neural networks, which take each pixel location as a token to attend to other pixels. This helps to model long-range dependencies across the image. Notably, this self-attention mechanism occurs between convolutional layers.

Very recently, self-attention has started to replaced convolution entirely. Ramachandran *et al.* [192] proposed stand-alone self-attention, a fully attentional model using only self-attention layers without convolutions. This was extended to Axial DeepLab [241], which learned positional encodings for pixels and added axial attention, iteratively attending across rows and columns. A different line of work saw the introduction of Vision Transformers [53] which splits the image up into patches, and applies self-attention between these patches at each layer. This helps to greatly save on computational cost compared to computing the attention of each pixel. DINO [25] show that self-supervised vision transformers generate attentional saliency maps that are comparable in performance to state-of-the-art supervised segmentation models.

Attention has also been applied within Generative Adversarial Networks (GANs) [73] for image generation. Zhang *et al.* [275] propose Self-Attention GAN, which makes use of a key, query, value set to perform attention similarly to traditional transformers [234]. Tang *et al.* propose AttentionGAN [230] for image-to-image translation, where the attention masks focus on key areas to translate.

2.3 Applications in Medicine

As with most fields, DNNs have seen extensive use in the medical domain [194]. Causability and explainability two very essential challenges that are necessary to address in order to use artificial intelligence in medicine [88]. Rudin *et al.* [200] argue that designing interpretable models is more desirable than explaining decisions of black box models with tools like Grad-CAM [206]

presented in Section 2.1.5.

Interpretability and visualisation are essential when using DNNs for a medical purpose [236]. As an abundance of data has been collected in recent years, it is natural to consider data-driven models to aid clinicians in making decisions. However, to process such large amounts of data, these models are often highly complex. Without an appropriate method to infer *why* a model arrives at a decision, models are not deemed trustworthy, which limits their usage within the real-world.

This section will give an overview of applications of DNNs in medicine, and the current running themes to facilitate the usage of DNNs for real-world applications. The areas of clinical data, medical image data, and mental health will be explored separately.

2.3.1 Clinical Data

Clinical data is a very popular data source to apply deep learning [212] for improved diagnosis [165] and risk prediction [250]. In particular, electronic health records can be a bountiful source of data to improve personalised medicine and improve quality of healthcare. Rajkomar *et al.* [191] demonstrate the potential of deep learning on EHR by developing an interpretable, scalable model that delivers insight on over 46 billion data points.

A large number of approaches for EHRs have been proposed. Miotto *et al.* [156] present an unsupervised approach, where each dimension of the feature vector corresponds to the number of times a patient is diagnosed with a specific disease. Nguyen *et al.* [163] propose deepr, which encodes a patients medical history as a sentence, as well as a record of the timeline, in order to use methods from natural language processing to understand the medical history. Kelly *et al.* [107] argue that these methods are not sufficient and identify three challenges for the use of machine learning on clinical data: bias identification, generalisation, and interpretability.

Gianfrancesco *et al.* [67] detail the multitude of potential biases that may be present in EHRs. In particular, individuals from vulnerable sub-populations are more likely to visit multiple institutions to receive care. However, this means their data is more likely to be spread across different organisations, which can make it more difficult for DNNs to fully utilise all the data. Furthermore, data points are often missing entirely [183]. Multiple imputation methods may be used to help

rectify the problem, but if missing values commonly occur within a particular sub-population, it is likely that the model exhibit bias and be less useful when applied to that sub-population.

As interpretability is essential to trust the output of a model, Choi *et al.* [38] propose RETAIN, a two-level neural attention model which processes EHRs to identify the importance and content of past visits, which then allows their model to predict the likelihood of heart failure. Similar to the works in this these, the attention mechanism is utilised to interpret the model, giving understanding to how it reaches a decision.

2.3.2 Medical Imaging

With the large success that DNNs have had in the computer vision field, most prominently using CNNs, they have also been applied extensively to analyse medical images. Piccialli *et al.* [184] split the use of DNNs on medical images into 6 main tasks: classification, detection, segmentation, reconstruction, registration, and dose estimation. Predictably DNN's success carries over into the medical domain [75]. DNNs already outperform clinicians at many tasks, such as diagnosing breast cancer from radiological scans [154] and recognising cardiac abnormalities from EGC data [195].

Data leakage is a huge concern with making sure that deep learning models do not cheat when making a decision. Winkler *et al.* [252] show that a CNN trained to classify skin cancer from skin lesions was classifying lesions based on the contours that have been drawn around cancerous skin lesions by clinicians. Kalmady *et al.* [103] identify that many fMRI data sets are corrupted with patients that are already taking anti-psychotic medication which is detectable within fMRI scans. Reported models then only need to detect signs that a patient is taking this medication rather than detecting the underlying disease.

To make models more interpretable, and help to ensure that data leakage is not occuring, a number of deep learning works have resorted to using attention mechanisms. Oktay *et al.* [167] build Attention U-net to segment organs on an abdominal computed tomography (CT) scan. This architecture could also be applied for disease diagnosis to segment regions of the image that the network considers important to arrive at a decision. Paschali *et al.* [179] equip a DNN with a multiple instance learning branch to output a logit heatmap of the neural activations.

2.3.3 Mental Health

DNNs within the field of mental health are usually concerned with two tasks: detection [209] and diagnosis [180]. Mental health issues, such as depression, often remain undetected unless the individual self-reports the struggles that they are undergoing. A naïve method to detect mental health issues is through designing and issuing mental health-specific forms for individuals to fill out.

This method is highly unreliable because individuals can easily mis-report their symptoms, a common occurrence due to the effects of having mental health problems, and the stigma attached to mental health issues in many cultures. Until recently, this unreliable method has been the primary way to detect mental health problems due to a lack of alternatives.

With the rise of social media, people have begun to share more information about themselves online. This information is usually unfiltered and more authentic to the true state of the individual. Eichstaedt *et al.* [56] were able to predict the diagnosis of depression from user language mined from Facebook posts. Benamara *et al.* [13] compose the eRisk2017 Reddit data set to detect the depression of users. Coppersmith *et al.* [41] go a step further and predict the likelihood of suicide risk. This demonstrates that it is also possible to use social media data to identify severity of depression, as well as detecting it.

When it comes to diagnosis, a vast majority of machine learning applications to diagnose mental health focus on the analysis of brain scans [211, 153]. Lee *et al.* [125] to identify the presence of insomnia. This is also the case for schizophrenia diagnosis [166, 177]. Qi and Tejedor [186] apply DNNs to magnetic resonance imaging (MRI) to identify subjects with schizophrenia and schizo-affective disorder. A layer-wise relevance propagation module was proposed by Yan *et al.* [265] to interpret a DNN that discriminated between schizophrenia patients with healthy controls. Nguyen *et al.* [162] use a slightly different method of detecting anomalies. They use generative models to inpaint image regions with what should be expected based on the distribution of the data. Any image that varies significantly from the inpainted image is judged to be an anomaly and therefore a potential risk. This represents a microcosm of a very popular field with many similar techniques used for varying applications. For a full survey on the use of deep learning to brain images, see [277].

2.4 Applications in Security

2.4.1 Face Verification

As with many other fields, DNNs have been the predominant machine learning method in recent years. The first major breakthrough came when DeepFace [226] attained the state of the art on the popular 'Labeled Faces in the Wild' (LFW) data set. Since then, deep learning techniques have dominated the field, with a multitude of methods even able to outperform humans [120].

For many classification tasks, the softmax loss is the dominant optimisation function. However, it is often not sufficient for verification and re-identification tasks because variations within the same class may be larger than differences between classes. The triplet loss can help to solve this problem, and has gained notoriety by further improving the state of the art [205, 178] for face recognition. Triplet loss research has typically focused on improving either the triplet mining algorithm or the loss function.

A major problem for face recognition systems is the lack of robustness to changes in pose. There are two main classes to handle this difficulty: *one-to-many augmentation* and *many-to-one normalisation* [243]. This challenge is exacerbated when there is also a change of appearance between images of the same individual, such as aging, growing of facial hair, or application of makeup.

One-to-many augmentation uses data augmentation and generation techniques such as GANs [73] to infer different poses from a singular face image. By training the system from such augmented data, the model can learn a relationship across pose, making it more robust to real cross-pose data [215, 282].

On the other hand, many-to-one normalisation aims to directly learn the mapping from non-frontal face images to a frontal face image. Zhou *et al.* [289] propose GridFace to learn this relationship directly with a CNN that is embedded with regularisation and rectification modules. Huang *et al.* [91] use a GAN with a local and a global pathway to synthesise a photorealistic frontal face image from an alternative view.

2.4.2 Person and Vehicle Re-identification

Historically, popular methods for person re-ID were typically comprised of two components: designing hand-crafted features and learning distance metrics [284, 295]. Most works focused on developing features invariant to variations in light, pose and viewpoint while using conventional distance metrics like the Mahalanobis distance [198], Bhattacharyya distance, and the l_1- and l_2-norms. Research has also been performed on a post-processing technique called re-ranking [286, 9].

Although similar to person re-ID, vehicle re-ID has received comparatively little attention. This is inconsistent with other computer vision tasks in the vehicle domain, like detection and classification, which have received increased attention in recent years. This lack of popularity can be attributed to the inferiority of large-scale vehicle re-ID data sets compared with their human re-ID counterparts. This is beginning to change as large-scale data sets, such as VeRi and VehicleID, have been released and have started to attract more research attention.

Research focus in re-identification has shifted towards deep learning methods, which are routinely used to obtain state-of-the-art results over a wide variety of challenges in computer vision and machine learning. Typically, two types of CNN model have been employed to solve the person re-ID task: the classification model that is used across a broad spectrum of computer vision problems [121] and, more commonly for re-ID, the Siamese model which takes multiple images as input, such as pairs [132, 187], triplets [205, 173], and quadruplets [33].

As there is typically more variance between viewpoints within vehicle re-ID compared to person re-ID (Figure 25), more creative methods have been proposed to obtain satisfactory results. Liu *et al.* [138] developed a two branch CNN to learn deep features and the distance metric simultaneously. Liu *et al.* [146] combined hand-crafted features and high-level attributes learned by a CNN with license-plate recognition and spatio-temporal information. Zhou *et al.* [292] trained a model on a toy car data set to in order to infer a multi-view vehicle representation from any input view. Due to the proficiency of deep learning at handling large-scale databases like the one we construct, we elect to utilise it in our experiments.

Recently, to improve robustness across data sets, there has also been focus on unsupervised re-ID by domain adaptation [207], because traditional supervised re-ID cannot generalise to additional

data sets. Fan *et al.* [58] develop a progressive unsupervised learning method that iterates between person clustering and CNN fine-tuning during training. Zhong *et al.* [288] explore three types of invariance that hinder the ability of the re-ID model to generalise to new domains: example invariance, camera invariance and neighbourhood invariance. Deng *et al.* [49] translate images to the target domain using CycleGAN [296] then enforce *domain-dissimilarity* between the translated image and other images in the data set. Ding *et al.* [51] use adaptive exploration to learn discriminative features in the target domain. Whereas unsupervised learning requires generalisation to unlabelled data, PVUD requires generalisation between data types.

Attention in Re-identification

A recent trend in person re-ID has been to design attention modules that can extract colour information from clothing [134, 266, 261, 285, 101], thereby ignoring the background information. These can be roughly categorised into *hard attention* and *soft attention*. HACNN, described in Section 2.2.2, demonstrates how soft attention and hard attention can be effective for re-identification, but also the potential that can be gained by using the two modules simultaneously.

Hard Attention: The most popular attention systems for person re-ID rely on hard attention, whereby regions are explicitly singled out as being more important. This can be done manually, e.g. part-based systems split the image into several parts to process separately, or automatically using external segmentation or pose extraction models.

The most success has been had with part-based methods. Yao *et al.* [268] introduce a part loss, i.e. a classification loss for each individual part. Fu *et al.* [59] propose horizontal pyramid matching, whereby the final feature representation of each image is obtained by concatenating the one-, two-, four-, and eight-part representations.

Other hard-attention mechanisms have been proposed for re-id. Wang *et al.* [246] estimate the pose of each images and extract skeletons to align features. Liu *et al.* [139] extract skeletons in a similar way and use these as a condition to generate images in a wide variety of poses. Wu *et al.* [258] extract a pose mask with a segmentation model to handle misalignment issues more comprehensively than by extracting a skeleton.

Soft Attention: Sun *et al.* [223] introduce a part-based convolutional baseline (PCB) with a refined

part-pooling mechanism. In most part-based systems, the parts remain static, however the refined part-pooling mechanism allows PCB to dynamically adjust the location of the parts to improve part-matching between images.

Jiang *et al.* [101] propose an attentional loss by comparing the spatial feature maps at different ResNet blocks and minimising the discrepancy between these. Chen *et al.* [34] use a self-attention mechanism to explore non-local relationships within re-ID images. He *et al.* [82] go a step further and propose transreid, whereby convolutional layers are replaced entirely by transformer layers that are driven by the self-attention mechanism.

Triplets

Compared to most classification problems, re-ID often contains many classes (individuals, vehicles, etc.) and few samples per class. This makes learning class-specific features difficult. To handle this problem, it is often beneficial to consider metric learning, usually in the form of the triplet loss [87] or centre loss [249]. The triplet loss in particular has seen extensive use for person [83, 36] and vehicle [115] re-id.

Triplets are generated by pairing query images with one image of the same identity and one with a different identity. Wang *et al.* [242] proposed to use the triplet loss function to learn image similarity. Cheng *et al.* [36] introduced an improved triplet loss function that decreases the distance of similar IDs and increases the distance of dissimilar IDs.

Building an effective triplet network is heavily reliant on the mining strategy. To challenge the framework to be able to handle tough cases, difficult triplets need to be mined, but choosing only the hardest triplets in the data set will result in a model that is not representative of the entire set of triplets. To strike the balance between finding difficult triplets while still generating a representative model, Hermans *et al.* [83] present *Batch Hard* mining, which selects only the hardest triplets across each batch selected during training. In a similar manner, Almazan *et al.* [4] select triplets that start off relatively easy but get more difficult as training progresses.

Chen *et al.* [33] add an additional term to the triplet loss to form a quadruplet loss. This term contains a second negative pair which helps to enlarge inter-class variations across the data set. Jiang *et al.* [101] demonstrate improved performance through adding a self-supervised attention loss to

the quadruplet loss. While performance is enhanced by these works, they all focus on improving the same aspects of the triplet loss. We instead tackle the under-researched feature representation and the distance function.Wu *et al.* [257] combine triplet loss with identification loss and centre loss. Tian *et al.* [232] mine more informative triplets via their re-weighting strategy.

Triplet-wise training has also been effectively applied for vehicle re-id. Zhang *et al.* [280] combined the triplet loss with a classification loss and also ensured negative samples in one triplet act as positive samples in another triplet. Bai *et al.* [10] fed groups of images into their triplet network to mitigate inter-class variance and propose a mean-valued triplet loss to enhance learning.

Mid-level Features

Yu *et al.* [270] concatenate features from earlier ResNet layers with the final layer representation for cross-domain image matching. However, although their approach works well when it uses the triplet loss for sketch-based image retrieval, their approach does not work well with the triplet loss for re-ID so they switch to a classification loss. Zhu *et al.* [298] also fuse mid-level features with final level ones as part of a two stream pose-based and part-based architecture. Zeng *et al.* [272] perform an extensive analysis on the performance of each layer to develop a hierarchical deep learning feature, which fuses features from several earlier layers. Although their method works well with their newly defined metric, their model is heavily engineered for person re-ID, thus would struggle to adapt for vehicle data. Lin *et al.* [136] align mid-level features to boost the performance of unsupervised re-id. This provides further evidence that mid-level features are an important tool for re-ID and not just specifically useful for supervised, person re-id.

2.4.3 Computer Vision on UAVs

A large body of research applying computer vision to imagery captured by UAVs has been developed, including visual segmentation [150], target tracking [135] and aerial re-ID [76, 231, 279]. However, the study of the same tasks where UAVs are the main object of interest has not been extensively investigated.

Most UAV-related computer vision research is focused on deep learning approaches for UAV detection and tracking [141, 43, 170]. In this context, some data sets have been created to investigate novel visual-based counter UAV systems. The Drone-vs-Bird Challenge data set [40] collects a

series of videos where UAVs usually appear small and can be easily confused with other objects, such as birds. Recently, the Anti-UAV data set [102] has been proposed to evaluate several tracking algorithms in both optical and infrared modalities. Pure computer vision tracking algorithms are exhaustively compared, but pre-trained models were not permitted in this challenge. This means there are significant improvements that could be made beyond the presented results. Isaac-Medina *et al.* [95] performed a comprehensive review of different detection and tracking frameworks on multiple UAV data sets.

Chapter 3

Fused Attention for Robust Interpretable Schizophrenia Diagnosis

DNNs cannot be used for real-world medical diagnosis without demonstrating interpretability and a robustness to data distribution shift. This chapter proposes a model which utilises attention to achieve both of these goals. The proposed model is application-agnostic, and can be used on any kind of binary tabular data. Stress tests are designed to test a range of models under data distribution shift to evaluate robustness and simulate their performance in a real-world setting. Parts of this chapter are published at a peer-review journal [175].

3.1 Introduction

According to the World Health Organisation, schizophrenia can cause a greater level of disability than any other physical or mental illness [256]. Diagnosis and treatment must be conducted as early as possible to improve outcomes, so that it does not reach this level of severity [79]. However, the symptoms of schizophrenia are similar to those caused by drug use, brain tumours, and other mental health problems like bipolar disorder. This makes schizophrenia notoriously difficult to diagnose, with clinicians going through an exhaustive process to rule out other potential causes. A minimum of six months of observation is required before a schizophrenia diagnosis can be

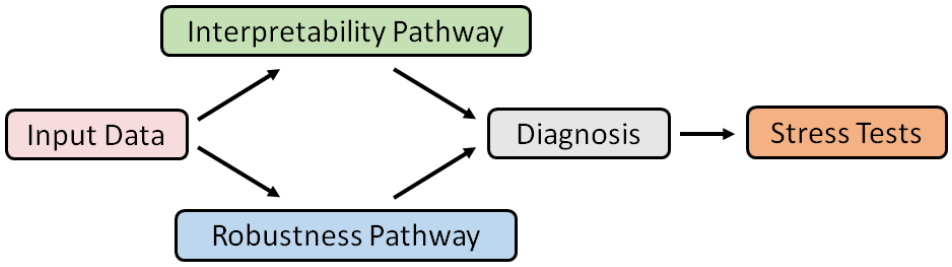


Figure 4: An overview of our proposed method. Data is processed via two pathways: one for robustness and one for interpretability. These pathways are complementary to one another, provide an insight into how the model arrives at a decision, and can generalise to new distributions.

provided, according to DSM-5 criteria [6]. Our model predicts the diagnosis after six months from one observation session with 98% accuracy.

Machine learning is increasingly being applied for medical diagnosis in the real world. Identifying problems at an earlier stage helps to manage them before they develop into a serious condition. Employing machine learning for disease diagnosis can also reduce a considerable number of required man-hours: machine learning systems can quickly process information and provide a recommendation. This would allow more patients to receive high-quality early-stage treatment.

To use machine learning to diagnose schizophrenia, we construct a data set that complies with current clinical assessment guidelines. Clinicians currently must assess patients according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [6] criteria A-F. We compose a data set of clinical observations of patients with a mental health condition. The collected attributes in the data set include attributes related to DSM-5 criterion A symptoms and attributes related to the symptoms of other mental health illnesses, to help the model to exclude these illnesses in the same way that psychiatrists must. Moreover, mental health clinics that use DSM-5 criteria will record psychiatric observational data similar to ours. Our model can therefore be applied to data that is already recorded, or after a clinical appointment, with minimal change to current processes required. These data set design choices have been made to demonstrate the ability of our proposed framework to function in the real world. Many machine learning models have shown good performance on laboratory-created data sets, but their real-world impact has been minimal [229]. This chapter aims to minimise the gap between research and practice.

One major drawback of DNNs is a lack of interpretability. A missed diagnosis could result in

death, whereas incorrectly diagnosing a patient with a condition could also result in complications (e.g. due to side-effects of needlessly prescribed medicine). For a clinician to have confidence in a recommendation from a deep learning system, they must be able to understand why the system has come to that recommendation.

We propose RobIn to improve the interpretability of the system and provide a high quality diagnosis in an efficient manner. It is important to consider *static* and *dynamic attention*. We wish to determine which features the network believes are important in general, and to understand how this changes based on specific values that features take, and combinations of those values. Our mechanism consists of a *squeeze and excitation* (SE) module [89] to compute the global feature importance, and a *self-attention* (SA) module [234] to decipher how features interact individually and with each other.

Squeeze and Excitation networks [89] compute very discriminative feature representations by learning which input features at each layer are most important. We posit that this fixed global feature importance allows the network to remain robust to data from new distributions. To show the effectiveness of our method, we design three stress tests at four different strength levels to alter the distribution of test data. Our stress tests consist of noise addition, data erasing and a combination of the two. We evaluate all compared models across these twelve robustness experiments to determine which would be suitable for deployment in a real-world clinical setting.

An overview of the chapter can be found in Figure 4. We provide the following contributions:

1. The collection of psychiatric evaluation data that enables machine learning training for automated diagnosis of schizophrenia.

2. A **Rob**ust, **In**terpretable (RobIn) deep network is developed to model the data to recommend a schizophrenia diagnosis. Our framework outperforms existing machine learning methods such as support vector machines, multi-layer perceptrons, and deep neural networks.

3. RobIn is made interpretable on two levels: the adaptation of channel attention to determine global feature importance, and self-attention for feature interactivity analysis.

4. A total of twelve stress tests are designed to evaluate model performance on a distribution that is not i.i.d. to the training data, to simulate model performance in the real world. Our

proposed model outperforms all other methods at remaining robust to perturbations.

We perform experiments on two settings: the standard 90/10 cross-validation protocol that is most commonly used with small data sets, and a 50/50 train/test split with 25 runs as a baseline with which to compare the robustness stress tests. We find that RobIn outperforms other methods on both of these test settings, and is also most robust to distribution shift proposed to test robustness.

The rest of the chapter is organised as follows. Related studies are outlined in Section 3.2. Section 3.3 provides more details on the data set. Section 3.4 outlines the attention mechanism helping clinicians to understand the outcomes of the DNN. Section 3.5 contains our experimental evaluation and stress tests. The chapter is concluded in Section 3.6.

3.2 Schizophrenia Background

3.2.1 Diagnosis of Schizophrenia

Schizophrenia has traditionally been diagnosed via specifications from the Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases (ICD), with the most recent versions being DSM-5 [6] and ICD-11 [108] respectively. These traditional specifications recommend a diagnosis based on the symptoms that a patient is exhibiting.

The Research Domain Criteria (RDoC) [44] is a modern approach to understand the underlying neurobiology associated with major mental health issues. Although a proportion of schizophrenia research has shifted from DSM-ICD to RDoC, we note that the RDoC website states "*RDoC is not meant to serve as a diagnostic guide, nor is it intended to replace current diagnostic systems.*" RDoC can be effective supplementary information but the necessary data can be costly to obtain and is therefore less accessible in developing countries, or for patients who must pay to obtain a brain scan but cannot afford to do so. For this reason, we design our diagnostic machine learning system using the traditional symptomatic approach, using DSM-5 symptoms.

Although DNN results within mental health seem promising, the real-world application of machine learning techniques on brain image data for schizophrenia diagnosis remains limited. There are several problems that are ubiquitous in the domain of machine learning for schizophrenia di-

DSM-5 Criteria	Attribute(s)	Label Information
Delusions	'Thought Content'	Persecutory/grandiose delusions
Hallucinations	'Thought Perception'	Auditory/visual/tactile hallucination
Disorganised speech	'Speech'	Normal, mute, incoherent
Disorganised behaviour	'Mood', 'Attention'	Good, poor, neutral
Negative symptoms	'Mental State Examination'	Kempt, unkempt, restless

Table 2: Data Set Attributes Relating to DSM-5 Criteria

agnosis:

- There is a large *training-application gap* [238], i.e. the trained models accurately describe the source data well, but the models are not shown to be transferable to the real world [229, 228], nor are there explanations of how their models could be used in a clinical setting. In contrast, our model is designed to be directly applicable to data that is routinely collected by schizophrenia diagnosticians. Furthermore, we design stress tests to demonstrate that our model would be robust to data from different sources.
- Data sets consist of only schizophrenic or healthy individuals but do not contain other unhealthy samples with bipolar disorder or drug overdoses [253]. In practice, a model trained on this data is only able to determine if a patient is healthy or unhealthy, but cannot diagnose schizophrenia. Our data contains patients with multiple causes for the observed symptoms ('bipolar affection disorder', 'complex partial seizure', 'drug related disorder', etc.) resulting in a more challenging and realistic data set.
- Brain image scans in schizophrenia data sets often contain patients who have already started taking anti-psychotic drugs. These drugs have a distinguishable effect on the brain [126], reducing the reliability of the data sets and the results.

3.3 Data

In this section, we motivate the collected data set, give an overview of its characteristics, and detail how it follows current diagnostic guidelines. We also discuss the selected subjects and the pre-processing performed on the completed data set.

3.3.1 Data Motivation

Artificial intelligence for schizophrenia diagnosis has received criticism for not being transferable to the real-world [229]. One reason for this is a disconnect between machine learning practitioners and schizophrenia diagnosticians. Machine learning suffers from the ubiquitous practice of competing for the highest scores on benchmark data sets, often giving less regard to real-world applicability.

- The collected data is standardised, based on symptoms from DSM-5 [6]. Attributes are taken directly from case files of patients, so the data will be very similar to that which is already collected in mental health clinics. Moreover, this data does not depend on advanced technology (the EEG attribute could easily be discarded), meaning that models developed from this data are more accessible to developing countries with limited resources. Note that our model requires minimal computational resources.

- Studies have shown a link between childhood trauma and schizophrenia in both high income [152] and developing [151] countries. Schizophrenia can be particularly dangerous in poorer regions of Africa, where the likelihood of trauma is greater, the availability of treatment is lower, and there is stigma attached to mental illness [23]. Most recent works that apply AI to diagnose schizophrenia use images from EEG or fMRI scans [253]. However, these scans are often unavailable in poorer regions and there is no evidence to suggest that models developed in high-income countries can generalise to data from developing countries, either because of hardware differences or genetic differences [77].

- The data allows us to predict a clinician's six-month diagnosis from one observational session. This greatly advances the speed that a potential schizophrenia diagnosis could be identified to help inform the clinician's future care of the patient.

- Most research on schizophrenia diagnosis only consider two classes: people with schizophrenia and people without schizophrenia. Mental health clinics rarely face this challenge as patients who do not have schizophrenia will usually have a similar mental health condition. The main difficulty of diagnosis is differentiating schizophrenia from other conditions rather than from healthy samples. Our data set contains samples with a broad variety of health issues to more accurately estimate the difficulty that clinicians face.

3.3.2 Data Acquisition

The data was collected from a total of 151 subjects, between 2013 and 2018 inclusive, from the Lagos University Teaching Hospital, Lagos, Nigeria. Our data set consists of psychiatrists' direct observations of their patients; consequently, our model gets the same information that psychiatrists have used to determine their diagnosis.

97 samples are positively diagnosed Schizophrenia patients, and the remaining 54 are other patients with different afflictions. It is important to note that the negative samples are patients suffering from other related illnesses; they are not healthy individuals as in most schizophrenia diagnosis data sets used for machine learning research. This makes schizophrenia diagnosis more difficult, but is a significantly more realistic challenge.

The medical records of the patients were obtained from their case files, after obtaining ethical approval from the Lagos University Teaching Hospital Health Ethics Committee. Each subject signed a patient consent form prior to their mental health assessments; the form permits the use of medical information for the purposes of research and education. All records were obtained anonymously in line with ethical approval guidelines, to retain the privacy of the subjects.

The full details of the collected attributes can be found in the appendix. Note that although we collect attributes of 'Age', 'Sex', 'Occupation History', and 'Marital Status', we do not use them in any of the reported experiments. RobIn is trained on observed features related to mental health, rather than the general characteristics of the individuals; that is, we train with features that represent an underlying cause of schizophrenia. In fact, we found that when we included the four aforementioned features, 'Marital Status' was often given a large amount of attention. We discovered that the model was exploiting a bias in the data set as many schizophrenic patients were divorced. This gives further credence to the importance of utilising interpretable networks to help uncover biases that are missed in the original data inspection.

With any human-labelled data set, there is a small chance of label error. However, if ML methods can consistently infer a diagnosis from input features, the overall data set can be seen as reliable enough to generalise to new data.

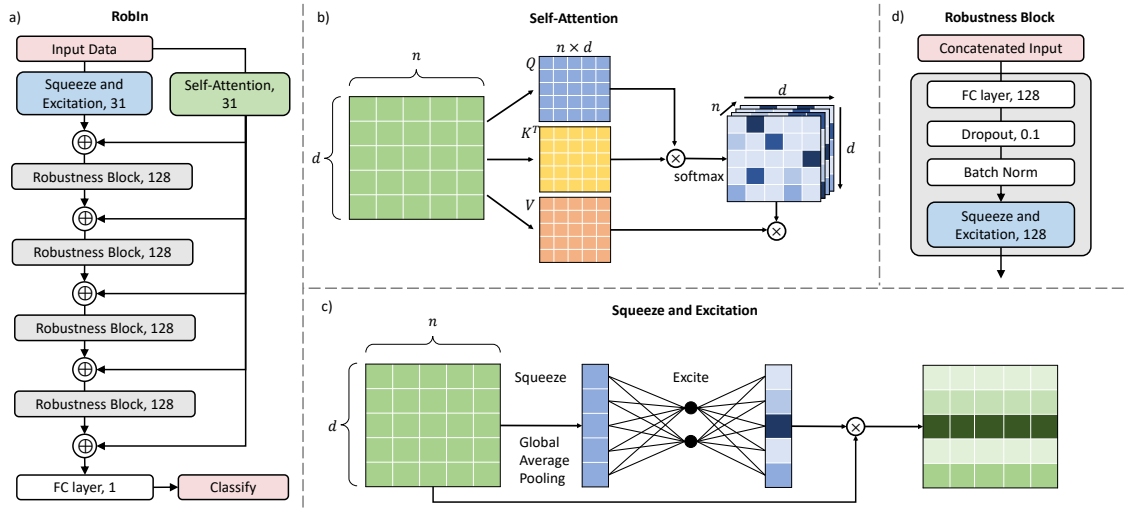


Figure 5: a) An overview of the entire network - input data goes down the robustness stream and the interpretability stream; b) self-attention mechanism: the input representation is converted into a key, query and value matrix, the cosine distance between each query and each key is found via a matrix multiplication with a higher activation signalling higher alignment between query and key; c) squeeze and excitation: each attribute $i = 1, \ldots, d$ is *squeezed* down to a representative number, a miniature neural network *excites* the squeezed information to evaluate how important each attribute is, then the initial data is multiplied by the importance scores; d) the robustness block we propose in this chapter.

3.3.3 Data Structure

The data set has 31 raw attributes that are used for learning, and a final diagnosis. The attributes collected reflect the criteria used by psychiatrists for Schizophrenia diagnosis, as officially recommended in DSM-5 documentations [6]. DSM-5 outlines diagnostic criteria A-F, which must be fulfilled for a patient to be diagnosed with schizophrenia.

Criterion A concerns the characteristic symptoms of the condition, and is where DNNs can be effective. Table 2 presents: the DSM-5 diagnostic criterion A, the attributes selected to reflect criterion A symptoms, and further information on the potential labels of these attributes.

We also collect attributes that commonly occur in other mental health conditions, including depression, bipolar disorder, and schizo-effective disorder. We do this because DSM-5 Criteria D-F are exclusions of substance misuse and other disorders such as such as schizo-affective disorder and autism spectrum disorder.

3.3.4 Data Pre-processing

Data pre-processing is an essential step in most machine learning pipelines to improve consistency of the algorithm. The collected data set contains a mixture of numeric and text data, which is difficult for a deep learning model to process. Categorical data is converted directly into numeric representations via *label encoding*: for every attribute in the data set, each unique value is assigned a representative number.

Of the selected 31 raw attributes, the majority of samples are missing at least one value. This scenario is common in medical data and is therefore well-studied because, for example, different clinicians record different data. To address this, we set all missing values to -1 before performing label-encoding. This method is easiest to apply to the real world where the majority of samples will also have missing values, rather than having to use imputation methods from external data sources that may not match up to the data being processed.

Experimental results showed that setting missing values to -1 resulted in better performance than attempting single-variate or multi-variate imputation techniques. This is likely because this format explicitly informs the system about data absence, allowing the attention mechanism to learn to assign less emphasis.

3.4 Methodology

Despite the excellent performance of deep networks, they suffer from not being interpretable. Features at each layer are generated automatically by the network, so it is not clear to humans what these mean. To use machine learning for health applications, it is important to have a degree of understanding of how the algorithm arrives at a conclusion.

We develop a robust, interpretable framework based on *Squeeze and Excitation* and *Self-Attention* to provide insight on the network's decision process. From the input layer, all input features get processed by a robustness pathway, governed by squeeze and excitation, and an interpretability pathway, indicated via self-attention. After each robustness block, we fuse the output vector with the self-attention, to ensure the interpretable mechanism is influential at all stages of the network. Fusing these attention modules allows the network to discover both *global importance* through squeeze and excitation, and *feature interactivity* through self-attention. We are therefore able to

interpret the network on multiple levels. Our overall framework can be found in Figure 5.

3.4.1 Channel Attention

The channel attention mechanism described in section 2.2.1 is designed for convolutional neural networks. We re-design channel attention to be used for regular neural networks taking on non-sequential data, in order to learn the importance of the input channels to help humans understand why a machine has come to a certain conclusion. This process is described here.

The information from each attribute (rather than from each channel) is *squeezed* into a representative descriptor via Global Average Pooling. Given the j^{th} attribute $a_j = [a_{1j}, \ldots, a_{nj}]$, where n is the number of samples, we form the attribute descriptor, c, via:

$$c_j = \text{squeeze}(a) := \frac{1}{n} \sum_{i=1}^{n} a_{ij}. \tag{3.1}$$

These feature descriptors form a vector $\mathbf{z} = [c_1, \ldots, c_d]$ where d is the dimensionality of each sample.

We then *excite* the network to determine which attributes are most important. $\mathbf{z}$ is passed through a fully connected layer, a ReLU, another fully connected layer and finally a sigmoid activation.

Formally, this excitation is written as:

$$\mathbf{s} = \text{excite}(\mathbf{z}) := \sigma\left(\mathbf{W_2}\delta(\mathbf{W_1}\mathbf{z})\right), \tag{3.2}$$

where $\mathbf{W_1}, \mathbf{W_2}$ are the parameters of the fully connected layers, δ is the ReLU function and σ is the sigmoid activation function.

3.4.2 Self-Attention

We utilise self-attention as described in Section 2.2.3. To interpret the model, self-attention provides more information than squeeze and excitation. Rather than showing just the raw importance of each feature, self-attention shows us the importance of every feature j to each feature i. For example, squeeze and excitation may suggest that a feature i is more important than feature j,

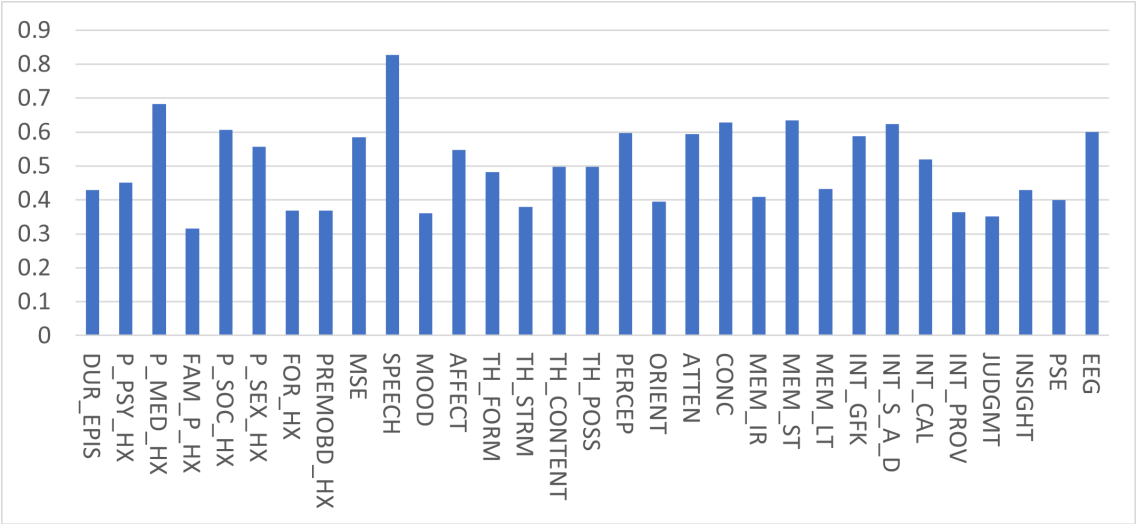


Figure 6: Global Feature Importance of the RobIn: as expected, no features are entirely discarded, but certain features such as speech, past medical history and concentration are thought to be important.

however it may be the case that j is important if i takes on a particular value (or is missing). Self-attention can handle this more complex circumstance.

3.4.3 Robust Interpretable Network

The architecture is designed with three principles in mind: performance, interpretability, and robustness. We marry the two attention mechanisms described, with squeeze and excitation contributing to performance and robustness, and self-attention mostly benefiting performance and interpretability.

In particular, squeeze and excitation has been widely shown to provide more informative features when used at each successive layer. We follow this standard design for the main branch of our network. Self-attention on the other hand is often used sparingly due to its computational cost being $\mathcal{O}(d^2)$ where d is the number of features. This occurs due to the matrix multiplication in Equation 2.21. Another issue is that we wish to apply self-attention directly to the data to be able to interpret the final result; however, self-attention applied in early layers often performs an averaging effect in early layers, and only starts to highlight important feature interactions at later layers [193]. This problem is exacerbated when multiple self-attention modules are used throughout the network.

To solve these problems, we include one self-attention mechanism in our framework, and con-

Table 3: Comparison with Baseline Machine Learning Techniques for the Diagnosis of Schizophrenia with 90/10 Cross Validation

Method	Acc.	F1-score	AuC	Precision	Sensitivity	Specificity
MLP	91.33 ± 3.63	92.84 ± 3.03	91.11 ± 3.85	89.83 ± 5.27	96.87 ± 2.60	85.36 ± 7.59
DNN	92.00 ± 2.70	93.19 ± 2.46	92.49 ± 2.85	93.19 ± 4.04	93.67 ± 2.86	91.31 ± 4.94
SVM	94.67 ± 3.14	95.55 ± 2.71	95.62 ± 2.52	99.00 ± 1.62	92.90 ± 5.00	98.33 ± 2.71
Tree	96.00 ± 3.31	96.84 ± 2.65	97.15 ± 2.38	$\mathbf{100.00} \pm 0.00$	94.29 ± 4.77	$\mathbf{100.00} \pm 0.00$
SENN	96.67 ± 2.91	97.39 ± 2.25	97.22 ± 2.51	98.89 ± 1.80	96.11 ± 3.31	98.33 ± 2.71
SANN	96.67 ± 3.33	97.12 ± 2.92	96.51 ± 3.32	96.64 ± 2.79	97.78 ± 3.61	95.24 ± 7.30
RobIn	$\mathbf{98.00} \pm 2.30$	$\mathbf{98.56} \pm 1.62$	$\mathbf{98.33} \pm 3.33$	99.00 ± 1.62	$\mathbf{98.33} \pm 2.71$	98.33 ± 5.00

catenate its output with the feature representation obtained from each layer. This means that the self-attention mechanism acts as a direct path from the input data set to the output classification. Therefore the weights of the self-attention mechanism are partially backpropagated directly from the classification layer, rather than the indirect path that is usually seen. This is important for interpretability because the attention heatmaps have to be powerful enough to contribute to the final classification, otherwise performance would heavily suffer. We also find that this self-attention module returns more descriptive, polarised attention heatmaps compared with a traditional implementation, so clinicians can attain more insight from studying them. To help reduce impact of overfitting as the data set is small, dropout is included, whereby a different portion of the parameters are dropped during each pass through the network. In experiments, we found a dropout level of 0.1 to perform well. Our full RobIn architecture is shown in Figure 5.

3.5 Evaluation

3.5.1 Evaluation Protocol

We consider two settings to test machine learning models on our schizophrenia data set:

1. 10-fold cross-validation,
2. A 50/50 train/test split repeated 25 times, to test the ability of all models with less training data, and to have more test data to perturb with the designed stress tests to more reliably evaluate the performance of all models under a different distribution.

We performed hyperparameter tuning on all models with manual search. In all cases, we report the strongest performance that we could obtain. We also report confidence intervals at the 95% significance level. For fairness, all experiments are performed with the same random seed. The implementation of all frameworks was performed in PyTorch on an NVIDIA Geforce GTX

Table 4: Comparison with Baseline Machine Learning Techniques for the Diagnosis of Schizophrenia on a 50/50 Train/Test Split (25 runs)

Method	Acc.	F1-score	AuC	Precision	Sensitivity	Specificity
SVM	83.57 ± 1.47	86.67 ± 1.28	83.38 ± 1.44	89.27 ± 1.76	84.78 ± 2.46	81.97 ± 2.96
MLP	84.43 ± 1.80	88.24 ± 1.35	82.14 ± 2.18	85.04 ± 2.49	$\mathbf{92.22} \pm 1.37$	72.07 ± 4.46
DNN	84.69 ± 2.30	88.37 ± 2.30	82.53 ± 1.85	85.34 ± 2.30	92.12 ± 1.51	72.93 ± 3.90
Tree	85.49 ± 1.85	88.15 ± 1.68	85.33 ± 1.92	90.29 ± 2.17	86.71 ± 2.62	$\mathbf{83.95} \pm 3.51$
SENN	85.97 ± 1.69	$\mathbf{89.68} \pm 2.32$	85.45 ± 1.99	89.68 ± 2.32	88.47 ± 1.88	82.43 ± 3.94
SANN	86.40 ± 1.82	89.36 ± 1.47	85.30 ± 2.35	88.66 ± 1.47	90.47 ± 1.60	80.14 ± 4.02
RobIn	$\mathbf{86.45} \pm 1.46$	89.19 ± 1.19	$\mathbf{86.08} \pm 1.75$	$\mathbf{90.51} \pm 2.17$	88.42 ± 1.83	83.73 ± 3.77

2080ti.

We propose three attention-based models to evaluate on these challenging tasks: our **Rob**ust **In**terpretable Deep Network, **RobIn**; a squeeze and excitation network with four-layers, **SENN**; and a self-attention network containing four consecutive self-attention layers (i.e. no fully-connected layers), **SANN**. These models are compared with standard approaches: a decision tree [259], **tree**, which attempts to discover a tree-like structure with decision points at each node and possible consequences at each branch to find a relationship between data and output; a support vector machine [42], **SVM**, which tries to find the hyperplane that maximises the margin between classes; a multi-layer perceptron [80], **MLP**, which is a neural network with one layer; and a deep neural network consisting of four layers, **DNN**. Note that we can consider the comparison between RobIn, SENN, DNN, and MLP an ablation study where we remove the self-attention mechanism, the squeeze and excitation layers and the additional hidden layers respectively.

We report the metrics: 'Accuracy', 'Precision', 'Sensitivity', 'Specificity', 'F1 Score' and 'AuC', where

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{3.3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{3.4}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{3.5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \tag{3.6}$$

are calculated via true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). From these measures we obtain the F1 Score and AuC:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{3.7}$$

The AuC is the area under a receiver operating characteristic (ROC) curve, plot on the unit square with axes sensitivity and (1-specificity). An AuC of 0.5 indicates random guessing.

3.5.2 Comparison with Baselines

The effectiveness of the compared models on the 10-fold cross-validation setting is evaluated in Table 3. RobIn attains the best results across the most descriptive measures: accuracy, F1-score and AuC. RobIn outperforms MLP and DNN to a statistically significant level on all three measures, and SVM on Accuracy and F1-score. Given that the decision tree also obtains a good performance, it is very difficult to get a statistically significant result. Nonetheless, we get a statistically significant improvement on F1-score at a 95% significance level, and get p-values of 0.086 and 0.121 for accuracy and AuC, respectively.

The effectiveness of the compared models over 25 runs on the 50/50 split is evaluated in Table 4. These results are shown as a pre-cursor to the robustness tests in Section 3.5.4. With a much smaller training set, we expect the traditional models to outperform the deep models. However, with the same random seed, RobIn outperforms all other models on Accuracy and AuC. SENN achieves a higher F1 score than RobIn; however, it does so with a very high variance, which indicates that RobIn is more consistent. In Section 3.5.4, we see that RobIn is indeed more robust and consistent as it is better at handling data pertubations. Although our results are not statistically significantly better than the decision tree, Section 3.5.4 shows that the scores from the decision tree and SANN are unreliable because they over-predict a positive diagnosis because the data set has more positive samples than negative i.e. they are exploiting bias in the data set to make a prediction rather than finding real statistical correlations from the data itself. This can already be seen Table 4 from the surprisingly high sensitivity of the MLP and DNN compared to their weak accuracy, but will become even more clear in Section 3.5.4. This means that the performance of these methods has been artificially inflated. Although they are capable of predicting from an i.i.d. test set, they cannot be trusted in a clinical setting.

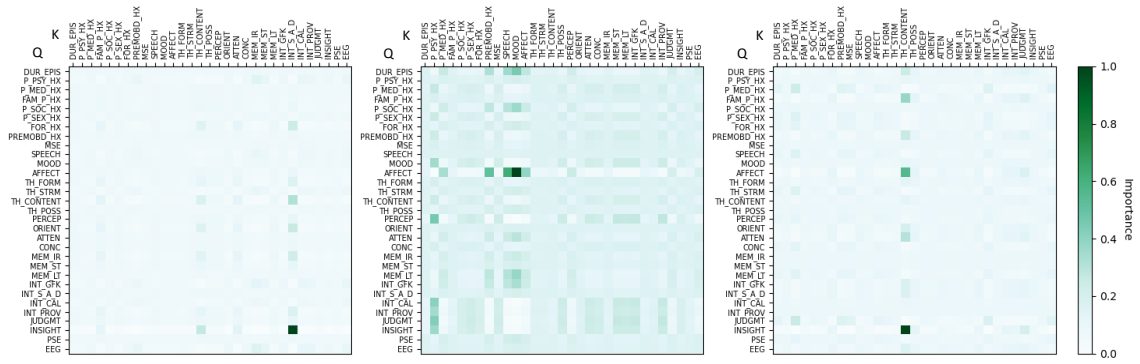


Figure 7: Heatmaps generated by the self-attention module. Darker squares indicate higher importance assigned by the queries (rows) to the keys (columns).

3.5.3 Interpretability

Global Attribute Importance

An example of the global attribute importance is shown in Figure 6. At this stage of the network, the squeeze and excitation mechanism needs to consider all attributes, and it therefore does not give any feature a particularly low score. However, we can see that the model determines speech ability to be very important in finding a diagnosis, while other attributes like past medical history, concentration, short-term memory, and EEG scan are also deemed important.

These scores remain identical for test data that is considered. This allows our system to remain robust to different stress tests that we conduct, which indicates that this global attribute importance is essential for the model to generalise to real-world data. As well as making the model more robust to data from different sources, this global attribute map also allows us to see the inner-workings of the model, therefore making the model much more interpretable compared to standard DNNs, consistent with Definition 1.2.1.

Attribute Interactivity

In Figure 7, we see examples of the attention activations of the self-attention mechanism. These heatmaps visualise the features that are deemed important to arrive at a diagnosis. In particular, we can observe that a model often finds that the insight of a patient is a useful query. Referring back to Definition 1.2.1, the internal mechanics of the model, visualised with these feature interaction maps, illustrate to human observers how the decision was made, indicating that the model is interpretable.

Note that because of the design of RobIn, more polarised activations are obtained because we connect the first layer with the final layer. In self-attention frameworks, the first layer activations are usually averaged while final layer activations are more polarised. However, this is not beneficial for interpreting feature importance because by the time polarised heatmaps are obtained, features have already undergone several layers of abstraction.

3.5.4 Robustness

DNNs are difficult to implement in the real world because, even though they perform well on unseen samples from the same data set, they struggle to generalise to unseen samples from a different source [45]. This is because the test set is independent and identically distributed (iid) to training set, but real-world data rarely has an identical distribution to small-sample collected data sets. We design three stress tests to randomly alter the distribution of the test set at varying strengths to evaluate the robustness of all models:

1. **Noise**: We sample noise $X \sim \mathcal{N}(\mu = 0, \sigma^2)$ and add this noise distribution to the test data to simulate different ways that a clinician might record a patient observation
2. **Data Erasing**: We randomly eliminate a fraction $\frac{1}{m}$ of the test data to simulate different ways that missing values could impact performance
3. **Noise + Data Erasing**: We combine 1) and 2) to provide a challenging, realistic test of model robustness

All models are evaluated on the original 50/50 train/test split with the same random seed, with the test set altered via the aforementioned perturbations. For each stress-test, we evaluate σ^2, $\frac{1}{m}$ at values $\{\frac{1}{10}, \frac{1}{5}, \frac{1}{3}, \frac{1}{2}\}$. Note that all perturbations are applied after the initial training data has been normalised between 0 and 1.

The performances under stress tests are presented in Figures 8 - 10. We found that accuracy and F1-score were poor indicators of model performance under these stress-tests because the data set is skewed with more positive samples than negative ones. As data became more noisy, models would often over-predict a positive diagnosis to obtain the highest accuracy. This can be seen from the specificity graphs, where the decision tree and the neural networks without squeeze and excitation modules all drop below 0.35. This indicates that models have learnt to make decisions

based on the data set characteristics, rather than the features provided. We report the accuracy and specificity to present this phenomenon, but base our robustness evaluation on the AuC - the only reliable evaluation metric because it is a function of specificity.

The figures show that the decision tree and SANN are not deployable because they show a tendency to over-predict a positive diagnosis. Those models are not suitable for real-world deployment because, counter to our data set, the proportion of people who actually have schizophrenia is very low.

Across all three stress tests, RobIn and SENN suffer least from perturbations of the test data, indicating that they are most suitable for deployment in the real world. This gives credence to our claim that the squeeze and excitation mechanism improves model robustness.

Additional Noise

As seen in Figure 8, when noise $X \sim \mathcal{N}(\mu = 0, 1/2)$ is added, the decision tree does barely better than random guessing with an AuC of 0.555, whereas RobIn maintains an AuC of 0.643 which is very impressive given the severity of the perturbation. Moreover, when $X \sim \mathcal{N}(\mu = 0, 1/5)$ perturbs the data, the AuC of RobIn falls only by 0.088, in contrast to the 0.186 drop by the decision tree. In the realistic scenario when $\sigma^2 = \frac{1}{10}$, all methods using attention see very little performance loss, whereas again the decision tree performs worst with a loss of 0.099, more than RobIn suffers with σ^2 is twice as large.

Removed Values

In Figure 9, the removal of values appears to be a much more difficult task, as shown by the sharp drop in AuC even when only $\frac{1}{10}$ values are removed. RobIn and SENN maintain the highest AuCs throughout the stress levels. All models struggle with half of the values removed and are close to random guessing. At the $\frac{1}{3}$ stress level, RobIn has 0.011 greater AuC than the next best, SENN. It is worth noting that the specificity of the SVM remains fairly constant as more values are removed; however, it has by far the lowest accuracy at all stress levels and only consistently outperforms MLP on AuC, despite the high specificity. It over-predicts a negative diagnosis, resulting in a high number of true negatives and a low number of false positives.

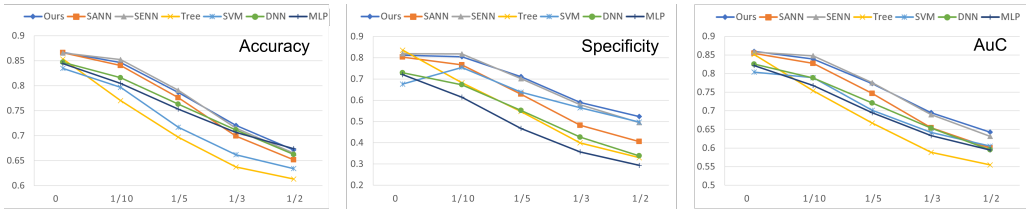


Figure 8: Robustness comparisons with the addition of noise, $X \sim \mathcal{N}(\mu = 0, \sigma^2)$, where the x-axis is σ^2.

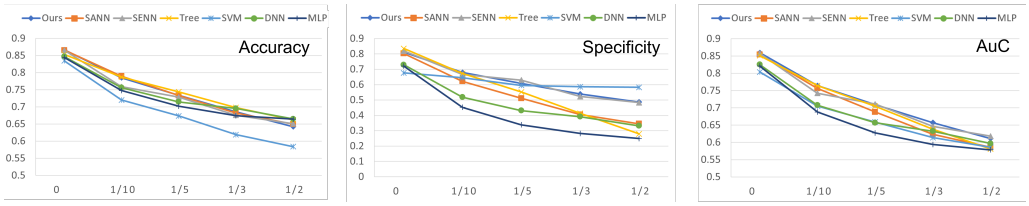


Figure 9: Robustness comparisons with the removal of data points where the x-axis signifies the fraction of values that were removed from the test data.

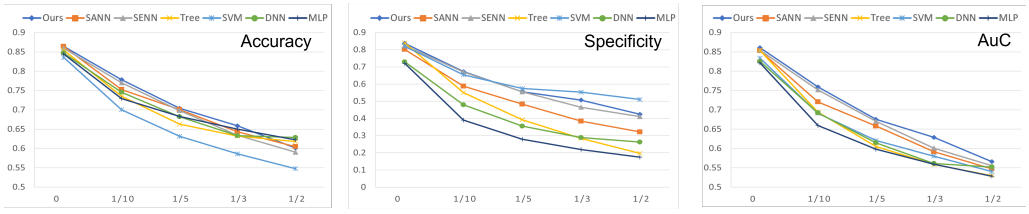


Figure 10: Robustness comparisons with the addition of noise and the removal of data points. The x-axis signifies the variance of the normal curve from which the noise was sampled and then added to the test data, and the fraction of values that were removed from the test data.

Combination

The combination of the previous two stress tests in Figure 10 is the most realistic test, where noise simulates difference of opinion between clinicians and the data point removal simulates values that are often missing. RobIn outperforms all non-attention based methods on AuC at $\frac{1}{10}$, $\frac{1}{5}$, and $\frac{1}{3}$ stress levels, with a 95% confidence interval. Furthermore, in the $\frac{1}{3}$ level, RobIn attains a 0.629 ± 0.0257 AuC, a statistically significant improvement on the next best method, SENN, which attains 0.601 AuC. RobIn even maintains the highest accuracy until the $\frac{1}{3}$ level, despite the propensity of other methods to cheat. It is clear that RobIn is the most robust method and most suitable for real-world deployment.

3.6 Conclusion

Machine learning has shown a tremendous amount of promise in detecting and diagnosing mental health conditions. However, machine learning practitioners have struggled to bridge the gap between theory and application for the task of schizophrenia diagnosis. To combat this, we have collected a data set that complies with current clinical guidelines outlined by the most recent edition of DSM-5. This data will be of a similar format to records that are already kept by clinicians for patient monitoring, so our model could be directly applied to electronic health records that are already available. Furthermore, unlike many commonly used data sets that only use healthy controls, ours differentiates between schizophrenic patients and those with other illnesses, making the task more difficult, but also more realistic.

We have performed extensive studies to show that a wide range of machine learning models can be applied to successfully classify whether a patient will be diagnosed with schizophrenia after six months from one session with the patient. We have developed a robust, interpretable network that outperforms these other machine learning methods. Our network contains a squeeze and excitation mechanism that works on a global scale to give an overview of feature importance and a self-attention mechanism that works on a local level to assess feature interactivity.

To help reduce the training-application gap, we have also comprehensively tested all methods with stress tests to evaluate their potential to generalise to new data sources. These robustness tests are essential to give some indication of which methods could be selected as a candidate to

be used in the real world. Our proposed method is the most robust of all compared models on all three stress tests. The evidence suggests that this is because of the static weights learnt by the squeeze and excitation mechanism. Therefore, we conclude that self-attention contributes to model interpretability and squeeze and excitation contributes to model robustness, whilst both contribute to overall performance.

Chapter 4

Multi-scale Attention for Robust Makeup Style Transfer

This thesis aims to demonstrate that attention mechanisms are an effective tool to improve deep network interpretability and robustness, to motivate researchers to incorporate them into their models, which in turn makes them more likely to be utilised in a real-world setting. To do so, attention mechanisms need to be evaluated on a broad spectrum of applications. Therefore, this chapter will study incorporating attention mechanisms into a very different model, with a very different application, compared to the previous chapter. The degree of difference from this chapter to the last helps to verify attention mechanisms can generally be incorporated into DNNs, rather than only being useful for a specific subset of applications.

Specifically, attention mechanisms are utilised to improve robustness for image-to-image translation, when the data trained is low-quality. The specific application studied is Makeup Style Transfer, but the proposed framework is application-agnostic. This chapter predominantly focuses on the robustness aspect of attention mechanisms, showing their ability to to *generate* realistic images when data is noisy. This is particular impressive given that DNNs without attention mechanisms struggle to even classify noisy face images [63]. Interpretability of attention mechanisms is less of a focus on this chapter, because image generation already obeys some properties of interpretability, in that a human can visually inspect where style has been transferred to understand what the model deems important. Parts of this chapter were published in a peer-reviewed

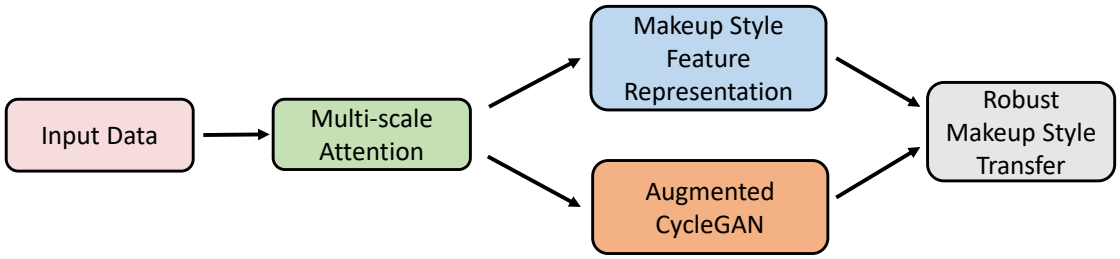


Figure 11: An overview of the use of attention to improve makeup style transfer robustness explored in this chapter.

conference [171].

4.1 Introduction

Current state-of-the-art makeup style transfer methods [28, 131, 30, 276] display a common trend when applied to low quality images: only lipstick colour is consistently transferred. Eyeliner and mascara occasionally make the transition. Foundation, eye shadow, blusher, fake tan, concealer, powder, and contours are largely disregarded. Bags under the eyes, that want to be concealed, are ignored. For real-world applications, such as makeup-invariant face verification [278] or beautification [130], this does not suffice. We postulate that the fault lies with the hard attention that current makeup transfer methods use to handle the difficulty of the task.

Convolutional neural networks struggle to generalise to different data sets [64]. This also extends to GANs. For example, PULSE [155] applied to external data, converted a downsampled image of Barack Obama into a white man. To minimise generalisation error and to help to defend against model bias, it is beneficial for the algorithm to be able to train on a large variety of data, whereas current state-of-the-art makeup style transfer methods can only train on high-quality lab data sets. In practical applications of makeup transfer, low-resolution faces are prominent, because faces often take up a small proportion of an image. After cropping, they appear at a low resolution. As this scenario will frequently occur, it is important that models can handle it. Because we cannot reliably depend on models to generalise to this low-resolution setting, we design our framework to be able to train directly on low-resolution data.

A starting point for makeup style transfer might be to use off-the-shelf image-to-image translation state-of-the-art, such as CycleGAN [296]. However, this performs poorly because the two domains are highly overlapping; both domains comprise of face images, with the greatest difference usually appearing on the lips and around the eyes. It is challenging to describe a makeup style since it consists of multiple non-requisite components. A face only wearing lipstick and the same face only wearing eyeliner should both belong in the makeup domain. This is difficult for standard models to gauge without being directly pointed to.

Many works use CycleGAN as a starting point for makeup style transfer, because it performs so strongly for most domain translation tasks. However, it needs to be heavily engineered to attain reasonable performance on makeup style transfer specifically. Most works utilise a part-by-part solution to apply a style from one face image to another. Current works develop a hard attention module where face parts likely to consist of makeup (eyes, lips and cheeks) are segmented and optimised upon separately [28, 131, 30]. However, to segment the image, these methods are dependent on the Face Parsing Algorithm (FPA) [218]. We demonstrate that FPA is limited in handling low-resolution face images due to the lack of detailed features to identify face parts. As a consequence, state-of-the-art makeup style transfer algorithms cannot be applied successfully to lower-resolution faces.

We discard the hard attention and develop an analogous soft attention module to transfer makeup style in a holistic manner, rather than piece by piece. To tackle the problem of identifying salient parts of low resolution images without FPA, a novel weighted multi-scale spatial attention module is proposed. The module consists of spatial attention with multiple convolutional kernels. These convolutional layers determine salient areas of the image at different scales, which are then processed by an intermediate channel attention module, which determines the importance of different scales and assigns respective weights. This attention module serves two main purposes. Firstly, attention at different scales can focus on transferring different aspects of a makeup style: smaller scales capture fine-grained information such as fake eyelashes and lipstick, whereas larger scales focus on transferring foundation and fake tan that appear across the entire face. Secondly, faces can appear at any size in an image, so it is important to be able to effectively handle different resolutions. Our attention module can dynamically adjust which convolutional kernels are assigned high weights, and therefore extract more information from lower quality images. This results

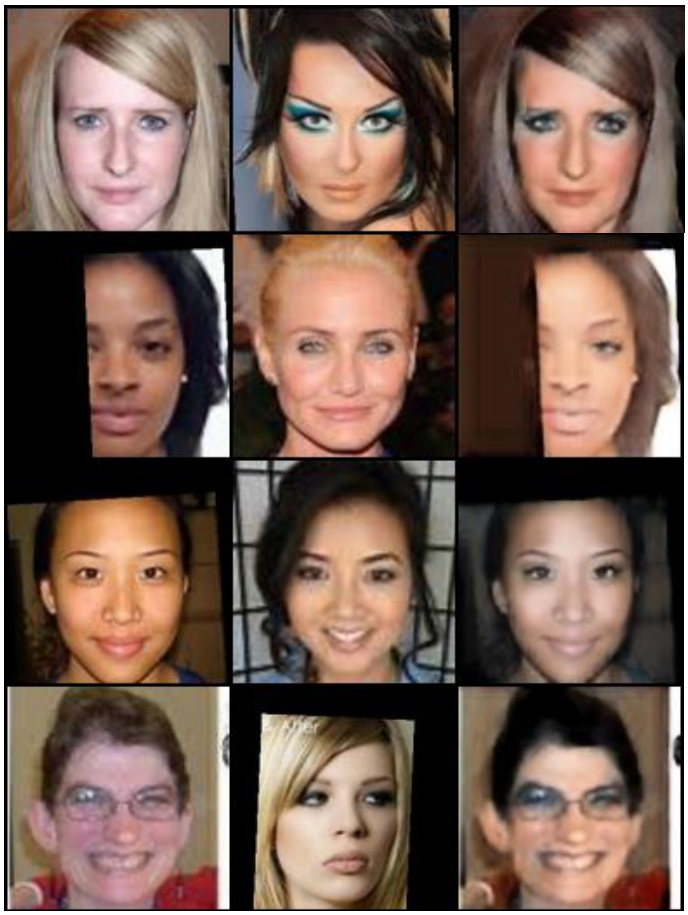


Figure 12: (a) low quality source images; (b) makeup images from which to extract the makeup style; (c) the inferred result. Our method is capable of handling noisy, partially cropped, real-world data.

in a better face representation, and a better encoded makeup style. As shown in Figure 12, our framework is able to transfer makeup style under a wide range of difficult conditions, including low-resolution images, cropped faces and radical makeup styles.

Quantitative results show that the weighted spatial attention module outperforms the state of the art at transferring makeup style on low quality data. We also provide qualitative examples to show that our model compares favourably to state of the art on difficult tasks.

In this chapter, the following contributions are provided:

1. We propose a new weighted multi-level spatial attention module to capture high-level and fine-grained style information. Such a mechanism is employed to encode the makeup style from the reference image and apply it to the source image with generative adversarial net-

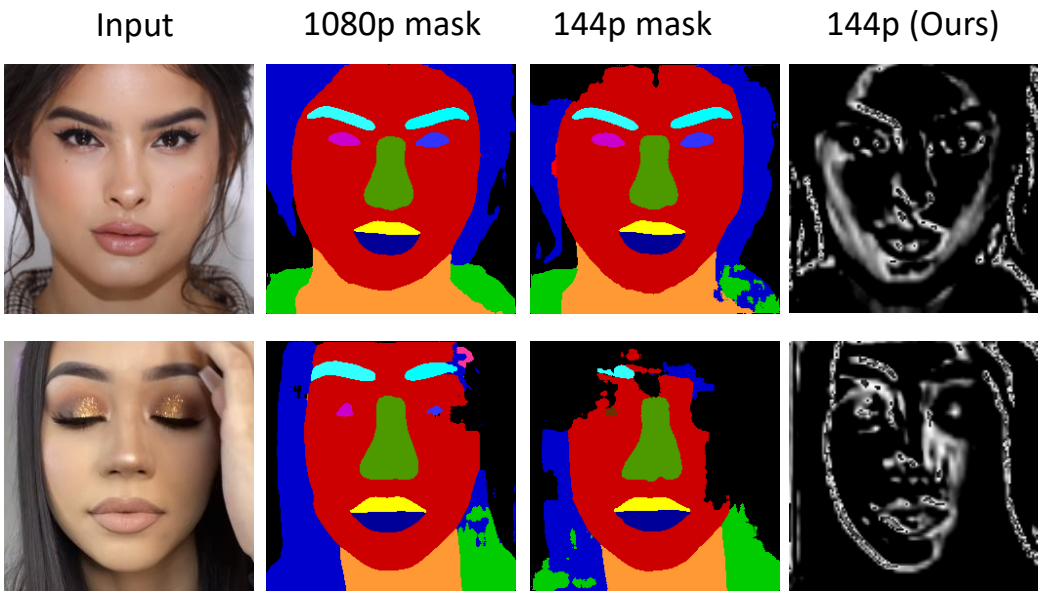


Figure 13: Hard attention, used in current state-of-the-art frameworks, on different resolutions. In the top row, due to facial pose and good lighting, the low resolution image can be segmented well. However, on row 2, closed eyes and occlusion from the hand causes segmentation failure. Multi-scale attention (lighter means higher weight) is more capable of handling these challenges and gives a more detailed attention map. Note that current state of the art is dependent on the attention maps, whereas ours still attains reasonable performance without attention.

works.

2. We demonstrate that state-of-the-art makeup style transfer techniques such as [131, 276] struggle to handle lower resolution data encountered in the real world. To handle this issue, an end-to-end framework based on Augmented CycleGAN is designed, with attention modules included within the generator, discriminator, and encoder.

3. We design a metric to quantitatively evaluate makeup style transfer.

The rest of the chapter is organised as follows: Some further background on makeup style transfer is outlined in Section 4.2. Section 4.3 describes the spatial attention mechanism and explores the full network for many-to-many image translation. Section 4.4 demonstrates our results compared to state-of-the-art methods. The chapter is concluded in Section 4.5.

4.2 Makeup Style Transfer Background

For digital makeup, an early work from Guo and Sim [52] can be used to transfer makeup style from one portrait to another. The source and target images are decomposed into three different

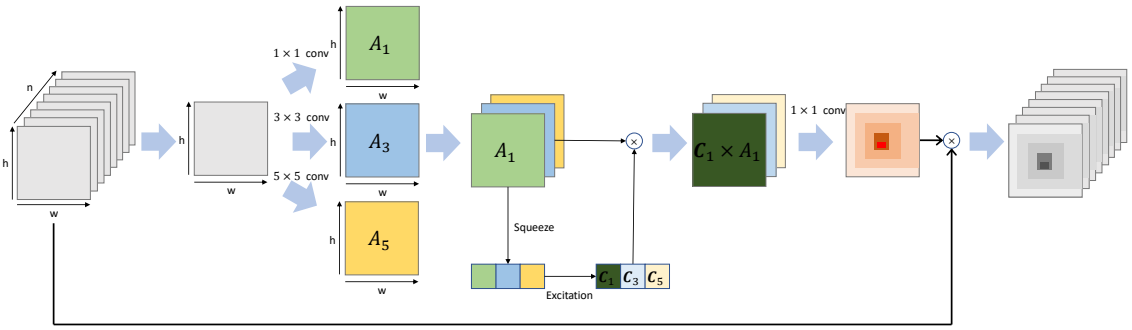


Figure 14: Our proposed weighted, multi-scale attention module. a) the input is squeezed along the channel dimension to obtain the representation matrix; b) the representation matrix is convolved through different sized kernels to extract the intermediate attention maps consisting of different scale information; c) the intermediate attention maps are concatenated and passed through a squeeze and excitation mechanism to assign each map a weight; d) this weighted multi-scale representation is passed through a final convolutional layer to obtain an $h \times w \times 1$ representation; this representation is multiplied by the input

layers: face structure layer, skin detail layer, and the colour layer. By altering the skin detail and colour layers, makeup style can be transferred. Xu et al. [263] proposed using face landmark detection to locate more important regions on the face and edit the skin colour and local details for each landmark. Li et al. [128] presented a physically-based model to alter the optical properties in the reflectance layers extracted from an image to simulate the digital makeup effects. More recent work by Liu et al. [142] proposed an end-to-end deep learning framework to i) recommend the suitable reference makeup style for the input image, ii) transfer the commonly used cosmetics (such as foundation, eye shadow and lip gloss) for different facial parts locally using the proposed *Deep Transfer Network*. The aforementioned approaches facilitate makeup style transfer based on underlying models or facial landmarks. Sub-optimal results will be produced if the model extraction and landmark detection are inaccurate.

CycleGAN [296] and other image-image translation works [297, 96, 3] demonstrated encouraging results on image-to-image translation tasks. PairedCycleGAN [28] improves the preservation of face identity by incorporating both a makeup transfer and a makeup removal networks. The face is separated into three parts, the eyes, lips and skin, and a generator-discriminator pair is trained for each part to capture unique characteristics. In addition to the typical cycle consistency loss and perceptual loss for ensuring the quality of the style transfer and realism of the resultant images, BeautyGAN [131] further includes the *makeup loss* to improve the appearance of the lips, eye

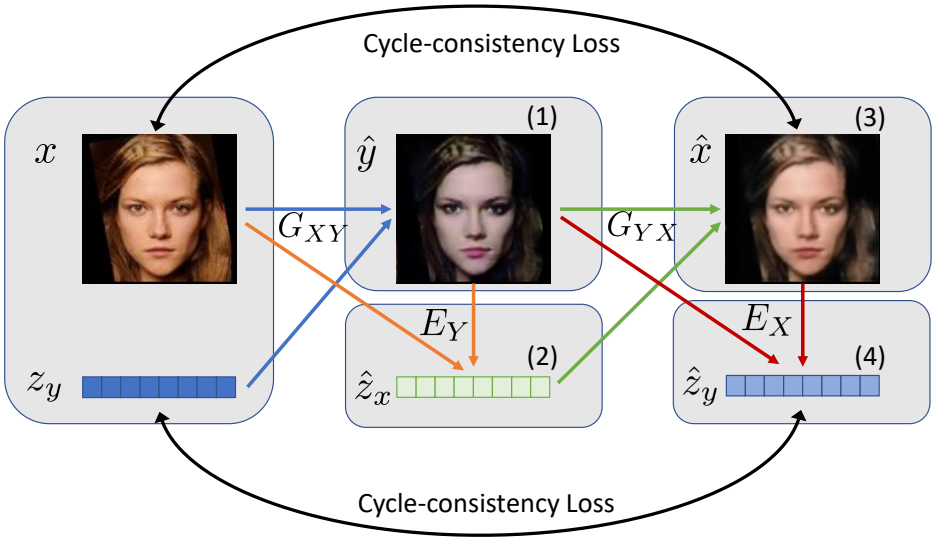


Figure 15: An overview of the Augmented CycleGAN baseline: a four step algorithm (denoted by blue, orange, green and red arrows, respectively) to maintain cycle-consistency for both the input image x and the input latent code z_y.

shadow and face regions. Zhang *et al.* [276] are able to not only transfer the makeup style from one image to another, but control the strength with which it is applied, or apply a hybrid of two different makeup styles. BeautyGlow [30] decompose makeup and non-makeup images into latent vectors, then combine them in the latent space. They then generate the new makeup image from the combination vector. Unfortunately, none of these methods have demonstrated the ability to work effectively on low-resolution images that are frequently encountered in real world applications of makeup style transfer.

4.3 Methodology

Many current state-of-the-art makeup style transfer frameworks are reliant on hard attention to segment the face into parts. Figure 13 shows that the same image taken at a low resolution can result in drastically worse performance. As real-world applications would benefit from efficacy on low-quality data, the state of the art should capably handle this data.

We propose a new *weighted multi-scale spatial attention* module that is composed of *spatial attention* and *channel attention* as an alternative to the face parsing algorithm. The spatial attention

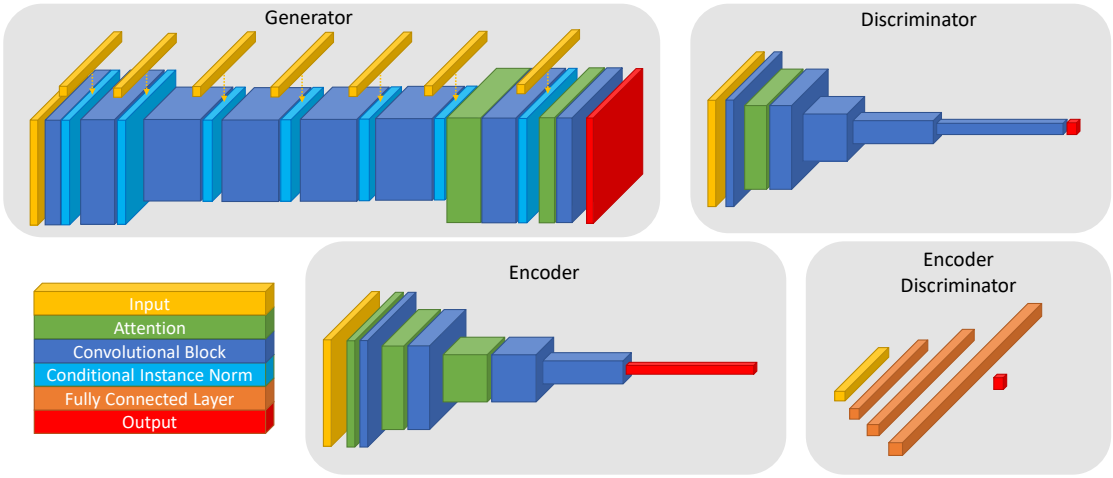


Figure 16: The full architecture of all of our networks

extracts saliency information from the image at different scales. The channel attention learns the relative importance of these scales to give these attention maps an associated weight. This design is motivated by the observation that one makeup style is composed of different aspects of makeup: foundation covers a large area over the face while eyeliner is only visible across a few pixels. Using multiple scales of spatial attention allows us to capture all of the information necessary for makeup style transfer. Computing spatial attention with a large kernel size may result in eyeliner being overlooked. Our module combines three different kernel sizes to help avoid this issue.

In the rest of this section, we will formally define the problem, then describe the proposed attention module that can capture makeup information of each image. The full procedure is outlined in Figure 14.

4.3.1 Problem Formulation

In this section, we will build on the formulation of CycleGAN give in Section 2.1.4

Given a set of images without makeup, X, and a set of images with makeup, Y, we aim to convert any image pair $(x \in X, y \in Y)$ into a new image $\tilde{x}_y$, that has transferred the makeup style from image y onto image x. Most image-to-image translation tasks apply an arbitrary style to image $x \in X$ to obtain $\tilde{x} \in Y$. For each $x \in X$, we only get one $\tilde{x} \in Y$. However, we want to apply the specific makeup style from $y \in Y$.

To accomplish this, we will use the Augmented CycleGAN [3] model as a baseline. Here, we

provide an outline of how it obtains a many-to-many mapping. As visualised in Figure 15, we simultaneously train two generators and two encoders with associated discriminators:

$$\left.\begin{aligned} G_{XY} &: X \times Z_Y \longrightarrow Y, \quad D_Y : \quad Y \longrightarrow \{0,1\}, \\ G_{YX} &: Y \times Z_X \longrightarrow X, \quad D_X : \quad X \longrightarrow \{0,1\}, \\ E_X &: X \times Y \longrightarrow Z_X, \quad D_{Z_X} : \quad Z_X \longrightarrow \{0,1\}, \\ E_Y &: Y \times X \longrightarrow Z_Y, \quad D_{Z_Y} : \quad Z_Y \longrightarrow \{0,1\}. \end{aligned}\right\} \tag{4.1}$$

The generators, encoders, and discriminators are trained with adversarial losses and cycle-consistency losses as presented in [296, 3], which we outline here. If not explicitly denoted, all symbols are defined as presented in Section 4.3.1.

Generator Losses: The loss for the image generator-discriminator pair is similar to a typical conditional GAN.

$$\begin{aligned} \mathcal{L}^{Y}_{\text{GAN}}(G_{XY}, D_Y) = &\mathop{\mathbb{E}}_{y \sim p_d(y)} \Big[\log D_Y(y) \Big] \\ &+ \mathop{\mathbb{E}}_{\substack{x \sim p_d(x) \\ z_y \sim p(z_y)}} \Big[\log(1 - D_Y(G_{XY}(x, z_y))) \Big], \end{aligned} \tag{4.2}$$

while the loss for the encoder network, which generates a 32-dimensional latent code from input faces is

$$\begin{aligned} \mathcal{L}^{Z_X}_{\text{GAN}}(E_X, G_{XY}, D_{Z_X}) = &\mathop{\mathbb{E}}_{z_x \sim p(z_x)} \Big[\log D_{Z_X}(z_x) \Big] \\ &+ \mathop{\mathbb{E}}_{\substack{x \sim p_d(x) \\ z_y \sim p(z_y)}} \Big[\log(1 - D_{Z_X}(\tilde{z}_x))) \Big], \end{aligned} \tag{4.3}$$

where $\tilde{z}_x = E_X(x, G_{XY}(x, z_y))$.

Cycle-consistency Loss: Both losses have an associated cycle-consistency restraint. For image generation, the loss is similar to cycle-consistency loss of CycleGAN.

$$\mathcal{L}^{X}_{\text{CYC}}(G_{XY}, G_{YX}, E_X) = \mathop{\mathbb{E}}_{\substack{x \sim p_d(x) \\ z_y \sim p(z_y)}} \big|\big|\tilde{x} - x\big|\big|_1, \tag{4.4}$$

where $\tilde{x} = G_{YX}\big(\tilde{y}, E_X(x, \tilde{y})\big)$ and $\tilde{y} = G_{XY}(x, z_y)$.

For the encoder, we reconstruct makeup style z_y via

$$\mathcal{L}_{\mathrm{CYC}}^{Z_Y}(G_{XY}, E_Y) = \underset{\substack{x\sim p_d(x) \\ z_y \sim p(z_y)}}{\mathbb{E}} \big|\big|\tilde{z}_y - z_y\big|\big|_1, \tag{4.5}$$

where $\tilde{z}_y = E_Y(x, G_{XY}(x, z_y))$.

The presented loss functions are combined as follows: the GAN loss, $\mathcal{L}_{\mathrm{GAN}}^{Y}$; the encoder generator equivalent, $\mathcal{L}_{\mathrm{GAN}}^{Z_X}$; the image cycle-consistency loss, $\mathcal{L}_{\mathrm{CYC}}^{X}$; and the encoder cycle-consistency loss, $\mathcal{L}_{\mathrm{CYC}}^{Z_Y}$ are combined with hyperparameters γ_1 and γ_2 to obtain the loss function in the non-makeup to makeup direction:

$$\begin{aligned} &\mathcal{L}_{\mathrm{GAN}}^{Y}(G_{XY}, D_Y) \quad + \quad \mathcal{L}_{\mathrm{GAN}}^{Z_X}(E_X, G_{XY}, D_{Z_X}) \\ &+\gamma_1 \mathcal{L}_{\mathrm{CYC}}^{X}(G_{XY}, G_{YX}, E_X) + \gamma_2 \mathcal{L}_{\mathrm{CYC}}^{Z_Y}(G_{XY}, E_Y). \end{aligned} \tag{4.6}$$

A symmetric equation is simultaneously optimised in the opposite direction. In our experiments, we assign $\gamma_1 = 1$, $\gamma_2 = 0.5$. γ_1 is given a higher weight because the task to reconstruct a 32-dimensional latent code through a cycle is easier than to reconstruct an image. However, we find that assigning γ_2 a reasonably large weight encourages the network to pursue more dramatic changes, even if they are less realistic. We prefer this because unaltered images are not at all useful for practical applications.

Overall, Augmented CycleGAN works on similar principles to CycleGAN, but also adds encoder-discriminator networks into the overall framework. These allow the framework to capture image-specific style, which enhances CycleGAN from a framework for domain translation into a framework for image-to-image translation. However, our early experiments concluded that Augmented CycleGAN alone is not sufficient to effectively transfer makeup between images. In the next section, the main architectural novelty, *Multi-scale Spatial Attention*, is introduced, so that the networks learn the important parts of the image to encode and transfer makeup style.

4.3.2 Multi-scale Spatial Attention

Spatial Attention aims to identify the most salient pixels. We develop a multi-scale spatial attention map to determine saliency at different granularities.

First, given a facial image, $p = (i, j), 0 \leq i \leq h, 0 \leq j \leq w$, are averaged across all channels, $s_p = \frac{1}{c} \sum_{k=1}^{c} p_k$, where c is the total number of channels to obtain a representative $h \times w$ feature map $\mathbf{Z}$.

We define the intermediate attention map, A_n, via $A_n = l_n(\mathbf{Z})$ where l_n is the convolutional layer with kernel $n \times n$. We process $\mathbf{Z}$ with three different convolutional layers, with $1 \times 1, 3 \times 3$, and 5×5 kernels simultaneously to obtain A_1, A_3, and A_5. Smaller kernels extract more detailed fine-grain information, such as eyeliner and lipstick that may be present. Larger kernels will be more adept at learning the importance of information that may cover larger areas, such as blusher applied to the cheeks.

The intermediate multi-scale attention is obtained by concatenating the attention maps in the channel dimension,

$$A = [A_1, A_3, A_5], \tag{4.7}$$

and bilinearly upsampled, so A has dimensions $h \times w \times 3$.

Finally, a 1×1 convolutional layer processes A to produce the final $h \times w \times 1$ multi-scale attention map, A_{MS}, which combines information from the weighted intermediate attention maps. A_{MS} is then multiplied by the initial input; this forces the spatial attention network to learn to assign greater values to more salient pixels. This process is outlined in Figure 4.1.

Our network is entirely self-contained and end-to-end. Unlike current state of the art, the framework is not reliant on any external models, algorithms, or software that can inhibit performance.

4.3.3 Network Architecture

The architecture of our full system as described in Equation 4.1 can be found in Figure 16. Rather than applying the attention model indiscriminately, we select where to apply attention to get maximum benefit without sacrificing efficiency.

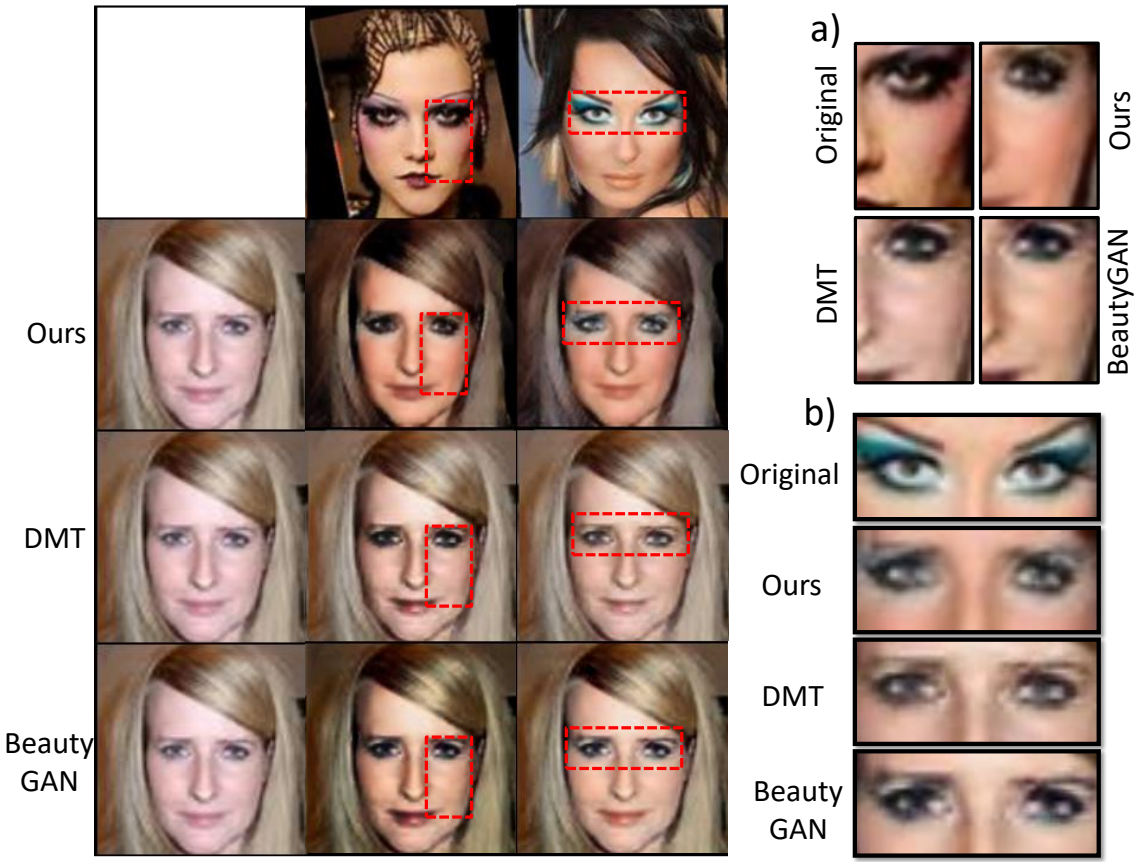


Figure 17: Comparison with DMT and BeautyGAN on challenging makeup styles: a) our method is the only one that captures skin tone, and best approximates the colour contours in the original image; b) our method best transfers fake eyelashes and comes closest to transferring the butterfly wings.

Encoder

We first explore the encoder as it is the network that benefits most from the attention module. The attention module is applied in early layers where the feature maps most resemble the input image. In fact, we apply attention directly to the image before feeding it into the network. We find that this significantly improves the encoder's ability to identify the makeup style to be encoded and incorporate it into the final feature representation.

Discriminator

The discriminator attempts to discover whether an image that has had the encoded makeup applied to it is real or fake, in the context of the makeup domain. By incorporating attention, the discriminator is more capable of identifying makeup in real images. In order to fool the discriminator, it becomes more important for the generator to apply makeup. However, if we add too

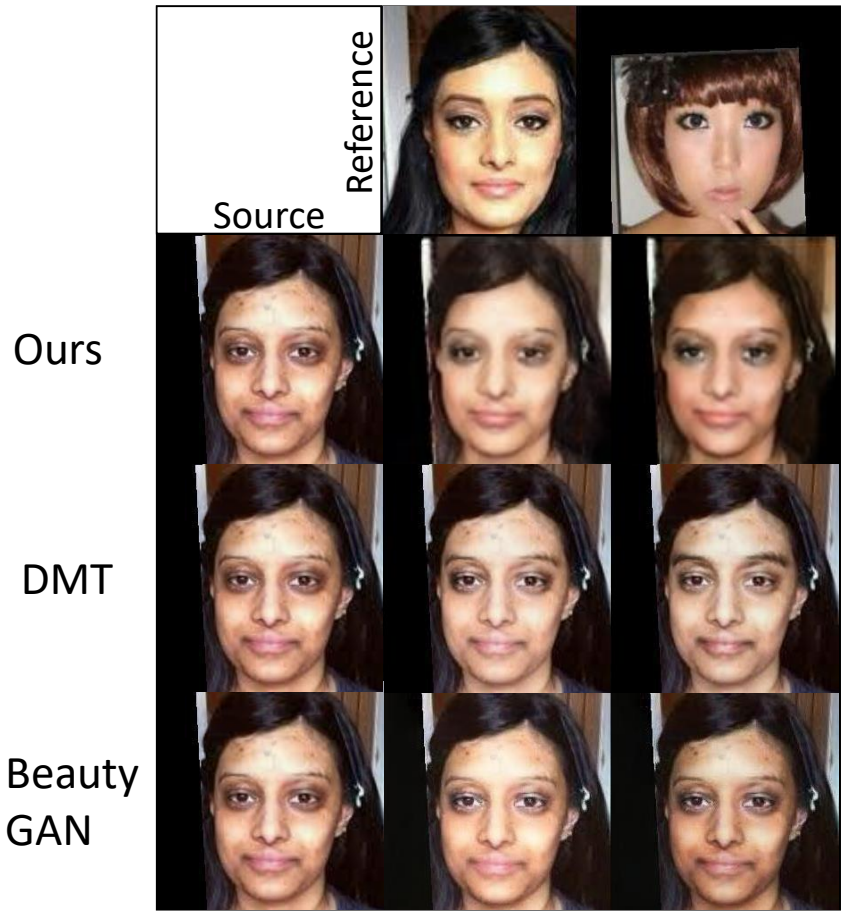


Figure 18: Comparison with DMT and BeautyGAN demonstrating our ability to cover blemishes compared with state of the art.

much attention, or incorporate it too early, the discriminator becomes too good at identifying fake makeup images, so the generator can't fool it. The ramification is that the system attempts only to maintain cycle consistency and very little image editing is performed. We settle on one weighted multi-scale attention module after the first convolutional block.

Generator

The inputs are a source image and a 32-dimensional latent code obtained from the encoder. This latent code is injected into convolutional feature maps at every layer via conditional instance normalisation [55]. Because the final layers contribute most to the generated image, we incorporate attention towards the end of the image generation process - before and after the final injection of the latent code of the makeup style. The attention module before the final style injection attends to areas where makeup should be applied, guiding the injected code towards those areas. The

attention module at the final layer attends to important areas of the face where makeup has been applied.

4.4 Evaluation

The experiments were performed on a workstation with four NVIDIA GeForce RTX 1070 Ti GPUs with 8GB of VRAM each. The training process took around 8 hours, while the testing process was within 5 seconds.

4.4.1 Evaluation Protocol

We train and evaluate our model on the FBD data set [278], a data set for makeup invariant face verification. It contains 2527 paired makeup and non-makeup images. We follow their pre-processing: the Viola Jones face detector [237] is applied to localise faces, then faces are cropped and aligned as per Shan *et al.* [208]. Pre-processing facial images to align facial landmarks is common; however, it has a propensity to introduce noise to images due to automatically scaling, skewing, cropping and zooming.

We also collect our own data set of 10 subjects from YouTube makeup tutorials, at 1080p and 144p, to allow us to perform quantitative evaluation, as explained in Section 4.4.3.

We compare against two state-of-the-art models: DMT [276] and BeautyGAN [131], because they are the best performing makeup style transfer frameworks that provide pre-trained models to test against.

Note that we are handicapped as DMT and BeautyGAN are trained on the MT data set [131], a lab created data set with mostly forward facing, high quality images with good lighting. In contrast, our models are trained on the low quality data set, which sometimes results in noise being added into the generated image.

4.4.2 Qualitative Evaluation

In Figure 17, we create a challenging task to transfer radical makeup styles onto a source image. Our framework outperforms the state-of-the-art methods, best transferring the skin-tone from fake tan and powder and in the makeup images. Ours also best approximates the colour distribution of

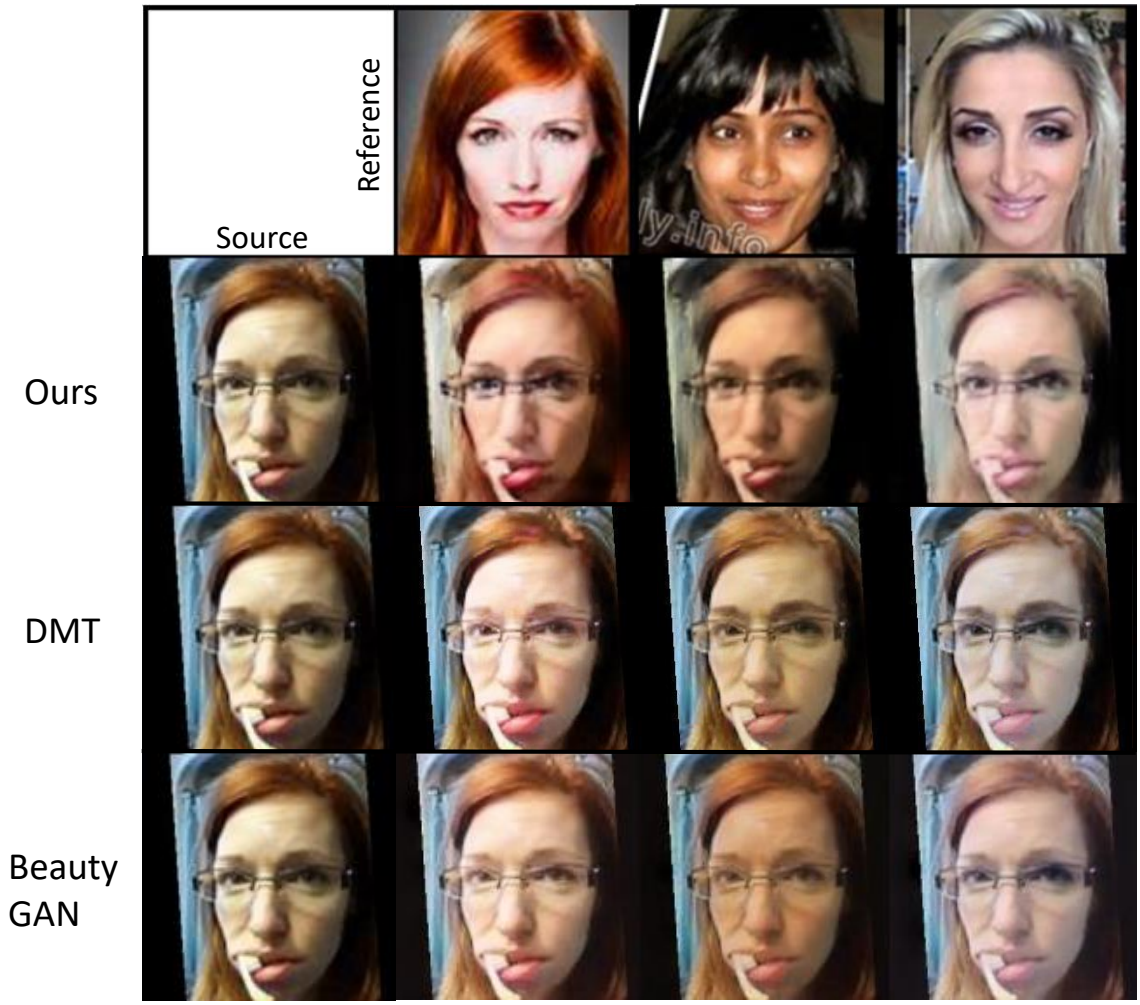


Figure 19: Comparison on source image with occluded lips and eyes.

the makeup face, applying blusher on the cheeks, whereas the other methods only apply a generic pale skin tone across the entire face. Other methods only apply a bold line around the perimeter of the eyes, whereas ours accurately applies fake eyelashes. We also best succeed at transferring the unconventional butterfly wings found in the source image. This shows that our soft attention is more accurately able to adapt to unconventional and extreme styles, compared to hard attention that is only capable of handling styles similar to what it has seen during training.

In Figure 18, we demonstrate our framework's ability to cover blemishes compared with state of the art. The first reference image shows the same subject as in the source image. By applying makeup, the subject has clearly chosen to conceal the blemishes on her forehead. Our framework is the only method capable of accurately transferring the makeup style in order to cover blemishes as desired. This highlights a major flaw with current state of the art methods. Because they use

Table 5: Comparison with state-of-the-art methods on the proposed Proportional Face Distance metric, measuring the accuracy of colour transferred from the reference image onto the source image for each face part.

Method	Eyes	Skin	Lips	Total
BeautyGAN [131]	0.230	0.086	**0.215**	0.532
DMT [276]	0.238	**0.084**	0.218	0.541
Ours	**0.197**	0.089	0.229	**0.515**

Lower numbers are better

Table 6: Attention mechanisms ablation study on the proposed Proportional Face Distance metric

Method	Eyes	Skin	Lips	Total
Ours	0.197	**0.089**	**0.229**	**0.515**
w/o Multi-scale Attention	**0.188**	0.105	0.236	0.529
w/o Any Attention	0.274	0.126	0.247	0.647

Lower numbers are better

hard attention to identify regions such as lips and eyes, they cannot adjust to different challenges. Our method is far more flexible in being able to handle outlier cases due to the holistic, soft attention approach.

Figure 19 provides a comparison on an image where both the lips and eyes are partially occluded. Ours best transfers the lip colour across all three images and is the only method able to apply fake lashes behind glasses in the first column. In the second column, DMT applies an unnatural pale green tinge that, far from beautifying the image, makes the subject appear unwell. Our framework can consistently apply makeup style on low-quality images without suffering from the problems that current state-of-the-art methods experience.

4.4.3 Quantitative Evaluation

Proportionate Face Distance Metric

The collected YouTube data set contains makeup and non-makeup images of subjects at 144p and 1080p. CelebAMask-HQ [124] was applied to the 1080p images (second column in Figure 13) to extract segmentation masks, and use them as ground truths of the 144p images for quantitative comparison. Each non-makeup image was then augmented with the makeup style from its own video. We did not add makeup styles from other videos to ensure that external factors, such as natural skin tone and different lighting, did not affect the results.

To perform quantitative analysis, we use the extracted masks to segment the eyes, face, and lips

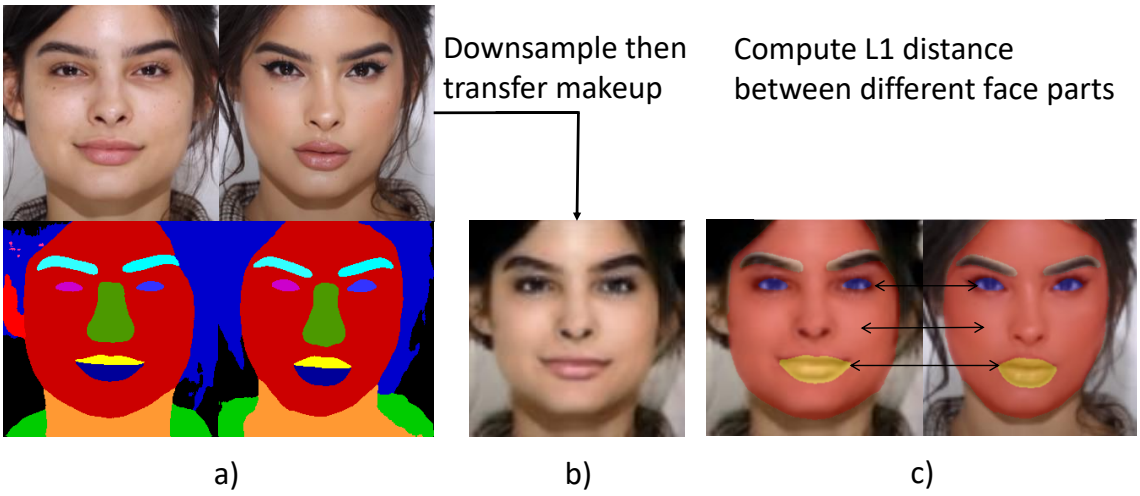


Figure 20: We design a quantitative evaluation metric for low resolution makeup style transfer: a) extract segmentation masks from 1080p images; b) downsample images to 144p and transfer makeup style; c) apply segmentation masks to the real and fake makeup image, compute colour histograms for each face part then calculate the L1 distance between similar face parts.

of the real 144p makeup images and the generated images. We then obtain the colour histograms of each segmented face part, and calculate the L1 distance, D, between colour histograms of equivalent face parts. This process is visualised in Figure 20. Because the skin takes up most of the pixels of the face, without any additional consideration, whichever method performed best at transferring skin tone would be adjudged to have performed best overall. We ensure a fair comparison by assigning a weight to each face part based on the inverse of the number of pixels that each face part has. This proportionate face distance is therefore found via

$$D_{\text{total}} = P\left(\frac{1}{p_{\text{eyes}}}D_{\text{eyes}} + \frac{1}{p_{\text{skin}}}D_{\text{skin}} + \frac{1}{p_{\text{lips}}}D_{\text{lips}}\right), \tag{4.8}$$

where P is the total number of pixels in the face mask of the generated image, p_i is the number of pixels in the masks of face part i, and D_i is the distance of the colour histograms of face part i, with $i \in \{\text{eyes}, \text{skin}, \text{lips}\}$.

The quantitative results in this section are based on D_{eyes}, D_{skin}, D_{lips}, and D_{total}.

Comparison with State of the Art

We compare against state-of-the-art methods on our YouTube data set in Table 5. Our model outperforms the other methods at transferring makeup style on low quality images with a propor-

tionate face distance of 0.515, 0.017 lower than the next best model. We obtain almost equivalent performance with state-of-the-art methods at transferring the makeup on the skin and lips, but significantly outperform both methods at transferring makeup around the eyes.

Note that these state-of-the-art methods have a signficant data advantage over us: they train on a larger data set, with higher quality data, and the data source we selected, YouTube makeup tutorials, is much more aligned with their data, in terms of the number of front facing, unoccluded images. They also have a modelling advantage, in that our model is a general framework that could be applied for any style transfer task, whereas their models can only be applied to makeup style transfer. Given the data advantage and modelling advantage that we have allowed state-of-the-art methods, it is already impressive that we attain similar performance on skin and lips, and even more impressive that we beat them on eyes and overall performance.

Eyeliner and fake eye-lashes are usually represented on an image by a small number of pixels so other style transfer techniques struggle to identify the fine grained style that needs to be transferred. The lowest scale of our attention module incorporates this information into the learned makeup style.

Ablation Studies

Table 6 shows the impact of dropping components of our attention module. Our model performs best at transferring the total face makeup, attaining the strongest performance on skin and lips. There is a surprising result when comparing performance at transferring style around the eyes - regular spatial attention performs better than our model. It appears that the spatial attention, without being able to attend at different granularities, has overfit to the data for transferring makeup around the eyes. Notice that this is to the detriment of transferring makeup for the rest of the face. Due to the presence of larger convolutional kernels, our multi-scale attention better identifies the importance of the skin and outscores regular spatial attention by 0.016. Furthermore, when training data samples from the spatial attention model were inspected qualitatively, the performance was significantly less consistent than the multi-scale attention model, in that style transfer performance was much more polarised between being very strong and very poor.

The model without any attention at all is considerably weaker at transferring all three face parts because, without attention, the background has a large contribution on the encoded makeup style.

As a result, the makeup style injected into the generator contains superfluous information.

4.5 Conclusion

In this chapter, we have developed an end-to-end framework for transferring makeup style that attains state-of-the-art performance on low-quality images. The framework does not suffer from the issues commonly seen among state-of-the-art methods, such as focusing only on lips, due to the developed novel weighted multi-scale spatial attention module.

One limitation of our method is that occasionally it can go too far with translating skin colour, and overly affect the background. Our framework favours riskier, more dramatic changes over safer ones. From an application perspective, it is more useful to have an extreme change than no change. If we wish to augment data for makeup invariant face recognition, extreme changes propose tough new challenges during training whereas little change does not assist training. Another limitation is due to training on low-quality data, the generator also learns to generate lower quality images compared to state-of-the-art models. A future work to rectify this is to train a super-resolution upsampling model to ensure that the generated image is at the same quality as the original input.

Chapter 5

Robust Feature Representation Learning for Person and Vehicle Re-identification

Similar to last chapter, a very different sphere of work is needed to further verify the usefulness of attention mechanisms across separate applications. This chapter will focus on feature representation learning for re-identification. This chapter focuses mostly on robustness, the following chapter will include more discussion on interpretability for re-identification. The main reason re-identification is selected to explore robustness is because the data availability is unique. First, there are multiple large-scale data sets available. More importantly, the tasks of person re-identification and vehicle re-identification have a lot of similarities, but also many challenging differences. Unifying these tasks is explored as a very challenging problem which is ideal to evaluate the robustness that attention mechanisms bring to a model.

This chapter will be an amalgamation of two separate works [174, 173]. Firstly, channel attention is demonstrated to significantly improve person re-identification (re-ID) performance. Secondly, to make the re-ID task more complex, the tasks of person and vehicle re-identification are unified and a novel loss function that attends to positive and negative pairs separately is constructed. Finally, the two papers are merged as out-of-distribution generalisation is tested, and channel attention is shown to be a key factor to remain robust in this difficult scenario.

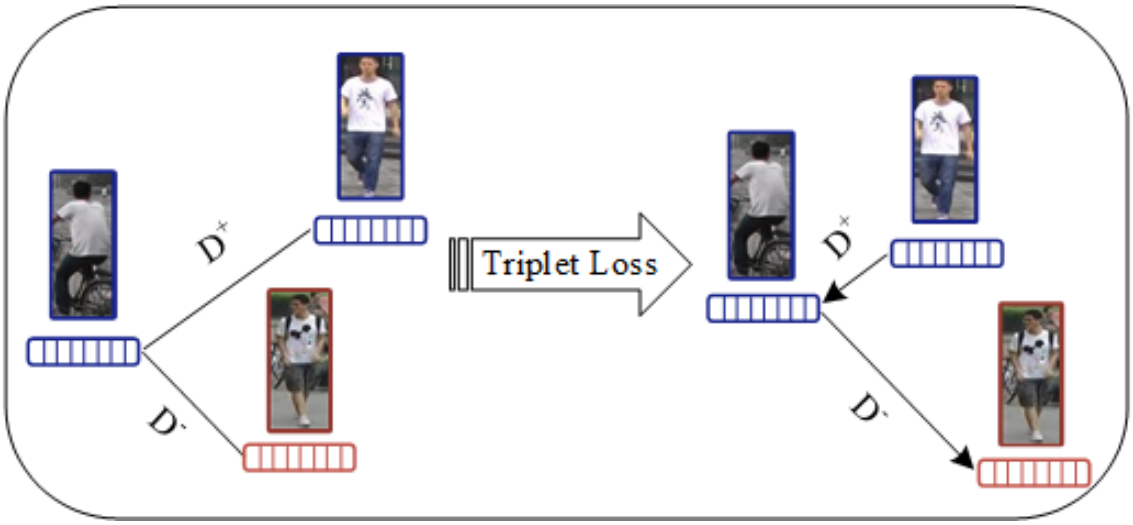


Figure 21: The triplet loss aims to reduce the distance of feature vectors from similar identities and increase the distance of feature vectors from dissimilar identities. We use channel attention in the form of squeeze and excitation units to get a better feature representation and improve the Euclidean distance by adding dynamic weights for each feature.

5.1 Introduction

5.1.1 Attention for Improved Triplet Loss

Re-identification is a core challenge for the computer vision community whereby a detection is required to be matched with another detection of the same object, typically from a different viewpoint. With the increasing volume of large-scale urban surveillance data, re-ID has started to attract a large amount of attention. In the past few years, deep learning techniques have received increased popularity due to significantly improving the performance of both pedestrian [33, 83] and vehicle [146, 292] re-ID.

Many person re-ID works make use of the standard convolutional neural networks with a cross-entropy loss [35]. More specific to re-ID, however, is the use of the triplet loss function [83, 36], either in place of or alongside the standard cross-entropy loss. The triplet loss, shown in Figure 21, enforces a distance margin, α, between the set of images of one person and all other images.

To date, most triplet loss works focus on mining better samples to improve the model generalisation [83, 4], or alter the loss function in order to increase the inter-class variance and decrease the intra-class variance [36, 33]. We instead focus on a) altering the deep learning architecture by adding channel attention to improve the feature representations learnt, b) adding dynamic weights

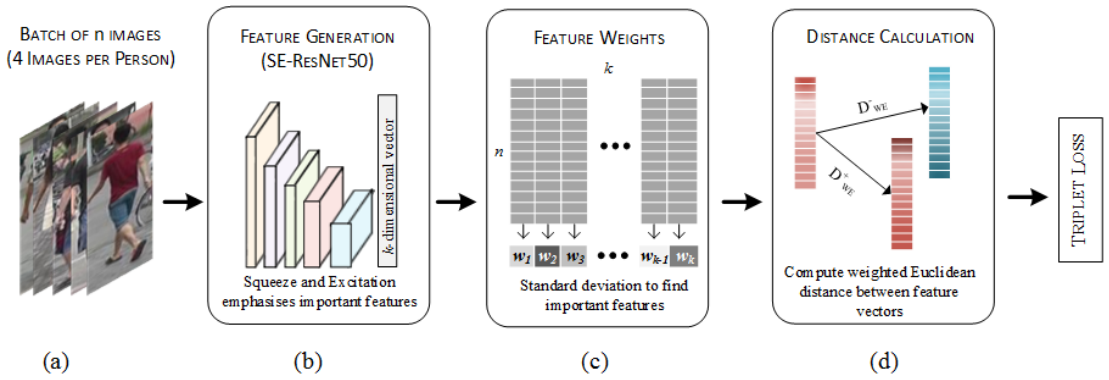


Figure 22: An overview of our architecture: (a) an input batch of n images is generated, (b) the batch is processed by SE-ResNet50 [89, 81] to generate one feature vector per image, (c) the standard deviation for each feature is computed then normalised to attain weights, (d) our improved triplet loss processes the mined triplets.

to the distance function with which the triplet loss compares these feature vectors. We show that these alterations improve re-ID precision individually. When implemented together, these adjustments complement each other, resulting in a performance improvement of over 9% mAP on the CUHK03 data set compared to the regular triplet loss.

Distance: The triplet loss, by its nature, attempts to decrease the distance between positive pairs of images while increasing the distance of negative ones. However, to date, little research has been done to assess exactly how this distance should be formulated. The Euclidean distance has been shown to perform well within the triplet loss function, thus has not received much scrutiny. We show that adding dynamic weights to the Euclidean distance can deliver considerable benefit when applied to the task of person re-ID.

The standard Euclidean distance considers all features as equally important. As shown in Figure 26 (c), our dynamically weighted Euclidean distance assigns an importance score to each feature derived from a feature's batch-wise standard deviation. Features with higher variance are more informative, thus assist the model to distinguish between images of different identities. To conceptualise this idea, if everyone in a batch wears a plain, white t-shirt, it is impractical to consider this information for re-ID. We assess the batch-wise feature vectors for high-level features that act in this manner and diminish their importance while highlighting more useful features.

Features: We would like our backbone architecture to generate feature representations of images which can best be exploited by the dynamically weighted Euclidean distance. In order to achieve

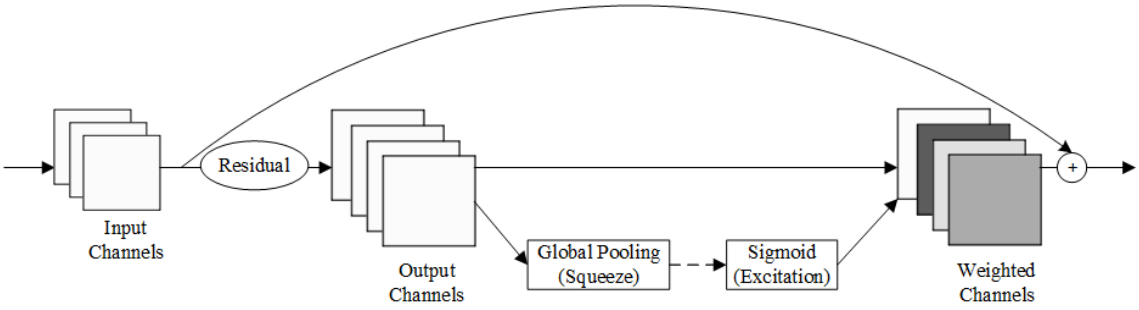


Figure 23: An overview of a ResNet block with a squeeze and excitation unit.

this, we use channel attention by adding SE units into our framework. These units act as weights to magnify important channels at each layer of the network while depreciating the value of less important channels. At deeper layers, these weights become more polarising to ensure salient features derived from the important channels are distinguishable from less important features.

As the less important features are mapped towards 0 by the SE units, they are more likely to have a low standard deviation and will therefore be assigned small weights by our dynamically weighted Euclidean distance.

The main contributions are as follows:

1. *Dynamically Weighted Euclidean Distance for Triplet Loss Feature Accentuation:* We introduce a weighted Euclidean distance which highlights features with high variation across the batch, in order to disregard features which are unimportant or susceptible to noise. This alone provides consistent performance improvement across all tested data sets.

2. *Feature Vector Generation with Channel Attention:* We are the first to adopt SE-ResNet 50 as the backbone architecture for the triplet loss. We demonstrate that the channel attention that SE units provide significantly boosts the performance of the triplet loss across a variety of data sets.

5.1.2 Unifying Person and Vehicle Re-identification

In the real-world, person and vehicle re-ID often need to be used together, e.g. when a person of interest boards a vehicle and gets off somewhere else. We would prefer re-ID systems to be able to handle this occurrence for continuous tracking. For this reason, Wei *et al.* [248] attempt to de-

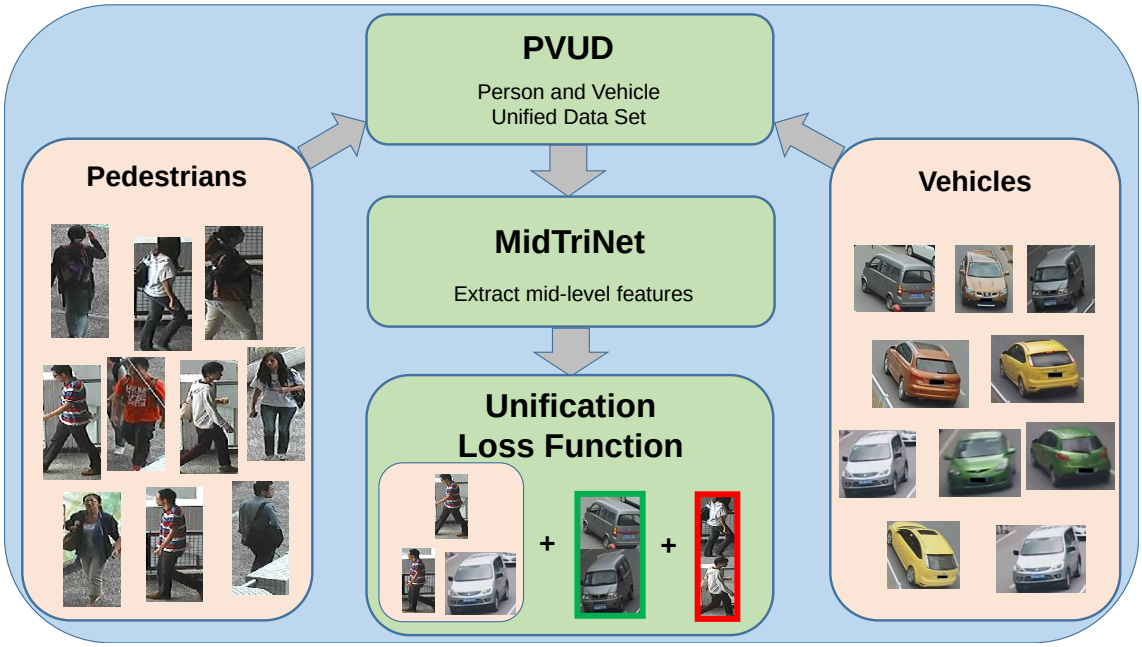


Figure 24: We propose a unified framework for pedestrians and vehicles re-identification using a new unified data set, PVUD, which challenges re-ID systems to be capable of handling both tasks simultaneously. Our framework includes MidTriNet to harness the power of mid-level features for re-ID, and a Unification Loss Function to better handle the mixed data stream.

velop an integrated application by using existing person and vehicle re-ID architectures. However, this does not truly unify the tasks, as the system accuracy depends on sub-systems, which is not optimal. It requires an additional component to classify between pedestrians and vehicles, which could introduce inaccuracies. We instead train person and vehicle re-ID in a unified manner. This approach allows us to discover underlying principles of re-ID, whereas handling the two systems separately does not allow us to explore this direction.

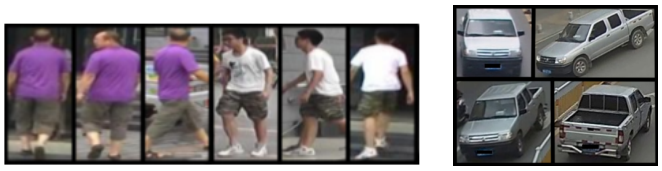
The challenges of re-identifying vehicles and persons have significant differences. For wide area video surveillance on humans, the same identity viewed from a different pose angle usually looks fairly alike. The shape of the detection remains upright and the colour information, predominantly extracted from articles of clothing, is of a similar pattern. The same condition cannot be satisfied for vehicles. Colour information can become far more distorted in different lighting due to the reflectiveness of the body of a car. The shape information of a car viewed from the front is significantly different than that viewed from a 45° or 90° angle. On the contrary, many high-end vehicle re-ID algorithms use license plate information [118, 143, 146], which is not applicable in the human domain. Moreover, pedestrians are more likely to undergo significant changes over time or viewpoint, e.g. a person's appearance is greatly altered after they put on a coat. In general,

changes to vehicles between viewpoints are high variance but predictable whereas the change in a person's colour representation is usually lower variance but prone to much more extreme outliers. We propose that there are *underlying principles of re-ID* that hold regardless of the composition of the object worked upon. Unifying person and vehicle re-ID allows us to explore and discover these underlying principles, precisely because they are so different.

Traditional works split the two tasks and design a network that can specifically target the individual task's respective challenges. However, we argue that this is inadequate. In the real world, urban surveillance videos provide a mixed stream of data, consisting of both vehicles and pedestrians, on which analysis is required. Mixing this data allows us to discover techniques and good practices, which are likely to extend to other re-ID tasks. For example, one may improve person re-ID performance by generating a better feature representation of pedestrians, e.g. by introducing squeeze and excitation modules. This tells us nothing about the framework's actual ability to re-identify an object, despite the accuracy increasing. Our data set helps to solve this issue.

We present an approach to unify the two tasks, summarised in Figure 24. We construct the Person and Vehicle Unified Data Set (PVUD) from other popular data sets, which is more representative of raw video surveillance data extracted from the real world. The data set is designed to be challenging and well-balanced, in order to prioritise re-ID systems that excel on both tasks. To the best of our knowledge, this is the first proposed re-identification data set containing both domains. We propose a triplet loss function that can be trained on either person or vehicle data and achieve state-of-the-art performance on each task. As the proposed framework is a form of metric learning, it does not require specific, domain-based design in order to re-identify objects. It is inherent in the framework to separate object classes in the same way that it separates different identities from one another, which makes it ideal to handle the challenges within the database that we design. We exploit information from mid-level layers which are more appropriate for the task of re-ID than the more abstract, final layer representations. In addition, we introduce hard negative and hard positive mining to our framework, with associated unification terms in the loss function, to improve its ability to handle multiple data streams.

We extensively test our proposed framework, attaining an 88.52% top-1 matching rate on PVUD, and competitive results with state-of-the-art methods on each of its components. The strong performance we obtain on the unified data shows that, contrary to discussion in [293], this is a realistic



(a) (b)

Figure 25: Matching people and vehicles contain different challenges: (a) Person shape and colour remains consistent across viewpoints; (b) Vehicle shape and colour changes drastically across viewpoints.

task on which to focus attention, particularly due to the presence of both pedestrians and vehicles within the vast majority of real-world, surveillance data.

The following contributions are made:

1. *The Person and Vehicle Unified Data Set (PVUD)* - Motivated by the composition of large-scale, surveillance data, we compose a challenging data set containing pedestrian and vehicle information to encourage the re-identification community to pursue the development of frameworks which are applicable to real-world data.

2. *Harnessing information from earlier layers* - We propose *MidTriNet*, a triplet framework which exploits information from mid-level layers. This information is more valuable than features from the deepest layers across both person and vehicle re-identification tasks.

3. *A unified framework* - We append unification terms to the triplet loss to derive a unification triplet loss function. We also introduce term-specific mining algorithms to discover the most important data for the unification terms to focus on.

5.2 Methodology

5.2.1 Channel attention with Dynamically Weighted Euclidean Distance

Although the triplet loss has seen extensive use in person re-ID, there has been little work to deviate from the standard Euclidean distance, despite it being a crucial element of the framework. We improve it by weighting each feature based on its importance.

To calculate which features are most discriminative, we use the $n \times k$ feature matrix output from the backbone of the network to calculate the standard deviation for each feature across the batch. This is shown in Figure 26 (c). The higher the standard deviation, the more variation in that

feature, and the more effective it is at helping the framework to tell people apart. These more meaningful features should thus be assigned a greater weight.

We use a softmax function regularisation on the standard deviations, then multiply by the total number of features to obtain the final weights. Overall, the weight, w_i, for the i-th feature can be calculated as

$$w_i = \text{softmax}\left(\text{s. d.}(\mathbf{F_i})\right) \times k, \tag{5.1}$$

where $\text{s. d.}(\cdot)$ is the standard deviation and $F \in \mathbb{R}^{n \times k}$ is the batch-wise feature matrix output by the backbone of the model with features $i = 1, \ldots, k$.

To ensure that the more important features are more prominent when calculating the distance matrix, we use the weighted Euclidean distance, D_{WE}, between two feature vectors, $\mathbf{x}$ and $\mathbf{y}$:

$$D_{WE}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k} w_i \left(x_i - y_i\right)^2}, \tag{5.2}$$

where w_i are the weights and k is the length of the feature vectors.

The standard triplet loss will separate embeddings to ensure that the distance between classes is greater than the hard margin α. Because we iteratively adjust the formulation of this distance, we are able to push classes apart even further, which leads to the model better representing the data.

Channel Attention Feature Embedding

The triplet loss evaluates the distance between feature representations, thus is very dependant on the quality of the feature vectors that are generated by the network. Furthermore, we would like these feature vectors to possess information which can be exploited by our dynamically weighted Euclidean distance.

We concentrate on improving the feature representations themselves by utilising channel attention [90] within ResNet50. Note that as channel attention maps unimportant features towards 0 throughout the network, they will typically have a low standard deviation. On the contrary, important features will be less impacted by the SE units and are therefore more likely to have a higher standard deviation. This means that our dynamic weights will be much more likely give

Table 7: Individual data set characteristics

Data Set	Train IDs	Test IDs	Images
Market-1501	750	751	32669
CUHK03	1372	95	14297
VeRI	576	200	40395
VehicleID	13134	13133	221763

Table 8: The composition of PVUD - The number of person and vehicle images are balanced to ensure the data set remains unbiased.

Data Set	Train IDs	Train Images	Test IDs	Test Images
Market-1501	751	12936	200	3486
CUHK03	1372	13176	95	921
VeRI	676	14632	100	2116
VehicleID	1500	12964	500	2085
Person Total	2123	**26112**	295	**4407**
Vehicle Total	2176	**27596**	600	**4085**

a large weight to features that are computed to be important by SE units, while still being able to identify features with high variance even though they are not determined to be salient by the network.

5.2.2 Person and Vehicle Unified Data Set

There is no publicly available data set for re-identification that contains objects from both person and vehicle classes. As re-ID frameworks are mostly applicable to surveillance data, which generally consists of pedestrians and vehicles, it is imperative for re-ID to be able to handle both streams simultaneously if it is to be applicable to real-world data. Moreover, testing on multiple domains concurrently allows us to be more confident that any adjustments made to the network are beneficial for the re-identification task in general, rather than just for a specific domain. To facilitate the research in this direction, we release a unified data set based on existing ones in the field.

We select the two most popular data sets in each domain - Market-1501 [283] and CUHK03 [132] for pedestrians, and VeRi [144, 145] and VehicleID [138] for vehicles. An overview of the raw data sets, containing the number of identities for training and testing, along with the total number of images can be found in Table 7. The final composition of PVUD can be found in Table 8.

Source Data Sets

CUHK03: The CUHK03 data set contains 14297 bounding boxes of 1467 persons. It contains

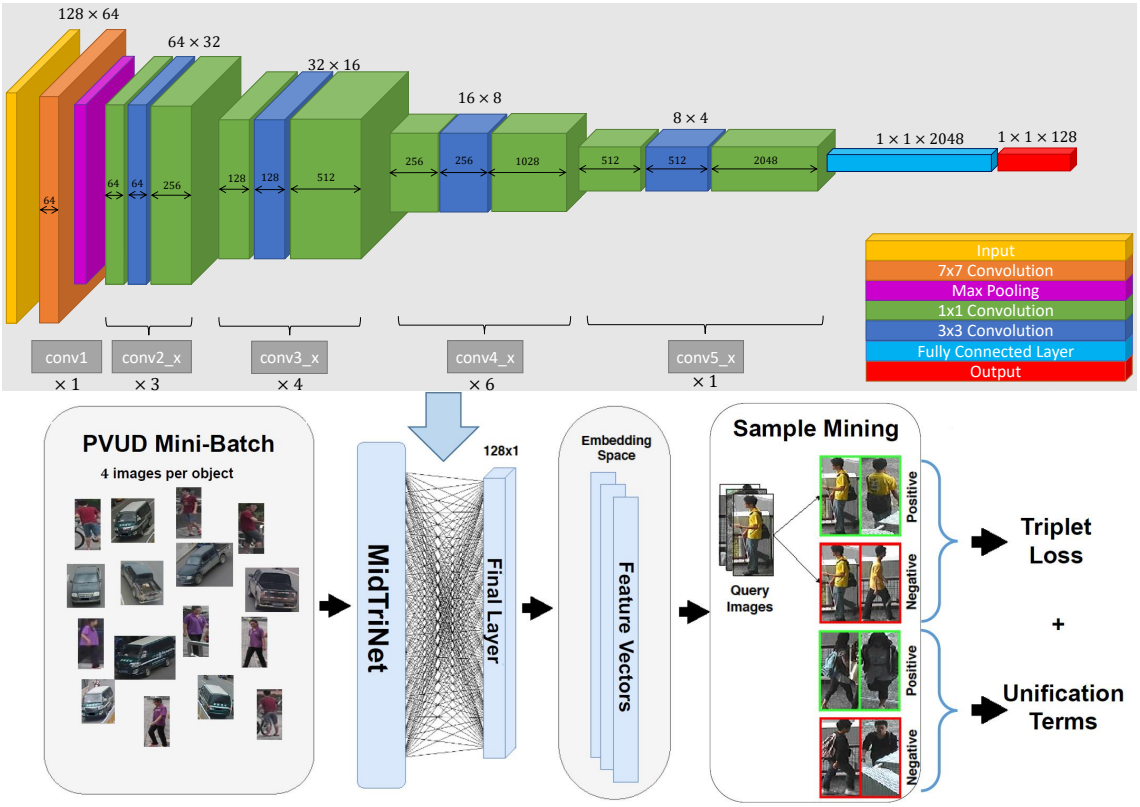


Figure 26: An overview of the architecture with unification terms. Each batch of images is processed with MidTriNet. We take the final layer of the network as the embedding space. We design unification terms specifically to make the network more robust against the mixed data that is present in PVUD and append them to the triplet loss function. Finally, we mine hard triplets, positive pairs and negative pairs to feed into our unification loss function.

two settings: one with manually annotated bounding boxes and one with automatically detected bounding boxes. We only consider the automatically detected setting as it contains some misplaced bounding boxes making it more challenging and more similar to what we would expect when applying re-identification to real-world tasks.

Market-1501: The Market-1501 data set contains 32668 automatically detected bounding boxes of 1501 individuals.

VeRi: The VeRi data set has 37,781 images of 576 vehicles for training and 11,579 images of 200 vehicles for testing. In order to obey the 'Balance' design principle, we move 100 vehicles from the test set to the train set. We also use a maximum of 20 images per vehicle. Rather than having standalone images from different viewpoints, VeRi contains 'tracks' of vehicles which are extracted as several consecutive frames from a video source. This means that images from

all angles are available. Thus, VeRi usually requires image-to-track calculation rather than the standard image-to-image metric that other data sets use. To maintain consistency across data sets, we use the image-to-image testing on PVUD.

VehicleID: The VehicleID data set has 221763 images of 26267 identities. VehicleID contains 'Small', 'Medium', and 'Large' settings for testing. As Market-1501, CUHK03 and VeRi are much smaller, for easier integration, we only take data from the 'Small' set. Contrary to the VeRi data set, VehicleID only contains images from the front and back of the vehicle.

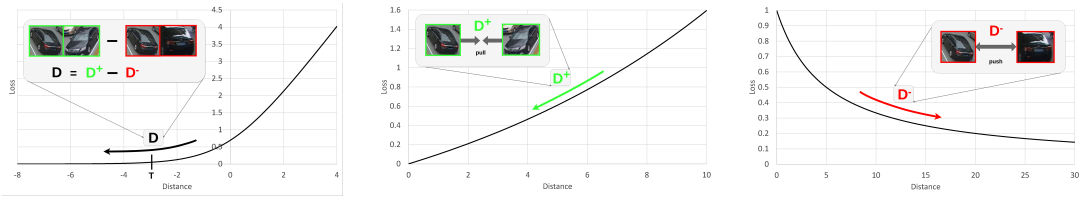
Design Principles

As discussed in [61], imbalanced data sets are inherently complex. When constructing this data set, it is important to ensure that person and vehicle data are equally balanced to accurately assess how strong a method is at re-identifying humans and vehicles simultaneously. As can be seen in Table 7, if we blindly conjoin the four data sets, there will be much more vehicle data than person data. This will result in the data set being biased towards vehicle re-identification methods, rather than methods which are effectively able to generalise across both tasks.

We lay out the following design principles to ensure the data set is as fair and balanced as possible without sacrificing difficulty. We provide full details of our constructed data set in Table 8.

Balance: A critical property of the data set is balance between different domains. In this regard, we have two options. We may either equate the number of vehicle IDs with person IDs in the data set, or the number of vehicle images with person images. We find that equating IDs leads to too many vehicle test images, which may result in weak person re-ID frameworks attaining an artificially high result. Balancing the number of images will facilitate a more challenging data set that better represents the real-world. Our studies shows that balancing IDs gives an mAP of 84.83%, whereas balancing images reduces the mAP to 77.51%.

Size: A data set with both pedestrians and vehicles is already challenging. We wish to take this challenge further. A larger testing set means more negative images to compare against, i.e. more likelihood to find a negative with a high similarity score, which makes testing more challenging. We also want to ensure that the data set is large enough for deep learning models, which demonstrate much greater efficacy at handling large-scale, real-world surveillance data. As we release



(a) softplus(D) (b) $\Psi(D^+) = \psi^{D^+} - 1$ (c) $\Phi(D^-) = \frac{\phi}{\frac{\phi}{10} + D^-}$

Figure 27: Visualisations of the softplus, Ψ and Φ functions used to calculate the overall loss function found in Equation (5.6)

the data set to motivate the re-ID community to design frameworks which can handle real-world data, it is imperative that it is suitable for deep learning frameworks. For these reasons, we select the design principle of maximising the size of the data set.

Random Sampling: In the interests of fairness, we randomly sample from the four comprising data sets rather than hand selecting examples.

Design Choices for VeRi:

VeRi requires image-to-track re-identification rather than image-to-image. Each track is composed of several consecutive frames from a video. We choose to include the entire track in order to more accurately model real-world surveillance videos. Many of the images used for training from VeRi are therefore similar to one another, so there is less effective information. This has two main consequences: (1) as VehicleID contains more effective training data, any framework must be capable of transferring knowledge between the data sets for accurate vehicle re-identification results, (2) models have to be robust against overfitting as VeRi training data can be very similar.

5.2.3 A Unified Framework for Person and Vehicle Re-identification

In this section, we will detail the implementation of MidTriNet and provide motivation for the design of our unification terms.

We present our architecture in Figure 26. The input batch is generated by taking four images of P identities, which are processed by MidTriNet and mapped into the embedding space. The distance matrix between all feature vectors is then calculated via a Euclidean distance and the hardest samples are mined. These samples are then fed into our novel loss function.

MidTriNet

Contrary to most deep learning classification tasks, mid-level layers have been shown to have similar importance as higher-level layers for constructing effective feature embeddings for re-ID [270, 272]. Re-identification relies on matching human-understandable information such as colour of clothes, and features are required to be viewpoint invariant. Mid-level information such as colours and textures, which are robust to viewpoint changes, are extremely useful information to discern whether an individual in the gallery set is the same as that in the query image. The very abstract features in the final layers are therefore not necessarily optimal for comparison, particularly within a triplet loss framework which attempts to differentiate between identities by directly comparing the feature representations of each image. To exploit the important information generated by the mid-level layers, we develop *MidTriNet*, which contains two major design choices throughout our experiments. These choices are supported by the ablation studies on the stride length (Table 18) and ResNet blocks (Table 19) provided in Section 5.3.3.

Layer removal: We remove the final two `conv5` blocks to strike a balance between the powerful representation ability that is characteristic of conv5 blocks and the re-identification task-specific efficacy of mid-level layers. Not only does this improve the feature embedding for re-ID, but also helps to protect the model against overfitting, which is extremely important for this data set as described in Section 5.2.2. Through our extensive experimentation, we find that removing the final two ResNet blocks works best.

Stride: To best exploit mid-level features for re-ID, we reduce the stride length in the `conv4` block from 2 to 1. This ensures that we have more informative feature maps at the important mid-level layers, which enriches the output of those layers to improve the final feature representation. The `conv5` stride length is typically 1 for this reason. Reducing the stride length of `conv4` to 1 allows us to focus on those features in the same way. This allows us to better compare the similarity between two images which benefits the network at all stages of training and testing.

Unification Terms

The triplet loss aims to simultaneously pull images of the same identity closer together whilst pushing away an image of a different identity. This can be difficult when dealing with unified data. Different data sets have different characteristics (camera intrinsics, lighting conditions, etc.),

so the feature representations are likely to be further apart from one another on average. This means that it is more difficult to find hard negatives (and thus hard triplets), so the model risks being unable to handle difficult situations when it comes to testing.

To counteract this, we mine the hardest negatives and positives across the batch. We design unification terms to separate hard negatives and compress hard positives, and append them to the loss function.

Loss Function

Let $\mathcal{T}$ be the set of triplets, where $t = (x_0, x_0^+, x_0^-) \in \mathcal{T}$ is a triplet comprising of a query image x_0, a positive image x_0^+ from the same identity as x_0 and a negative image x_0^- from a different identity. Let $f(x)$ be the feature vector of an image x. $\mathcal{H}^+$ is the set of the hardest positive pairs $h^+ = (x_1, x_1^+)$ with lowest similarity and $\mathcal{H}^-$ is the set of negative pairs $h^- = (x_2, x_2^-)$ with highest similarity (likewise, the hardest negative pairs). We set $\mathcal{H}^+ = \mathcal{H}^- = \mathcal{T} = 4P$ where P is the number of identities in each batch. Throughout this section, we refer to D as the distance between two feature representations.

The first term we use is a modified triplet loss function presented in [83]. Their analysis shows that replacing the traditional hard margin α from (6.2) with the softplus function $\mathrm{softplus}(D) = \log(1 + \mathrm{e}^D)$ is beneficial. The softplus function is shown in Figure 27(a). Overall, we have

$$\begin{aligned}\mathcal{L}_t = \sum_{t \in \mathcal{T}} \mathrm{softplus}(||f(x_0) - f(x_0^+)||_2 \\ - ||f(x_0) - f(x_0^-)||_2).\end{aligned} \tag{5.3}$$

The second term focuses on pulling together the positive pairs. We design the function $\Psi(D) = \psi^D - 1$, where $\psi > 1$ is a constant, in order to heavily punish large distances between positive pairs as seen in Figure 27(b). This forces the network to pull images from the same class together during training in order to keep the loss minimal. In our experiments we use $\psi = 1.1$. The positive unification term is written as:

$$\mathcal{L}_p = \sum_{h^+ \in \mathcal{H}^+} \psi^{\left(||f(x_1) - f(x_1^+)||_2\right)} - 1. \tag{5.4}$$

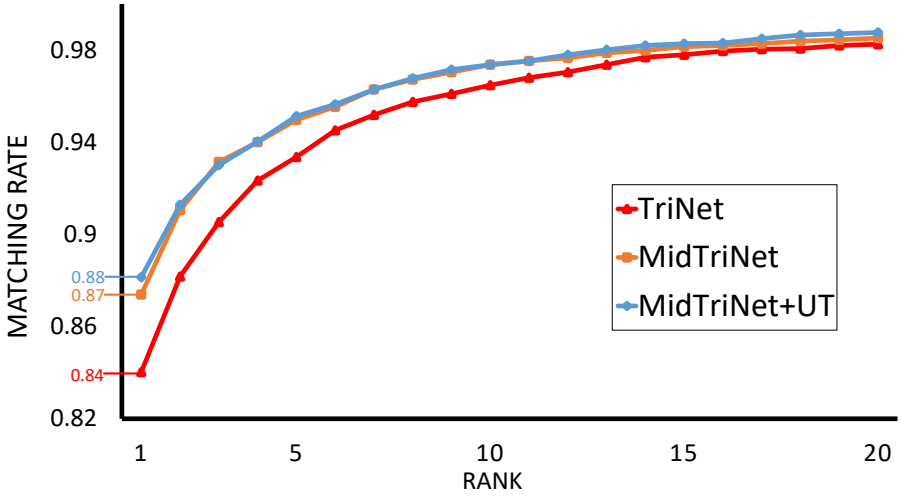


Figure 28: CMC curves for tested models on PVUD

Note that this is especially important in the vehicle domain. As discussed in Section 6.1, vehicle shape can change drastically in different viewpoints. One of the reasons why our network is so robust to this shape deformation is that we force the model to learn from additional hard positives, of which a large proportion will typically be vehicles in a very different pose.

The third term works similarly but aims to push negative images away from each other. We adopt $\Phi(D) = \frac{\phi}{\frac{\phi}{10}+D}$ to punish negative pairs with small distances and reward pairs with large distances as seen in Figure 27 (c). Throughout our experiments, we set $\phi = 30$. The negative unification term is written as:

$$\mathcal{L}_n = \sum_{h^- \in \mathcal{H}^-} \frac{\phi}{\frac{\phi}{10} + \left(||f(x_2) - f(x_2^-)||_2\right)}. \tag{5.5}$$

From Equations (5.3), (5.4) and (5.5), we obtain our unification loss function:

$$\mathcal{L}_\mathcal{U} = \alpha_t \mathcal{L}_t + \alpha_p \mathcal{L}_p + \alpha_n \mathcal{L}_n, \tag{5.6}$$

where α_t, α_p and α_n, are weights for their relative losses. Empirically, we found that setting $\alpha_t = 0.05$, $\alpha_p = 0.5$ and $\alpha_n = 0.5$ performs best.

Sample Mining

One of the most important elements of building a framework which utilises a triplet loss function is effective mining. We require it to effectively match vehicles with significant distortions in shape,

thus it is imperative that the model is trained on the most difficult samples available. Likewise, we wish for it to be able to handle outlier scenarios, e.g. where a person is wearing a bag, thus having a highly different appearance in different viewpoints. To challenge the framework to be able to handle these tough cases, sufficiently difficult triplets need to be mined. However, if the model is only trained on the most difficult triplets, it will not be representative of the entire data set and could struggle on easier examples.

Let p be the identity of the image $x_{p,i}$ in the batch, B, and let $f(x_{p,i})$ be its feature vector, where $p = 1, \ldots, P$ and $i = 1, \ldots, 4$. Each query image $x_{p,i}$ is paired with its hardest positive image x^+ and hardest negative image x^-, where:

$$x^+ = \max_{x_{q,j} \in B} \left(||(f(x_{p,i}) - f(x_{q,j})||_2 \right), \qquad \text{such that } p = q, \tag{5.7}$$

$$x^- = \min_{x_{q,j} \in B} \left(||(f(x_{p,i}) - f(x_{q,j})||_2 \right), \qquad \text{such that } p \neq q. \tag{5.8}$$

Together, we obtain the triplet $t_{p,i} = (x_{p,i}, x^+, x^-)$ and these form the set of triplets, $\mathcal{T}$, where $\mathcal{T} = 4P$.

In a similar manner, we scan across the entire distance matrix to find the set of hardest positive pairs, $\mathcal{H}^+$, and the set of hardest negative pairs, $\mathcal{H}^-$, with $\mathcal{H}^+ = \mathcal{H}^- = 4P$.

5.3 Evaluation

Evaluation Protocol

We perform experiments on the two most commonly used data sets to evaluate deep learning methods for person re-ID, CUHK03 [132] and Market-1501 [283]. In addition, we also provide results on VIPeR [74] to demonstrate that our method can considerably improve performance even on very small data sets, which many deep learning frameworks struggle on. We report the mean average precision (mAP) and top-1 matching rate (rank-1) scores for CUHK03 and Market-1501, and rank-1, rank-5 and rank-10 scores for VIPeR.

The rank-x matching rate is defined as the percentage of query images with a correct match within

the highest x ranks. The precision, P_x, of a framework at rank x is written as

$$P_x = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}. \tag{5.9}$$

To obtain the average precision for a given query, the average of the precision scores at each *true* positive in the ranking list is calculated. That is,

$$AP = \frac{1}{N} \sum_{i=1}^{N} P_i^+, \tag{5.10}$$

where N is the number of true positives in the gallery and P_i^+ denotes the precision at the i-th true positive in the ranking list. The mAP is then calculated by taking the mean of the average precision of all images in the query set. Throughout our experiments, we fix the margin $\alpha = 0.3$.

Many works boost the performance of their framework with a post-processing technique such as re-ranking [286] or a data augmentation procedure like random erasing [287]. Although our results would improve, we do not use re-ranking or random erasing in any of our experiments as it does not help to evaluate the core performance of the network.

Note that, to fairly compare the effect of the dynamically weighted Euclidean distance, we use it within the triplet loss function but don't apply it during the mining phase.

We also give results for *MidTriNet* and *MidTriNet+UT* (Unification Terms). MidTriNet is the baseline TriNet model, with the addition of the design choices described in Section 5.2.3 to harness mid-level features: reduction of the stride length in the `conv4` block, and removal of the final two `conv5` blocks. MidTriNet+UT includes the additional terms from Section 5.2.3 which specifically help the system to handle the mixed data in our unified data set.

All experiments are performed on a single NVIDIA GeForce GTX 1070 Ti GPU. Our model takes around 1 hour, 30 minutes to train on Market-1501 and around 35 minutes to train on CUHK03. We note that the system can further be optimised by tuning hyperparameters.

5.3.1 Person Re-identification

CUHK03: The CUHK03 data set contains 14297 bounding boxes of 1467 persons, with 767 identities used for training and 700 identities used for testing.

CUHK03 has two evaluation settings: the *labelled* setting contains bounding boxes that are manually annotated, and the *detected* setting contains bounding boxes that are automatically detected. We perform all of our experiments on the detected setting as it is a more realistic setup, which contains misplaced bounding boxes making the problem more challenging. It is more similar to what we would expect when applying re-ID to real-world tasks.

Market-1501: The Market-1501 data set contains images of 1501 people from six different cameras. The data set is split into 12936 images of 751 identities for training and 19732 of 750 identities for testing. We use the single query setting throughout all of our experiments.

VIPeR: The VIPeR data set consists of 632 pedestrians captured by two cameras. Deep learning methods typically do not report performance for this data set so we train all models ourselves with a batch size of 32. As VIPeR only contains one image per person in each camera, we replace the mAP metric with rank-5 and rank-10 precision scores.

Comparison with Baselines

We present our results with the baseline methods in Table 9. We select the baselines as ResNet50 [81] and SE-ResNet50 [89] with a cross-entropy loss, and TriNet [83] as our model is comprised of these elements. All triplet loss models in Table 9 are trained with the hard margin $\alpha = 0.3$ for direct comparison.

We comprehensively outperform baseline methods across all data sets. In particular, on CUHK03, we enhance the mAP of the triplet loss by 9.4% and the rank 1 performance by 9.3%. We also demonstrate considerable performance improvement on Market-1501 and VIPeR. The results show that both elements of our framework provide a significant contribution to enhance the re-ID precision.

Comparison with State-of-the-art

We further compare with state-of-the-art models (without re-ranking or random erasing) on the selected three data sets. In particular we note that our simple alterations are enough to give us the second highest mAP score of any core framework on the CUHK03 data set. We also notice that the state-of-the-art deep learning methods struggle to compete with ours on a small data set such as VIPeR, which demonstrates the robustness of our model.

Comparison with baseline methods							
Data Set	**CUHK03**		**Market-1501**		**VIPeR**		
Method	mAP	rank-1	mAP	rank-1	rank-1	rank-5	rank-10
ResNet50	26.3	26.6	68.3	85.8	11.1	32.6	44.0
SE-ResNet50	37.8	38.6	72.4	87.9	17.4	40.8	51.3
TriNet*	48.8	51.4	67.9	83.4	38.3	67.7	80.4
Ours: DWE TriNet	54.8	56.1	69.7	84.2	39.2	73.7	83.2
Ours: SE TriNet	52.9	54.7	73.1	**88.1**	40.2	69.6	80.4
Ours: SE+DWE TriNet	**58.2**	**60.7**	**74.2**	88.0	**44.9**	**75.6**	**86.1**

Table 9: Comparison with baseline methods. *Trained with a hard margin $\alpha = 0.3$.

CUHK03 (767/700) split		
Method	mAP	rank-1
DPFL [35]	37.0	40.7
SVDNet[222]	37.2	41.5
HACNN [134]	38.6	41.7
MLFN [29]	47.8	52.8
TriNet [83]	48.8	51.4
TriNet + RE [287]	50.7	55.5
DaRe [245]	51.3	55.1
PCB* [223]	57.5	63.7
HPM* [59]	57.5	63.9
MGN* [240]	**66.8**	**66.0**
DWE TriNet (Ours)	54.8	56.1
SE TriNet (Ours)	52.9	54.7
SE+DWE TriNet (Ours)	58.2	60.7

Table 10: Comparison with state of the art on the CUHK03 data set with the new split. *Use part-based information

CUHK03: Our results on the CUHK03 data set can be found in Table 10. It can be observed that the weighted Euclidean significantly boosts the performance on CUHK03.

We attain the second highest performance across all models on mAP. Our simple alterations are shown to outperform very sophisticated, state-of-the-art models that exploit spatial attention. In particular, we outperform the state-of-the-art methods PCB [223] and HPM [59]. The only method that exceeds ours, MGN, is heavily engineered. It takes different sized portions of the original image as input, which has been shown by multiple works to substantially improve performance. We note that (i) we can add this technique to our framework, (ii) they use a triplet loss in their model, which could be improved by adopting our formulation.

The most appropriate state-of-the-art method from Table 10 for comparison is Random Erasing [287], as it has become one of the most popular techniques within re-ID and also uses a triplet

Market-1501 (Single Query)		
Method	mAP	rank-1
DeepTransfer [65]	65.5	83.7
JLML [133]	65.5	85.1
TriNet [83]	67.9	83.4
TriNet + RE [287]	71.3	87.1
DaF [271]	72.4	82.3
DPFL [35]	73.1	88.9
HACNN* [134]	75.7	91.2
PCB* [223]	**81.6**	**93.8**
DWE TriNet (Ours)	69.7	84.2
SE TriNet (Ours)	73.1	88.1
SE+DWE TriNet (Ours)	74.2	88.0
MidTriNet (Ours)	73.4	87.8
MidTriNet+UT	74.0	88.9

Table 11: Comparison with baseline methods on the Market-1501 data set with the single query setting. For fair comparison, we don't include results which use re-ranking. *Use part-based information

loss. Our method comprehensively outperforms it, improving on its rank-1 accuracy by 10%. We further note that even if we keep the backbone architecture as ResNet50, simply changing the Euclidean distance function to our dynamically weighted Euclidean distance boosts performance more than Random Erasing. This further demonstrates the significance of the enhancements we have implemented and that the distance formulation is a crucial component which should not be overlooked when developing a triplet loss framework.

Market-1501: The results on the Market-1501 data set are presented in Table 11. We see that including squeeze and excitation blocks within the backbone architecture and adding dynamic weights into the Euclidean distance both enhance the framework. Our modified framework exceeds many state-of-the-art methods.

We note that although DPFL [35] and HACNN [134] beat us on Market-1501, their results on CUHK03 are much weaker, which indicates their models are heavily optimised towards the Market-1501 data set and not capable of generalising well. PCB [223] uses a part-based method which, as previously discussed, substantially improves performance and is compatible with our framework.

VIPeR: Performance on VIPeR can be found in Table 12. We outperform the state-of-the-art deep learning methods by 3.1% on the rank-1 matching rate. This demonstrates that our enhancements

VIPeR			
Method	rank-1	rank-5	rank-10
MLFN [29]	28.2	50.9	62.3
TriNet [83]	38.3	67.7	80.4
PCB* [223]	41.1	70.3	84.5
TriNet + RE [287]	41.8	71.2	83.5
DWE TriNet (Ours)	39.2	73.7	83.2
SE TriNet (Ours)	40.2	69.6	80.4
SE+DWE TriNet (Ours)	**44.9**	**75.6**	**86.1**

Table 12: Comparison with popular deep learning methods on the VIPeR data set. *Use part-based information

Method	rank-1	rank-5
XVGAN [292]	60.20	77.03
NuFACT [146]	76.76	92.79
VAMI [294]	77.03	90.83
PROVID [146]	81.56	95.11
HA-CNN [134]	83.00	92.41
TriNet [83]	83.25	95.23
MidTriNet (Ours)	88.56	**96.90**
MidTriNet + UT (Ours)	**89.15**	93.74

Table 13: Comparison on VeRi - Our method outperforms state-of-the-art and unification terms improve the rank-1 matching rate due to the diversity of the data set.

are very robust, even on data sets that do not have enough data for deep learning. In particular, we see that methods such as MLFN, despite performing well on popular deep learning data sets, do not have the ability to generalise as well as ours.

5.3.2 Vehicle Re-identification

In addition to PVUD, we test our framework on individual data sets. For person re-ID, the Market-1501 [283] and CUHK03 [132] data sets are selected. Meanwhile for vehicle re-ID, we use the widely used VeRi [144, 145] and VehicleID [138] data sets.

VeRi: The VeRi data set differs from other re-ID data sets as it maps temporally close images in the gallery onto tracks. The re-identification is computed from the query to the entire track (image-to-track) rather than just to gallery images (image-to-image). We follow the standard procedure for computing the similarity between a query image and a track, by calculating the similarity between the query image and all images on the track and then to take the maximum.

VehicleID: For the VehicleID data set, we follow the standard procedure as described in [138]. Given an identity i with N_i images in the test set, $\max(6, N_i - 1)$ images of identity i are placed

Table 14: Comparison on VehicleID - Our method outperforms state-of-the-arts, while UT has minimal effect on this saturated data set.

	Test Size = 800		Test Size = 1600		Test Size = 2400	
Method	rank-1	rank-5	rank-1	rank-5	rank-1	rank-5
VAMI [294]	63.1	83.3	52.9	75.1	47.3	70.3
BIER [169]	82.6	90.6	79.3	88.3	76.0	86.4
DREML [264]	88.5	94.8	87.2	94.2	83.1	92.4
DRDL [138]	49.0	73.5	42.8	66.8	38.2	61.6
TriNet [83]	91.5	**97.9**	89.3	96.1	85.5	94.2
MidTriNet (Ours)	**92.5**	97.6	**90.6**	**96.6**	**86.5**	94.6
MidTriNet+UT (Ours)	91.7	97.7	90.1	96.4	86.1	**94.8**

into the gallery set, and the remaining images are put into the query set.

Comparison with State-of-the-art

VeRi: We present our results on the VeRi data set in Table 13. Our method clearly outperforms the state-of-the-art at the vehicle re-ID task. We obtain a rank-1 score of almost 6% higher than the next best result.

VehicleID: Our results on the VehicleID data set can be found in Table 14. Our MidTriNet model consistently attains the highest mAP and rank-1 results across all three settings. Our methods considerably outperform state-of-the-arts, achieving 4% rank-1 improvement over the best method not to use a triplet loss.

Comparison with Baselines

PVUD: Our results on PVUD can be found in Table 15. It can be observed that the unification terms introduced in Section 5.2.3 boost the mAP by 0.92% and increase the top-1 matching rate by 1.13%. Moreover, we attain significant improvement over the standard TriNet on both networks. We also compare with ResNet where the triplet loss is replaced with cross-entropy loss, and two other state-of-the-art re-ID methods: Parts-based Convolutional Baseline (PCB) and Harmonious Attention Network (HA-CNN) [134] [223]. PCB is designed to specifically handle person data, and was trained with a ResNet-50 backbone for fair comparison. HA-CNN uses an attention module that learns from the data that is provided, so it could be applied to vehicles as easily as it is to persons.

Both methods provide a significant reduction of performance compared to standard ResNet, implying that attention diminishes performance in both cases. For PCB, this is because the part

Table 15: Results on our unified data set PVUD - The unification terms (UT) improve performance when the data is comprised from different domains due to the diversity of the data.

Method	mAP	rank-1
ResNet [81]	69.1	79.8
HACNN [134]	47.1	56.1
PCB [223]	64.7	73.3
TriNet [83]	74.5	85.2
MidTriNet (Ours)	76.6	87.4
MidTriNet + UT (Ours)	**77.5**	**88.5**

Table 16: Comparison on individual data sets when trained with PVUD - MidTriNet significantly outperforms TriNet and the unification terms improve performance when the training data is comprised of both vehicles and pedestrians.

Data Set	Market-1501		CUHK03		VeRi		VehicleID	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
TriNet	65.95	82.39	83.24	84.69	65.63	76.96	66.15	80.05
MidTriNet (Ours)	68.57	84.17	**87.25**	89.29	68.08	81.20	69.55	84.65
MidTriNet + UT (Ours)	**68.90**	**84.20**	86.77	**90.31**	**69.25**	**82.32**	**70.56**	**85.33**

models cannot adequately handle vehicle data as discussed in Section 2.4.2. PCB encodes better feature representations of pedestrians but cannot generalise to vehicle data that it was not designed to handle. Therefore, PCB obtains moderately better performance at re-identifying pedestrians within PVUD but drastically worse performance at re-identifying vehicles. This results in a net performance decrease of 4.4% when applied to the overall data set. In contrast, HA-CNN has demonstrated strong performance when trained on each domain individually. However, the attention mechanism of HA-CNN becomes confused when trying to handle two drastically different data types simultaneously. This results in a sharp decrease in performance. We discuss the impact of the individual unification terms in more detail in Section 5.3.3.

We compare against the baseline TriNet with two separate settings in Table 16. First, we train and test on individual data sets in the standard way. Secondly, we train all models on PVUD and test on the individual data sets to analyse how robust models are at handling data from different sources. This is very challenging. The model is required to be able to use information from training on one data set to test on another. Despite this, we see very little performance loss on any of the data sets. Both of our models considerably outperform the baseline on both settings for all data sets.

Further, we can see that the unification terms benefit performance across all data sets when the models are trained on PVUD. This demonstrates that the additional sample mining for the unification terms helps to create a much stronger model for handling mixed data.

Table 17: Ablation study on batch size - We see that the larger the batch, the stronger our performance.

Number of Identities	mAP	rank-1
18	72.10 ± 0.19	83.89 ± 0.30
32	76.78 ± 0.16	87.55 ± 0.24
36	$\mathbf{77.51 \pm 0.15}$	$\mathbf{88.52 \pm 0.22}$

Table 18: Ablation study on stride lengths of the `conv4` block - We see that reducing the stride length creates more informative mid-level features which boosts performance.

Block Strides	mAP	rank-1
2, 2, 2, 1	75.79 ± 0.17	86.77 ± 0.25
2, 2, 1, 1	$\mathbf{77.51 \pm 0.15}$	$\mathbf{88.52 \pm 0.22}$

5.3.3 Person and Vehicle Unified Dataset

PVUD: For PVUD, we take subsets of the standard train/query/gallery splits of each of the four individual data sets. These are procured by following standard evaluation protocol of individual data sets.

Note: as described in Section 5.2.2, the training set of PVUD contains some instances from the VeRi test set. For fairness, we exclude these when we train on PVUD and test on VeRi to analyse the robustness of our system.

On all data sets presented, MidTriNet significantly outperforms the baseline TriNet model. This consistent performance enhancement across domains provides conclusive experimental evidence that harnessing mid-level information is an underlying principle of re-ID.

Ablation Studies

In this section we present our ablation studies to demonstrate the benefits of a) mid-level information for re-ID, b) unification terms. All experiments in this section are performed on PVUD. We include confidence intervals at a 95% confidence level to demonstrate the significance of our design choices. We calculate these confidence intervals using the guidance for information retrieval tasks in [219].

Table 17 shows our ablation studies on the batch size. In particular, we find that larger batch sizes attain greater re-identification performance. This is because we can mine harder triplets, negatives and positives for our loss function so the framework learns more efficiently.

Our results with different stride sizes are presented in Table 18. We see that reducing the stride

Table 19: Ablation study on removing ResNet blocks - The final composition of MidTriNet (3,4,6,1), with two `conv5` blocks removed, significantly outperforms the others, validating our hypothesis that mid-level features perform best.

ResNet Blocks	mAP	rank-1
3, 4, 4, 1	76.59 ± 0.16	87.55 ± 0.30
3, 4, 6, 1	**77.51 ± 0.15**	**88.52 ± 0.24**
3, 4, 6, 3	76.11 ± 0.17	86.72 ± 0.25

Table 20: Ablation study on unification terms when trained on PVUD - Both unification terms are effective and they have a complementary effect when used together.

α_t	α_p	α_n	mAP	rank-1
1	0	0	76.59 ± 0.16	87.39 ± 0.24
1	1	0	77.02 ± 0.15	88.03 ± 0.23
1	0	1	76.93 ± 0.16	87.60 ± 0.24
0.05	0.5	0.5	**77.51 ± 0.15**	**88.52 ± 0.22**

from 2 to 1 in the third ResNet block boosts both the mAP and rank-1 performance by over 1.7%. This shows that the more informative mid-level feature maps are very important in boosting re-ID performance.

Table 19 shows that removing the final two `conv5` blocks significantly boosts performance. This supports the notion that mid-level features are more suitable than final level features for a triplet loss re-ID framework.

We perform ablation studies on our unification terms in Table 20. We see that both the positive and the negative term contribute to the overall score. When the negative term is excluded, the positive term provides a performance improvement of 0.43% on the mAP metric. Likewise, when the positive term is excluded, the mAP is 0.34% higher than the standard MidTriNet. We arrived a the highest performance with the unification terms weighted equally and very large compared to the standard triplet loss term.

5.3.4 Out-of-distribution Generalisation

Out-of-distribution generalisation is the ultimate test of a machine learning model as it is the best way to simulate how a model may perform in the real world. We take advantage of PVUD to evaluate the ability of models to generalise out of distribution. Each model is trained on images from CUHK and VeRi, and tested on images from Market-1501 and VehicleID. This is a particularly challenging machine learning scenario because each data set is collected in different lighting conditions, from different elevations, with different camera models. Furthermore, the vehicle poses

Table 21: Comparison with baselines for transfer learning. Methods are trained on images from CUHK03 and VeRi, and tested on images from Market1501 and VehicleID.

Method	mAP	rank-1
TriNet [83]	11.6	25.4
SE TriNet	12.6	27.0
HACNN [134]	16.9	35.1
ResNet [81]	22.9	42.2
SENet [89]	25.3	45.9
SE+DWE	**25.5**	**46.7**

collected from VeRi and VehicleID differ significantly.

We aim to evaluate the improvement that the addition of attention can make to evaluate it's potential impact on re-identification in the real world. Our results are provided in Table 21.

Notably, TriNet and SE TriNET perform particularly poorly at learning a new distribution. Although models with a triplet loss function learn a manifold that is appropriate for data from a particular source, it appears that different data sources will have considerably different manifolds, which restricts the ability of TriNet to generalise to unseen data sources. This indicates that triplet networks are likely to struggle to generalise to real-world scenarios unless an enormous amount of labelled training data from a large variety of sources is collected.

The introduction of channel attention, dramatically improves the performance of deep models, as seen with SENet and SE+DWE. In particular, channel attention with dynamically weighted Euclidean attains the strongest performance. Channel attention gives more weight to relevant features, which results in spurious features relating to background and lighting having less influence. Therefore, during training, the model learns based on human features relating to clothes and skin. Therefore, when testing on data from a different source, it also focuses on this information, which allows it to remain robust in this difficult setting. SE + DWE has a 4.5% performance improvement at its best guess, compared to the best performing model without attention, and is correct nearly half the time.

One interesting thing to note is that HACNN, which uses channel attention and spatial attention performs poorly. One possible reason for this is that the spatial attention does not do a good job of isolating the foreground objects, because the training data is composed of both persons and vehicles. HACNN also struggles with the regular PVUD data set for this reason, and the problem is exacerbated when test data comes from a different source than training data.

5.4 Conclusion

In this chapter, we have evaluated the effects of feature attention on the triplet loss function. We achieved this in two different ways: via assigning dynamic weights into the distance function used by the triplet loss, and by incorporating a backbone architecture with channel attention to emphasise important features throughout training. We demonstrate that both alterations alone boost performance of the triplet loss and complement each other for a significant improvement in precision when used together.

A balanced, challenging data set was constructed by combining the two most popular person re-ID and vehicle re-ID data sets. A triplet loss framework that beats or is competitive with state-of-the-art methods on both tasks and also attains high performance on our newly designed data set was designed. We proposed MidTriNet, to demonstrate that utilising mid-level features is an underlying principle of re-ID. Our design to exploit them boosts performance across all data sets. The unification terms presented in this chapter have been demonstrated to benefit the network, specifically when targeting mixed data streams. As a future work, we wish to explore this potential by deriving more complex mechanisms which target multi-domain data.

Finally, we have shown that attention is a major factor to improve a model's ability to generalise to data from a new source. This highlights how essential it is for the use of re-identification frameworks in the real world, where it is crucial that a model can perform effectively across a multitude of different lighting conditions and video qualities.

Chapter 6

Self-Attention for Robust Feature Representations of Unmanned Aerial Vehicles

In the previous chapter, the predominant focus was the ability of attention mechanisms to improve the robustness of DNNs for human and vehicle re-identification. This chapter extends this work in two ways. First, the novel task of unmanned aerial vehicle re-identification is proposed. Second, after again showing that attention-based networks are most robust to challenging data, the interpretability of attention networks is studied, to understand what a model considers important to create feature representations.

This chapter presents the novel task of unmanned aerial vehicle re-identification. This is a particularly challenging task because unmanned aerial vehicles are very small dynamic objects can that be viewed from any angle, highlighting the necessity of robustness to pose changes. Comprehensive experiments are conducted across a wide range of models to discover the best strategies to handle this challenge. Parts of this chapter are published at a peer-review conference [172].

6.1 Introduction

Unmanned aerial vehicles (UAVs) are becoming more accessible and more powerful through technological advancement. Their small size and manoeuvrability allows for a wealth of applications, such as film-making, search and rescue, infrastructure inspection, and landscape surveying. However, the malicious or accidental use of UAVs could pose a risk to aviation safety systems or privacy. This necessitates the development of counter-UAV systems. Due to the recent development of computer vision and deep learning, vision-based UAV detection and tracking systems have become more robust and reliable [95, 102].

There are two major issues with existing vision-based counter-UAV systems: firstly, many systems are only built for a single camera – once a UAV leaves the range of capture, the captured information can no longer be re-used; secondly, to help prevent ID-switching and handle occlusion, many tracking frameworks rely on a generic re-identification (re-ID) module [254], which cannot comprehensively handle the complex challenges that come with re-identifying UAVs [95]. Dedicated study to effectively re-identify UAVs is essential to solve both problems. To enable a cross-camera UAV system, effective re-ID is needed to match observed UAVs from one camera to another from different angles, poses, and scales. Generic re-ID mechanisms within off-the-shelf tracking frameworks can be improved by designing a bespoke UAV re-ID system to handle these extreme changes.

There has been a large body of research in re-ID for pedestrians [269] and vehicles [47]. Most state-of-the-art person re-ID research typically employs engineering solutions to improve performance, such as a 'bag of tricks' [149], which identifies several key re-ID principles to adhere to. Other works exploit the relatively static colour profile of pedestrians across views with part-based systems [223, 59]. In contrast, vehicles have drastically different appearances across views, so this information must incorporated into the model [294]. For UAV re-ID, even more consideration is required.

UAVs undergo a considerably greater change in scale than pedestrians or vehicles. UAVs are physically smaller than pedestrians and vehicles, and can be detected in cameras from a large distance, meaning they often appear very small in images. UAVs are also captured at a greater variety of angles, meaning more extreme shape deformation must be matched between views.

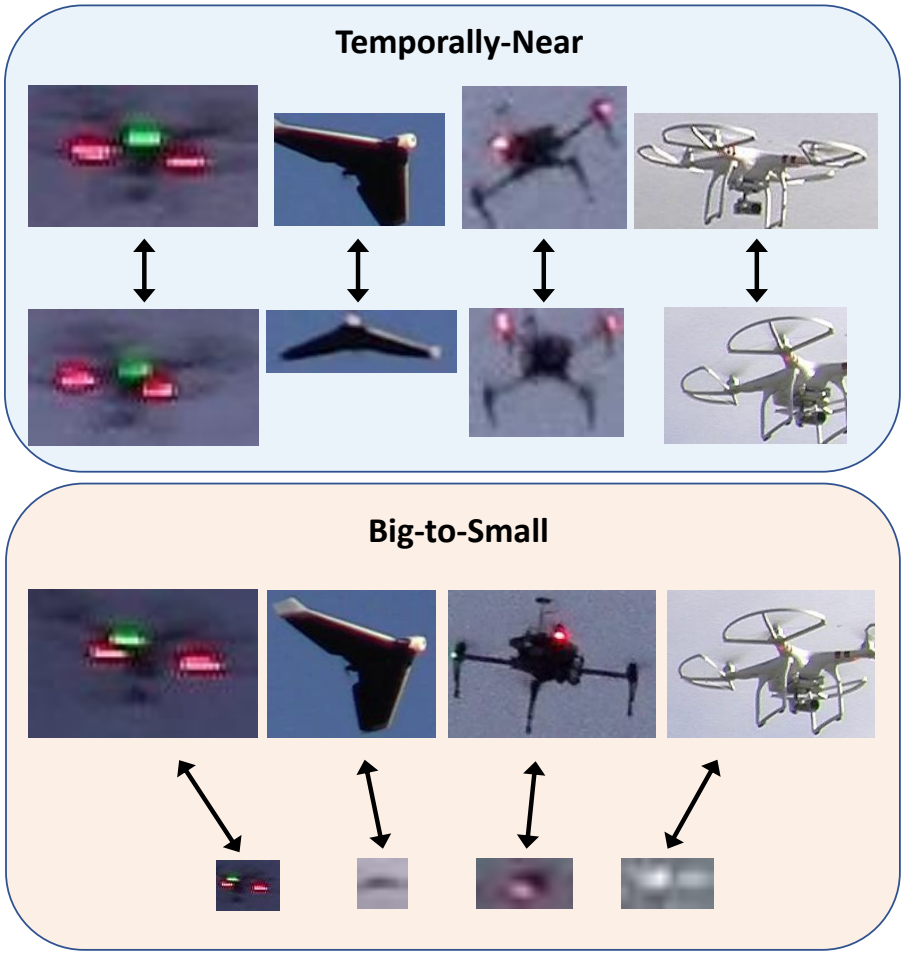


Figure 29: The two re-ID settings we explore. Temporally-Near models the difficulties of tracking UAVs, whereas Big-to-Small simulates cross-camera or temporally distant challenges of matching UAVs.

Because UAVs fly, they can appear from any angle on the sphere, compared to pedestrians and vehicles, that are typically captured from a 0-30°elevation. A study is required to evaluate the performance of existing re-ID systems on these challenges that UAVs provide.

However, to the best of our knowledge, there has been no research on UAV re-ID. In the absence of a true multi-view UAV data set, we propose *UAV-reID*, to train machine learning systems and evaluate re-ID frameworks for UAVs. To simulate re-ID challenges, UAV-reID has two settings: *Temporally-Near* aims to evaluate the performance across a short time distance, as re-ID modules within tracking frameworks must successfully identify the same UAV in subsequent frames within videos; *Big-to-Small* evaluates re-ID performance across large scale differences. The results inform re-ID performance of matching UAVs across two cameras, or across a large timescale within the same camera. Figure 29 visualises these settings.

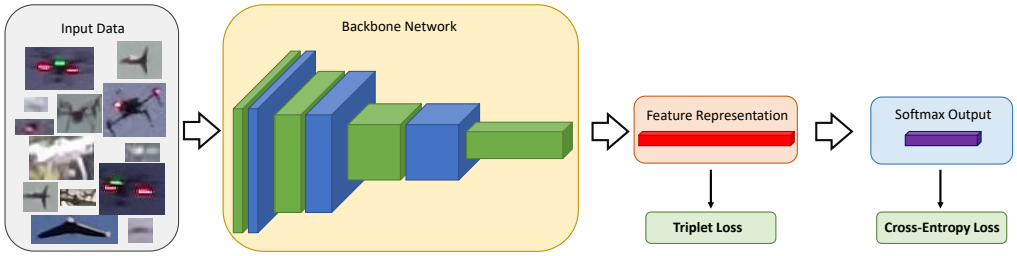


Figure 30: An overview of the pipeline for all of our experiments. Input data from the proposed UAV-ReID data set is processed by the given backbone network to obtain a feature representation. This feature representation is used in the triplet loss, and also goes through a softmax classification layer to be used in the cross-entropy loss. The backbone networks we evaluate are presented in Section 6.2.1

We conduct a benchmark study of state-of-the-art deep neural networks and re-ID-specific frameworks, including ResNet [81], SE-ResNet [89], SE-ResNeXt [260], Vision Transformers (ViT) [53], ResNetMid [270], Omni-scale Network (OSNet) [291], Multi-level Factorisation Network (MLFN) [29], Parts-based Convolutional Baseline (PCB) [223], and Harmonious Attention Network (HACNN) [134]. We test all baselines with a cross-entropy loss, a triplet loss, a combined loss, and a multi-loss.

Experimental results show that existing re-ID networks cannot transfer seamlessly to UAV re-ID, with the best setup achieving 81.9% mAP under Temporally-Near and 46.5% under Big-to-Small. ViT is the most robust to extreme scale variance.

The contributions of this chapter are summarised as follows:

- proposal of the novel task of UAV re-ID to match UAVs across cameras and time frames, to improve visual security solutions on UAVs
- We create the first UAV re-ID data set, *UAV-reID*, to facilitate re-ID system development and benchmarking
- We design two settings, *Temporally-Near* and *Big-to-Small*, to evaluate performance under conditions where re-ID is used in a practical environment
- We create the first extensive benchmark of state-of-the-art re-ID frameworks in the domain of UAVs, and critically evaluate their strengths and weaknesses, obtaining 81.9% mAP on Temporally-Near and 46.5% mAP on Big-to-Small.

6.2 Methodology

6.2.1 Deep Neural Network Backbones

Deep neural networks (DNNs) are machine learning systems that use multiple layers of non-linear computation to model the complicated relationship between the input and output of a problem. Convolutional neural networks (CNNs) are particularly suited for image-based object identification and tracking in computer vision applications. Firstly, CNNs can capture object features irrespective of their spatial locations on an image, due to the shift-invariance of convolution kernels. Secondly, modern CNNs can detect objects of complex shapes, sizes, and appearance by stacking multiple convolution kernels to learn powerful feature representations. We describe a selection of state-of-the-art CNNs and generic re-ID frameworks that we evaluate for UAV re-ID. Our overall framework is shown in Figure 30.

ResNet

Residual neural networks [81] are a popular variant of CNNs that connect adjacent layers of a network (residuals) with an identity mapping. Learning residuals unlocks the ability to train significantly deeper architectures to obtain more powerful features. In our experiments, we use the 18-layer, 34-layer, and 50-layer configurations.

SE-ResNet

ResNets are powerful but can still be improved by learning and re-weighting the hidden convolutional feature maps using attention. The popular Squeeze-Excitation (SE) network [89] introduces a channel attention mechanism to identify and appropriately weight important feature maps.

SE-ResNeXt

Another line of improvement for ResNet is ResNeXt [260], which maintains the identity skip connection while splitting the feature mapping of each layer into multiple branches. This increased dimension of network representation power has shown to be more effective for image recognition and object detection.

ViT

Transformers have recently become ubiquitous in natural language processing. Motivated by this, Dosovitskiy *et al.* [53] migrated transformers into computer vision to propose *Vision Transformers*. This architecture learns the relationship among all image patches for downstream tasks. We evaluate ViT with image patches of size 16×16 with the 'small' (8-layer) and 'base' (12-layer) configurations.

ResNet50-mid

A common practice of image representation learning in computer vision is to take hidden features from the penultimate CNN layer as image embeddings. Yu *et al.* [270] explore fusing embeddings from earlier layers to improve the performance of cross-domain image matching. Fusing representations from different layers has proven successful for other computer vision tasks on small objects [140], highlighting its potential within UAV re-ID systems.

OSNet

There have also been CNN architectures specifically designed for object re-ID. Zhou *et al.* [291] propose an omni-scale network, which improves re-ID performance by learning to fuse features of multiple scales within a residual convolutional block. Each stream in the block corresponds to one scale to learn and the outputs of all streams are dynamically combined to create omni-scale features. Considering the expansive array of scales at which UAV can appear, OSNet is well-suited to the challenge of UAV re-ID.

MLFN

Multi-level Factorisation Network [29] is similar to OSNet in that it tries to capture discriminative and view-invariant features at multiple semantic levels. The main difference is that it composes multiple computational blocks, each containing multiple factor modules and a selection gate to dynamically choose the best module to represent the input.

PCB

Different from holistic feature learning, Sun *et al.* [223] propose a *parts-based convolutional baseline* (PCB), which uniformly splits each input image into multiple parts. As the appearance consistency within each part is usually stronger than between parts, it proves easier to learn more robust and discriminative features for person re-ID. A part pooling module is added to deal with outliers.

HACNN

Li *et al.* [134] propose a *harmonious attention network*, which tackles the challenge of matching persons across unconstrained images that are potentially not aligned. HACNN uses layers that incorporate hard attention, spatial attention and channel attention to improve person re-ID performance on unconstrained images.

6.2.2 Loss Functions

This chapter will explore the ability of cross-entropy loss, triplet loss, combined loss and multi-loss. For convenience, an overview of cross-entropy loss and triplet loss is included here. (For more details, see Section 2.1.3.) The multi-loss is then explained in more detail.

Cross-Entropy Loss

The cross-entropy (CE) loss function is the standard loss that is used in most machine learning classification tasks. The negative log-likelihood between the true class labels and predicted class labels is minimised:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{x \in \mathcal{X}} y_x \log f(x;\theta), \tag{6.1}$$

where a network f with parameters θ predicts the class of an input x with a true class index y_x.

Triplet Loss

The triplet loss is a metric learning technique that decreases the distance between positive pairs of images and increases the distance of negative pairs. We denote a triplet, $t = (x, x^+, x^-)$, where x is the query image, x^+ is an image of the same object, and x^- is an image of a different object.

The triplet loss function is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{triplet}} = &\sum_{t \in \mathcal{T}} \max((||f^*(x;\theta) - f^*(x^+;\theta)||^2 \\ &- ||f^*(x;\theta) - f^*(x^-;\theta)||^2 + \alpha), 0),\end{aligned} \tag{6.2}$$

where $\mathcal{T}$ is the set of mined triplets, $||\cdot||^2$ is the Euclidean distance, and the feature representation $f^*(x;\theta)$ is obtained by passing input x through network f with parameters θ, and taking the representation before the softmax classification layer. Negative images are pushed away from positive images by a margin of α.

6.2.3 Combined Loss

In many re-ID works, combining the two losses can lead to performance gains. In the person re-identification domain, work has shown that giving these terms equal weight gives the most consistent performance [149]. This setting is followed for UAVs:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{triplet}}. \tag{6.3}$$

Multi Loss

Following the success of [18], we further evaluate the performance of a multi-loss function that has demonstrated superior performance to more well-established loss functions within the person re-ID domain. This loss is formulated as a weighted sum across cross-entropy loss, $\mathcal{L}_{ID}$, ranked list loss, $\mathcal{L}_{RLL}$, centre loss, $\mathcal{L}_{\text{centre}}$, and erasing-attention loss, L_{E_att}, as follows:

$$\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_{RLL} + \beta \cdot \mathcal{L}_{\text{centre}} + \mathcal{L}_{E\_att}. \tag{6.4}$$

As such, all losses receive equal weighting other than centre loss which serves to support $\mathcal{L}_{RLL}$, and thus receive weight β. We define $\mathcal{L}_{ID}$ as cross-entropy loss with additional label smoothing [225]. $\mathcal{L}_{RLL}$ can be considered a direct alternative to triplet loss, and learns a hypersphere for each class additionally to triplet loss behaviour. Learning the hypersphere helps avoid intra-class data distribution that might be apparent within triplet loss, and particularly impactful when training

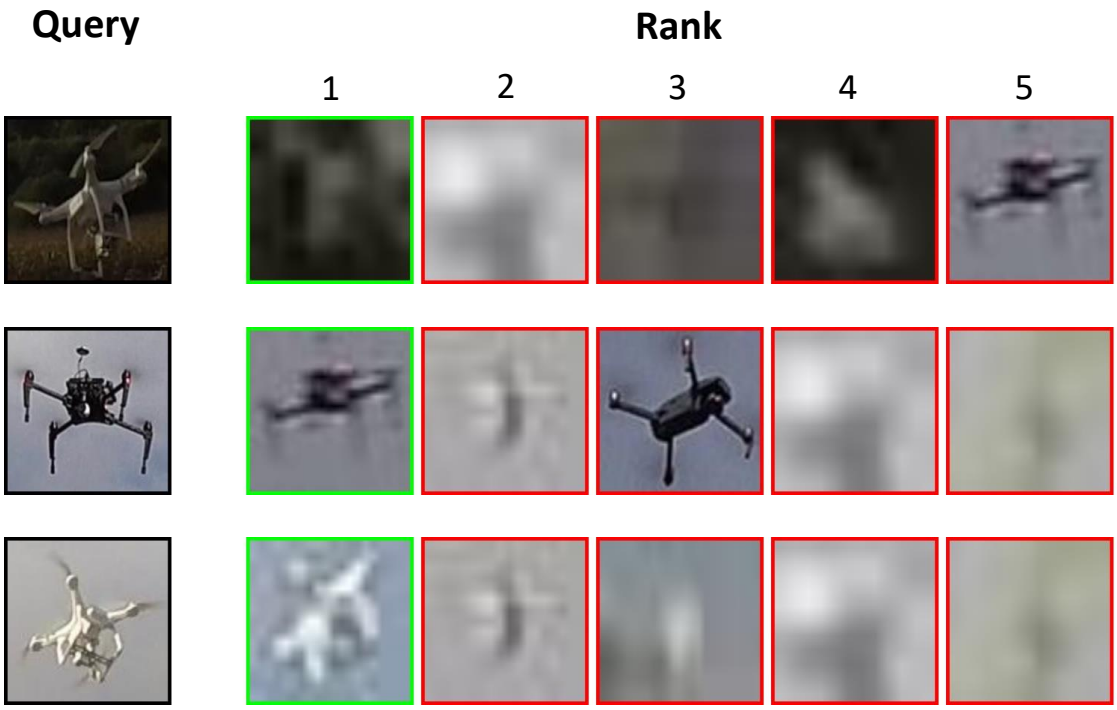


Figure 31: Examples from ViT with a combined loss on Big-to-Small. A green box indicates a correct re-ID. ViT can extract salient features from very low-resolution images to match UAVs across scale.

with limited data. Finally, $\mathcal{L}_{E\_att}$ introduces additional attention to image samples that receive erasing under random erasing augmentation [287] such that its impact is increased, as implemented in [18, 181]. This is particularly important when data availability is constrained so the effects of over-fitting are minimised during training; learning will be maximised from features extracted from erasing-augmented images that are less likely to contribute to UAV regions.

6.3 Evaluation

6.3.1 Data

UAV-re-ID is designed to evaluate two practical applications of re-ID. All data is captured from the 61 videos in the Drone-vs-Bird data set.

Temporally-Near

Given a UAV video with t frames, we consider UAVs in frames $\frac{t}{5}$ and $\frac{2t}{5}$. This temporal distance is close enough that UAVs remain at a similar size in most cases, but far enough for UAVs to

Table 22: Methods Tested on the 'Temporally-Near' setting.

Model	CE			Triplet			CE + Triplet		
	mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	rank-5
ResNet-18	**81.9**	**77.4**	77.4	72.7	61.3	*74.2*	71.7	58.0	77.4
ResNet-34	77.1	70.1	74.2	74.6	71.0	71.0	74.4	61.3	83.9
ResNet-50	75.9	71.0	71.0	*75.5*	71.0	71.0	76.7	67.7	77.4
SE-ResNet-50	77.1	71.0	80.6	74.1	67.7	*74.2*	79.4	74.2	*80.6*
SE-ResNeXt-50	75.8	71.0	77.4	66.8	61.3	64.5	76.2	74.2	74.2
ViT Small	75.6	67.7	74.2	74.1	64.5	*74.2*	75.6	64.5	74.2
ViT Base	79.2	*74.2*	77.4	73.2	67.7	*74.2*	*81.3*	**77.4**	*80.6*
ResNet50mid	78.0	71.0	**87.1**	74.0	67.7	*74.2*	76.1	67.7	77.4
OSNet	71.0	61.3	70.1	73.8	67.7	71.0	75.7	71.0	71.0
MLFN	69.9	61.3	71.0	73.4	67.7	67.7	65.7	58.1	61.3
PCB	*80.8*	*74.2*	**87.1**	73.2	67.7	67.7	81.4	**77.4**	*80.6*
HACNN	72.1	64.5	71.0	77.7	71.0	77.4	74.5	64.5	77.4

Bold denotes the highest values in the table, red denotes the highest in each column, blue denotes the second highest in each column

appear from a different viewpoint. This simulates the task that a re-ID module embedded within a tracking framework must perform, whereby UAVs undergo a limited transformation.

Big-to-Small

We obtain the largest and smallest UAV detections across the whole video. This simulates the task of matching known UAVs (for which we have rich visual information) with UAVs detected from a long distance. As such, we can identify the far-off UAV, and whether it poses a potential threat. For a DNN to perform well at this task, they need to be extremely robust to perturbations in size. It is very challenging to ensure that the representation of a large drone remains similar to the representation of the same drone when it is much smaller.

6.3.2 Evaluation Protocol

During training, we convert images to size 224×224 and augment images via random flipping, random cropping and random erasing [287]. The test set is split into a *query set* and a *gallery set*, with 31 identities each. Given a query image, q, the re-ID framework ranks all gallery images, g_i in order of likelihood that $g_i = q$, i.e. they contain the same UAV.

Table 23: Methods Tested on the 'Big-to-Small' setting.

Model	CE			Triplet			CE + Triplet		
	mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	rank-5
ResNet-18	40.3	*32.3*	41.9	36.9	25.8	32.3	37.5	25.8	32.3
ResNet-34	33.7	22.6	29.0	37.9	29.0	35.5	38.8	25.8	35.5
ResNet-50	37.8	22.6	*51.6*	39.0	29.0	35.5	42.9	29.0	35.5
SE-ResNet-50	38.0	25.8	*51.6*	*42.5*	29.0	45.0	41.4	29.0	38.7
SE-ResNeXt-50	40.0	29.0	35.5	31.9	16.1	29.0	38.8	29.0	32.3
ViT Small	43.1	35.5	35.5	39.0	22.6	*41.9*	40.9	29.0	38.7
ViT Base	40.5	29.0	**54.8**	36.9	22.6	32.3	**46.5**	**35.5**	45.2
ResNet50mid	38.4	25.8	*51.6*	42.3	32.3	32.3	*43.2*	*32.3*	38.7
OSNet	38.0	25.8	35.5	34.5	19.4	35.5	33.2	19.4	32.3
MLFN	38.1	22.5	38.7	36.8	25.8	32.3	33.9	22.6	25.8
PCB	*41.3*	*32.3*	35.5	43.7	32.3	*41.9*	38.2	25.8	32.3
HACNN	36.0	19.4	45.2	39.4	25.8	32.3	41.2	25.8	*41.9*

Bold denotes the highest values in the table, red denotes the highest in each column, blue denotes the second highest in each column

We use the standard 'mean average precision' (mAP), 'rank-1' and 'rank-5' metrics to evaluate our framework against the state-of-the-art methods. The rank-r matching rate is the percentage of query images with a positive gallery image within the highest r ranks. The precision at rank r, P_r, compares the number of true positives (TP) with the total number of positives in the top r ranks:

$$P_r = \frac{\text{TP}}{\text{TP+FP}}, \tag{6.5}$$

where FP is the number of false positives. As we only have one gallery image per query image, the mAP is calculated via

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{r_q}, \tag{6.6}$$

where the correct identity of q is found at rank r_q.

All experiments were performed using the torchreid framework [290] on an NVIDIA RTX 2080 Ti GPU.

6.3.3 Results

Results on the 'Temporally-Near' and 'Big-to-Small' settings can be found in Table 22 and 23, respectively. ViT Base with CE+Triplet loss comprehensively outperforms all other methods on the Big-to-Small setting, and has third highest mAP on the Temporally-Near setting. From Figure 31, rows two and three, we observe that ViT returns a similar ranking list on query UAVs that have different colour. We posit that ViT is better at capturing shape information due its global self-attention mechanism, compared to convolutional methods that rely on a local receptive field. Similar to ViT, PCB also splits the input image into parts and performs consistently well across both tasks. This indicates that a part-based strategy can be effective for UAV re-ID.

For Temporally-Near, the best rank-1 matching rate of 77.4% from generic architectures such as ResNet-18 and ViT is a strong baseline, but all models achieve respectable performance. Interestingly, the best performance with a triplet loss is obtained by HACNN, another attention framework, though it seems to struggle with a cross-entropy loss, and its triplet performance does not match ResNet-18, ViT Base or PCB. Nevertheless, when larger data sets are collected, triplet loss may become more viable; HACNN would appear to be a strong candidate if that is that case. For real-world tracking systems, re-ID is performed with only a few possible matches, rather than the entire test data set. These methods should therefore be strong enough to be used in real-world systems immediately.

As expected, Big-to-Small is more challenging than Temporally-Near due to the extreme variation in scale. Big-to-Small has top rank-1 and rank-5 matching rates of just 35.5% and 54.8%, respectively. ViT again performs very well here compared to other methods with ViT small getting the best performance with cross-entropy loss, and Vit Base achieving top performance overall. Once again HACNN appears to seriously struggle with a cross-entropy loss. Although ViT demonstrates potential, this setting requires further research to develop UAV-specific architectures to match objects across pose and scale. Across both tables, the only re-ID specific framework that appears to perform well is PCB. This suggests that re-ID frameworks are over-engineered for the specific task of re-identifying humans and entirely new models would need to be designed for UAV re-ID.

The general observations from Tables 22 and 23 are that the re-ID-specific networks generally do

Table 24: Methods Tested Using the Not 3D Re-ID framework [18]

Model	Temporally-Near			Big-to-Small		
	mAP	rank-1	rank-5	mAP	rank-1	rank-5
ResNet-18	74.3	67.7	71.0	36.4	25.8	29.0
ResNet-34	70.1	64.5	67.7	37.8	29.0	32.3
ResNet-50	79.5	*74.2*	77.4	38.5	29.0	32.3
SE-ResNet-50	72.1	64.5	71.0	40.2	*32.3*	35.5
SE-ResNeXt-50	72.0	67.7	67.7	39.4	29.0	35.5
ViT Small	79.2	71.0	77.4	39.6	29.0	32.3
ViT Base	77.0	71.0	77.4	41.6	29.0	38.7
ResNet50mid	78.7	71.0	77.4	45.6	35.5	41.9
OSNet	81.5	77.4	80.7	35.2	22.6	29.0
MLFN	74.3	67.7	71.0	40.8	*32.3*	41.9
PCB	*80.5*	*74.2*	80.7	39.3	29.0	32.3
HACNN	74.1	67.7	74.2	41.6	*32.3*	35.5
IBN-A	72.0	64.5	67.7	*41.9*	*32.3*	35.5

Red denotes the (joint) highest in each column, blue denotes the (joint) second highest in each column

not perform as well as generic networks. One reason for this is that extensive hyperparameter tuning is performed on generic networks to maximise performance on ImageNet. ReID-specific networks, although pre-trained on ImageNet, tune hyperparameters to maximise performance on re-ID data sets, so have less functional ability to be applied to new tasks. However, PCB, which uses a ResNet-50 backbone (optimised for ImageNet), does still attain strong performance. Furthermore, in almost all cases, cross-entropy loss performance exceeds triplet loss. Further, the combined loss is occasionally unable to yield higher performance than cross-entropy alone. It is a common occurrence however, that triplet loss performance improves as the number of classes within the data set increases. Furthermore, because UAV-reID only allows one-to-one matching, we cannot harness the power of hard-positive mining. We expect that triplet loss will generate better results, and perhaps exceed cross-entropy, when a more comprehensive data set is made available.

The results from the Not-3D ReID framework (Table 24) corroborate our findings. Indeed, the additional loss functions incorporated into one Multi-Loss aggregation function are generally unable to improve results, perhaps owing again to the lack of effective hard-positive mining and few available classes. In this regard, we can firstly observe that the IBN-A network does not outperform the other networks in either challenge. Secondly, the Not-3d ReID network is only able to improve upon ResNet50 and ViT Small generic re-identification networks for Temporally-Near,

but yields consistently stronger results for the re-ID specific networks. Compared to earlier results, OSNet shows far stronger performance with the Not-3D loss on Temporally-Near, getting the highest performance on all metrics with this loss, and pretty comparable with ResNet-18 with cross-entropy loss. Re-ID specific frameworks ResNet50mid, OSNet, MLFN, HACNN all improve performance on the Big-to-Small data set, while ViT and ResNet models appear to perform worse. Another interesting insight is that networks that achieve good results on the Temporally-Near challenge are not necessarily well-suited for the Big-to-Small challenge; the best performing networks for Temporally-Near (OSNet, PCB, ResNet50) are disjoint from those suited to Big-to-Small (ResNet50-mid, Vit Base, MLFN).

One general observation from the presented results is that the networks designed for person re-ID mostly do not translate well to UAV re-ID, with the exception of PCB. However, incorporating more modern combinations of loss functions, like in the Not-3D ReID framework can help to alleviate this. This suggests that the additional losses used in that framework interact well with re-ID specific models. Overall, ViT Base with the combined loss from Equation 6.3 performs best across the two tasks, getting the highest score on Big-to-Small, and the second highest score on Temporally Near. Interestingly, PCB is also a strong performer, but slightly less robust on the Big-to-Small task. It is worth noting that as part of their respective frameworks, both ViT and PCB split the image up into parts before performing further processing. These results indicate that this may be important step for robust modelling of dynamic objects such as UAVs.

6.3.4 Interpretability

As well as the general robustness that vision transformers exhibit, one major advantage is that they allow us to inspect the attention maps to understand how they reach a decision. The vision transformers we have used consist of twelve attention heads, which attend to different parts of the image. Because our UAVs are pre-cropped, most images take the form of foreground object on background image. Because there is not too much relevant information to attend to, many attention heads are similar to each other. In this section, we present interesting visualisations which illuminate the inner workings of vision transformers for re-ID.

Vision transformers split the images into 16×16 image patches, and the relative importance between image patches is learnt. Vision transformers also add a classification (CLS) token, which

Figure 32: Attention visualisation of the transformer mechanism within ViT on the Big-to-Small setting. Attention from the CLS token is presented, i.e. indicating the global importance of image regions. Different attention heads attend to different parts of the image, forming a more robust feature representation.

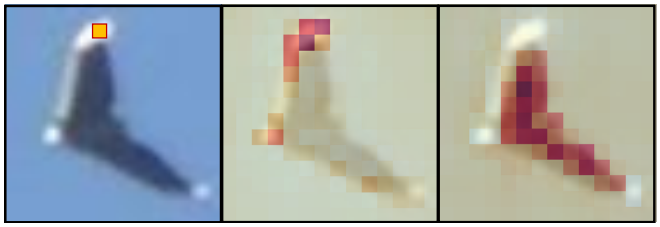


Figure 33: Attention visualisation of the transformer mechanism within ViT on the Big-to-Small setting. Attention from the 16×16 yellow image patch is presented. Attention can highlight salient image patches relevant to the selected image patch.

attends to the entire image. Figure 32 is a visualisation of four different attention heads of the CLS token.

The first attention map attends to the entire UAV, the second attends to its legs, the third to the propellers and the top of the UAV. This demonstrates clearly how it is encoding features and what the final feature representation consists of. The fourth attention map isolates the background. Even though the background is complicated, the attention head identifies that the drone is the foreground object, and considers the clouds and the trees together. This gives confidence that ViT has a good understanding of the image, and that the feature representation is composed in a structurally sound manner.

Figures 33 and 34 visualise attention from specific image patches, which are indicated via the yellow box with red border on the leftmost image. In Figure 33, one attention head identifies that the selected image patch is on a white stripe of the UAV and attends heavily to this stripe. This is a strongly identifiable feature which can be used for re-identification to help differentiate between other UAVs of a similar shape. The second attention head shown also identifies that the selected image patch belongs to the foreground object, so attends heavily to the rest of the UAV.

Figure 34: Attention visualisation of the transformer mechanism within ViT on the Big-to-Small setting. Attention from the 16×16 yellow image patch is presented. Attention can highlight salient image patches relevant to the selected image patch.

Figure 34 presents additional examples to further demonstrate the understanding that ViT generates. On the left, the query patch occurs on the UAV, and the resulting attention strongly segments the UAV from the background. On the right, the query patch occurs on one of the propellers, and the attention head attends to each of the other propellers. One of the advantages of transformers over traditional convolution is being able to learn non-local relationships between image patches to obtain a stronger feature representation. These visualisations demonstrate this process in action.

6.4 Conclusion

We have performed a benchmark study to use deep learning techniques for UAV re-identification. Vision transformers consistently achieve strong performance across both tasks, significantly outperforming all convolutional methods at matching large UAV detections with small ones. A range of methods can re-identify UAVs over a short time period with high precision. Of these methods, ResNet-18 (mAP 81.9) appears to be easiest to fit into tracking frameworks due to its high performance and relatively small model size.

Although the Big-to-Small setting is very challenging, vision transformers have a lot of potential to handle extreme scale transformation. Vision transformers replace the typical CNN architecture with self-attention layers, which appear to be more robust to significant changes in scale. Furthermore, they allows us to visualise the interactions that are occurring, which makes the model more interpretable, so we can understand how the feature representation is generated. This helps to solve both the robustness and interpretability problems that are commonly seen with DNNs.

Our future work involves the collecting a large multi-view UAV re-ID data set, and developing an

improved vision transformer by incorporating techniques used in convolutional neural networks to handle scale changes, such as concatenating outputs from different layers.

Chapter 7

Conclusion

This thesis has addressed the utilisation of attention mechanisms to improve interpretability and robustness of deep neural networks, to improve the quality of learnt feature representations. Despite the significant paradigm shift towards deep networks in research, real-world uptake has been slower because supervised deep networks are less robust to lower-quality data and models can be deemed untrustworthy. The aim of this thesis was to demonstrate the ability of attention mechanisms to bridge this gap between theory and practice.

This thesis has explored the impact that attention mechanisms can make when learning representations over a multitude of applications. Attention mechanisms have been tested within deep neural networks (DNNs), convolutional neural networks (CNNs), and Generative Adversarial Networks (GANs). Attention mechanisms prove to be highly flexible, improving performance regardless of the base framework within which they are applied. Furthermore, the idea of replacing convolution with attention in the form of Vision Transformers (ViTs) has been explored [172]. ViT gives the benefits of attention mechanisms while also encoding non-local relationships between image regions.

Attention has been shown to improve robustness of models in a variety of difficult conditions. By identifying important features, data perturbations are relatively less impactful on models with attention compared to other models [175]. In particular, other models have a tendency to favour the dominant class as the level of the attack increases, whereas models with attention can still extract relevant information from perturbed data. Attention has also shown its ability when working with

low-quality data, including low-resolution images [171] and extreme scale changes [172]. Finally, attention models exceed other models at out-of-distribution generalisation, which is most needed for DNNs to be applied in the real world.

Furthermore, by inspecting saliency heatmaps produced by attention mechanisms, the models can be interpreted to improve trust in the deep learning model. Again, this property is irrespective of the framework within which the attention mechanism is applied. In DNNs, when working on only tabular data, the attention map shows interactivity between different features whilst on image data, the attention map shows salient regions that the models believes is important. This is even the case within style transfer. In [171], we optimise eight neural networks simultaneous, and still obtain visualisations to see how the makeup style is encoded. This interpretability strongly helps to facilitate trust in DNNs.

Empirically, we have demonstrated the general effectiveness of attention to build more robust and interpretable deep neural networks, without sacrificing performance, to facilitate the transition of deep learning into the real world. Importantly, attention mechanisms take minimal effort to incorporate into current state-of-the-art learning frameworks. In the introduction, Roberts *et al.* [197] was discussed, where 2212 studies of AI models to assist with the covid-19 pandemic were analysed, but precisely zero proposed models were suitable for real-world deployment. Two major reasons for this were a lack of robustness and interpretability. This thesis gives evidence that attention mechanisms can significantly reduce the gap between research of DNNs and real-world deployment of DNNs. As DNNs have frequently been shown to outperform humans on individual tasks, this is an incredibly important step to take.

The detailed novel contributions of this thesis are outlined in the following section.

7.1 Thesis Contributions

- A novel fused attention network was proposed for the task of schizophrenia diagnosis [175]. Channel attention was incorporated into the robustness pathway to improve the ability of the network to handle perturbations of data. Self-attention was utilised in the interpretability pathway to allow visualisation of the interaction between features to understand what values within a data sample contribute most to the diagnosis recommended by the network.

We designed twelve stress tests to evaluate the robustness of compared models to noise and missing values. We found that channel attention contributes significantly to ensuring the model remains robust, by focusing the network on the most important features and minimising the weight of spurious features.

- A novel multi-scale attention mechanism was designed to perform makeup style transfer on low-quality data [171]. Whereas most state-of-the-art methods can only handle carefully constructed, high-quality data sets, our attention mechanism allows the model to be robust enough to handle real-world data. In particular, a feature representation of the makeup style can be robustly encoded. The attention map within the encoder can then be visualised, making the encoder interpretable. A novel quantitative evaluation is proposed to accompany the qualitative evaluation, and the proposed multi-scale Augmentated CycleGAN outperforms state of the art.

- Extensive experiments have been conducted on re-identification. The discriminative power of channel attention has been demonstrated for person re-identification [173]. To further examine robustness of different models, the task of unified person and vehicle re-identification was proposed by creating the PVUD data set [174]. A novel loss function is introduced, with individual terms that attend specifically to positive and negative pairs, respectively. Lastly, the performance of different methods for out-of-distribution generalisation was examined in order to evaluate how robust models are when test data comes from a different source than the training data. Notably, the proposed model with channel attention outperforms other state-of-the-art methods without attention. As real-world data is too voluminous to label everything in order for effective training, the task of generalising out of distribution is essential. The experiments presented signal that attention is an essential mechanism to include to better generalise to real-world data.

- The task of unmanned aerial vehicle (UAV) re-identification has been proposed [172]. Two scenarios, 'Temporally-Near' and 'Big-to-Small', were studied. Temporally-Near simulates the common task of linking two UAVs in nearby frames of a video sequence to assist tracking software, whereas Big-to-Small simulates the task of matching far away UAVs to a library of known UAVs, in order to identify potential threats. A variety of models were evaluated to identify the best models for each task. We find that attention-based models are

much more robust to extreme variations in scale, with Vision Transformers (where convolution is replaced with attention) performing particularly well. Vision Transformers also allow us to visualise the inner workings of the model, to understand how feature representations are generated. This allows the model to be interpretable, which gives much more trust on the outcome of the model.

7.2 Limitations and Future Work

- The primary limitation of the work on schizophrenia is the lack of data samples and sources. The planned future work is collect more data from different clinicians, to evaluate if the proposed model continues to outperform others on a more varied data set. Collecting data from different clinicians, and hopefully clinicians from different countries, would allow the exploration of out-of-distribution generalisation. The presented experiments aim to simulate this generalisation capability, but having the explicit data from different sources would allow stronger evaluation of robustness of all models. Fundamentally, even with proposed stress tests, data from multiple sources needs to be considered before a model can be used in practice. The experiments on re-identification indicate that attention helps models to be robust to data from a new source, which, along with the stress test results, gives confidence that the proposed model will be able to handle this more challenging scenario.

  Furthermore, the current data set only deals with observations that clinicians have recorded. This is in line with the 5^{th} edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Research Domain Criteria (RDoC) is a more modern methodology for evaluating mental health issues, which takes advantage of brain scan images to better understand the underlying neurobiology. Currently, RDoC is not able to serve as a diagnostic guide, so we wish to devise a multi-modal model which considers clinical observations and brain scans to combine the advantages of both systems. Again, attention will play a core role to ensure the effectiveness of multi-modal models, as it allows clinical data to attend to image data and vice versa, resulting in a model that has a better understanding of the entire picture, rather than a model which considers these separately. Attention also allows visualisation of the important parts of the brain scan, as in Chapter 6, making a more interpretable, trustworthy model.

- One-to-many augmentation is a well-established method to improve the robustness of face verification models when presented with face images from different poses [215]. Because our multi-scale Augmented CycleGAN outperforms state-of-the-art makeup style transfer augmentation models, it is a natural extension to synthesise many makeup images for each non-makeup image, to apply one-to-many augmentation for makeup-invariant face verification. Furthermore, in contrast to other state-of-the-art makeup style transfer models, the proposed model is cyclic, so makeup removal can also easily be performed. This means that a many-to-one normalisation strategy can also be utilised to further improve makeup-invariant face verification performance.

- Our experiments demonstrated that vision transformers were extremely successful at matching across scale, even more successful than convolutional models designed to be robust to scale changes. It is therefore natural to further explore the potential of vision transformers to handle this task. Several principles have been designed to improve the ability of convolutional models to be more robust to scale changes, such as and feature pyramid representations [119] exploring different receptive fields [291]. By amending vision transformers in these ways, we wish to obtain similar improvements.

  We also wish to collect a larger UAV re-identification data set on which to perform evaluations. As our current data set is relatively small, the triplet loss does not have much effect on improving the performance, but we wish to evaluate if this is still the case as more classes are added and the problem becomes more fine-grained.

- Although this thesis has demonstrated the ability of attention mechanisms to improve model interpretability and robustness, there are many modelling decisions that have to be made. For example, attention mechanisms have demonstrated their ability to be interpretable; however, this has been viewed through a binary lens. There may be more expressive interpretable mechanisms that can be used instead of, or alongside, attention mechanisms. Similarly, framework-based changes can also benefit interpretability and robustness. For example, one of the current state-of-the-art family of frameworks is Normaliser-Free Networks (NFNets) [19], a framework designed purely around removing batch normalisation. These frameworks are more robust than others because batch normalisation forces the model to behave differently during training compared to at inference. On the other hand, spectral

normalisation [158] resulted in much more robust GANs because of their better training performance. Changing the loss function can also make a large difference. In Chapter 6, re-ID frameworks mostly improved performance with a modern re-ID loss function while generic frameworks suffered. Future work involves exploring framework-level modifications that can be made for improved interpretability and robustness. One particularly promising direction is modifying frameworks to incorporate retrieval. Retrieval mechanisms are inherently interpretable, and have already been shown to improve performance on generative tasks [17]. Retrieval models are well-studied themselves, particularly in the context of challenges were robustness is key, such as zero-shot learning [147]. However, it is unclear whether there has been any works adding retrieval to a discriminative model, which would require a major framework change, as in [17].

- Attention mechanisms have empirically shown to improve interpretability, and robustness. A major research question that remains is whether this is a coincidence or if these are inherently linked. Intuitively, it makes sense that attention mechanisms discover important features resulting in a more robust model, and the result attention maps are interpretable. An example of this mutually beneficial relationship is chain of thought prompting [247, 112], where a model is asked a simple maths question in a zero-shot setting and is unable to answer correctly. When the model is told 'Lets think step by step', it shows its working in an interpretable fashion and is able to answer the zero-shot question. Beyond intuition, a future work is to layout a theoretical framework which underpins these empirical observations. This theoretical framework would give more confidence of the generalisability of attention mechanisms, and help them to become more commonplace in real-world applications.

- Overall, attention mechanisms have demonstrated strong ability to improve the interpretability of neural networks, and to make neural networks much more robust, facilitating their uptake into the real world. However, they do not completely solve the problem, as it is impossible to learn the underlying structural relationship between all features given a finite set of data. To this end, a major future challenge to further the work done in this thesis is to combine attention mechanisms with elements of causal learning. One example is causal graph discovery [69] where the underlying causal relationship between features needs to be learnt. The self-attention mechanism, where every feature attends to every other feature in

an asymmetric manner, is a sensible mechanism to explore in order to improve the ability discover these causal relationships.

- Out-of-distribution generalisation remains a major challenge. Our experiments in Chapter 5 show that attention can improve the ability of models to be robust to data from a new source. However, even though attention improves performance, the performance remains relatively low compared to data from the same source. Again, causal learning appears to be an appropriate next step to combine with attention to go further with extracting essential features and minimising the influence of superfluous ones.

References

[1] Towards trustable machine learning. *Nature Biomedical Engineering*, 2(10):709–710, 2018.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

[3] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cycleGAN: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

[4] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-ID done right: towards good practices for person re-identification. 2018.

[5] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[6] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. 2013.

[7] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[9] Song Bai and Xiang Bai. Sparse contextual activation for efficient visual re-ranking. *IEEE Transactions on Image Processing*, 2016.

[10] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Lingyu Duan. Group Sensitive Triplet Embedding for Vehicle Re-identification. *IEEE Transactions on Multimedia*, 9210(c):1–14, 2018.

[11] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[12] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[13] Farah Benamara, Véronique Moriceau, Josiane Mothe, Faneva Ramiandrisoa, and Zhaolong He. Automatic detection of depressive users in social media. In *Conférence francophone en Recherche d'Information et Applications (CORIA)*, 2018.

[14] Daniel S Berman, Anna L Buczak, Jeffrey S Chavis, and Cherita L Corbett. A survey of deep learning methods for cyber security. *Information*, 10(4):122, 2019.

[15] Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet, and Sebastien Martin. The price of interpretability. *arXiv preprint arXiv:1907.03419*, 2019.

[16] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022.

[17] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR, 2022.

[18] Toby Breckon and Aishah Alsehaim. Not 3d re-id: Simple single stream 2d convolution for robust video re-identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5190–5197, 01 2021.

[19] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR, 2021.

[20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[21] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018.

[22] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[23] Jonathan K Burns. The social determinants of schizophrenia: an african journey in social epidemiology. *Public Health Reviews*, 34(2):8, 2012.

[24] Ovidiu Calin. *Deep learning architectures*. Springer, 2020.

[25] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[26] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[27] Christopher F Chabris and Daniel J Simons. *The invisible gorilla: And other ways our intuitions deceive us*. Harmony, 2010.

[28] H. Chang, J. Lu, F. Yu, and A. Finkelstein. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 40–48, June 2018.

[29] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[30] H. Chen, K. Hui, S. Wang, L. Tsao, H. Shuai, and W. Cheng. Beautyglow: On-demand makeup transfer framework with reversible generative network. In *2019 IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition (CVPR), pages 10034–10042, June 2019.

[31] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[33] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] Wenbai Chen, Yue Lu, Hang Ma, Qili Chen, Xibao Wu, and Peiliang Wu. Self-attention mechanism in person re-identification models. *Multimedia Tools and Applications*, pages 1–19, 2021.

[35] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[36] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2016.

[37] B. L. P. Cheung and D. Dahl. Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 222–225, March 2018.

[38] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett,

editors, *Advances in Neural Information Processing Systems 29*, pages 3504–3512. Curran Associates, Inc., 2016.

[39] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, Mar 2020.

[40] Angelo Coluccia, Alessio Fascista, Arne Schumann, Lars Sommer, Marian Ghenescu, Tomas Piatrik, Geert De Cubber, Mrunalini Nalamati, Ankit Kapoor, Muhammad Saqib, et al. Drone-vs-bird detection challenge at ieee avss2019. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2019.

[41] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018.

[42] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[43] Celine Craye and Salem Ardjoune. Spatio-temporal semantic segmentation for drone detection. In *2019 16th IEEE International conference on advanced video and signal based surveillance (AVSS)*, pages 1–5. IEEE, 2019.

[44] Bruce N Cuthbert et al. The RDoC framework: continuing commentary. *World Psychiatry*, 13(2):196, 2014.

[45] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[47] Jianhua Deng, Muhammad Saddam Khokhar, Muhammad Umar Aftab, Jingye Cai, Rajesh Kumar, Jay Kumar, et al. Trends in vehicle re-identification past, present, and future: A comprehensive review. *arXiv preprint arXiv:2102.09744*, 2021.

[48] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[49] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[51] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1):1–19, 2020.

[52] Dong Guo and T. Sim. Digital face makeup by example. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–79, June 2009.

[53] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[54] Trafton Drew, Melissa L-H V˜o, and Jeremy M Wolfe. The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological science*, 24(9):1848–1853, 2013.

[55] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *ICLR*, 2017.

[56] Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. Facebook

language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[57] Kenneth Eriksson, Donald Estep, and Claes Johnson. Lipschitz continuity. In *Applied Mathematics: Body and Soul*, pages 149–164. Springer, 2004.

[58] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018.

[59] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, 2019.

[60] Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *arXiv preprint arXiv:2006.05690*, 2020.

[61] E. A. Garcia and H. He. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 21(09):1263–1284, sep 2009.

[62] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[63] Robert Geirhos, David H. J. Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker, 2018.

[64] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems 31*, pages 7538–7550. 2018.

[65] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep Transfer Learning for Person Re-identification. 2016.

[66] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[67] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

[68] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[69] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[70] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019.

[71] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

[72] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[73] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[74] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 262–275, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[75] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep

learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5):1153–1159, 2016.

[76] Aleksei Grigorev, Zhihong Tian, Seungmin Rho, Jianxin Xiong, Shaohui Liu, and Feng Jiang. Deep person re-identification in UAV images. *EURASIP Journal on Advances in Signal Processing*, 2019(1):54, 2019.

[77] S Gulsuner, DJ Stein, ES Susser, G Sibeko, A Pretorius, T Walsh, L Majara, MM Mndini, SG Mqulwana, OA Ntola, et al. Genetics of schizophrenia in the south african xhosa. *Science*, 367(6477):569–573, 2020.

[78] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR, 2020.

[79] Heinz Häfner and Kurt Maurer. Early detection of schizophrenia: current evidence and future perspectives. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, 5(3):130–138, Oct 2006. 17139339[pmid].

[80] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[81] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[82] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021.

[83] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[84] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework.

[85] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[86] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[87] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015.

[88] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[89] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[90] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[91] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 2439–2448, 2017.

[92] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[93] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018.

[94] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[95] Brian K. S. Isaac-Medina, Matthew Poyser, Daniel Organisciak, Chris G. Willcocks, Toby P. Breckon, and Hubert P. H. Shum. Unmanned aerial vehicle visual detection and tracking using deep neural networks: A performance benchmark. 2021.

[96] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[97] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2020.

[98] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

[99] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635, 2021.

[100] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[101] Minyue Jiang, Yuan Yuan, and Qi Wang. Self-attention learning for person re-identification. In *BMVC*, page 204, 2018.

[102] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-UAV: A large multi-modal benchmark for UAV tracking, 2021.

[103] Sunil Vasu Kalmady, Russell Greiner, Rimjhim Agrawal, Venkataram Shivakumar, Janardhanan C. Narayanaswamy, Matthew R. G. Brown, Andrew J. Greenshaw, Serdar M. Dursun, and Ganesan Venkatasubramanian. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *npj Schizophrenia*, 5(1):2, 2019.

[104] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

[105] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[106] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[107] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, Oct 2019.

[108] Brigitte Khoury, Cary Kogan, and Sariah Daouk. *International Classification of Diseases 11th Edition (ICD-11)*, pages 1–6. Springer International Publishing, Cham, 2017.

[109] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[110] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[111] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[112] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[113] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[114] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[115] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019.

[116] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[117] Chiman Kwan, Bryan Chou, and Li-Yun Martin Kwan. A comparative study of conventional and deep learning target tracking algorithms for low quality videos. In *International Symposium on Neural Networks*, pages 521–531. Springer, 2018.

[118] Mahmood Ashoori Lalimi, Sedigheh Ghofrani, and Des McLernon. A vehicle license plate detection method using region and edge based methods. *Computers & Electrical Engineering*, 39(3):834 – 845, 2013. Special issue on Image and Video Processing Special issue on Recent Trends in Communications and Signal Processing.

[119] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 536–552, 2018.

[120] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pages 189–248. Springer, 2016.

[121] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[122] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[123] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[124] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[125] Mi Hyun Lee, Nambeom Kim, Jaeeun Yoo, Hang-Keun Kim, Young-Don Son, Young-Bo Kim, Seong Min Oh, Soohyun Kim, Hayoung Lee, Jeong Eun Jeon, et al. Multitask fmri and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder. *Scientific reports*, 11(1):1–13, 2021.

[126] Tyler A Lesh, Costin Tanase, Benjamin R Geib, Tara A Niendam, Jong H Yoon, Michael J Minzenberg, J Daniel Ragland, Marjorie Solomon, and Cameron S Carter. A multimodal analysis of antipsychotic effects on brain structure and function in first-episode schizophrenia. *JAMA psychiatry*, 72(3):226–234, 2015.

[127] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[128] C. Li, K. Zhou, and S. Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4621–4629, June 2015.

[129] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

[130] Jianshu Li, Chao Xiong, Luoqi Liu, Xiangbo Shu, and Shuicheng Yan. Deep face beautification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 793–794, 2015.

[131] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 645–653, New York, NY, USA, 2018. Association for Computing Machinery.

[132] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[133] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2194–2200. AAAI Press, 2017.

[134] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-

identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[135] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[136] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification. pages 1–13, 2018.

[137] Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2):146–160, 1976.

[138] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[139] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[140] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

[141] Mingjie Liu, Xianhao Wang, Anjian Zhou, Xiuyuan Fu, Yiwei Ma, and Changhao Piao. Uav-yolo: Small object detection on unmanned aerial vehicle perspective. *Sensors (Basel)*, 20(8):2238, 2020.

[142] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. Makeup like a superstar: Deep localized makeup transfer network. *2016 International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[143] Wu Liu, Xinchen Liu, Huadomg Ma, and Peng Cheng. Beyond Human-level License Plate

Super-resolution with Progressive Vehicle Search and Domain Priori GAN. *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*, pages 1618–1626, 2017.

[144] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. *Proceedings - IEEE International Conference on Multimedia and Expo*, 2016-August, 2016.

[145] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 869–884, Cham, 2016. Springer International Publishing.

[146] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018.

[147] Yang Long, Li Liu, Yuming Shen, and Ling Shao. Towards affordable semantic searching: Zero-shot retrieval via dominant attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[148] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[149] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[150] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108–119, 2020.

[151] Sumaya Mall, Jonathan M Platt, Henk Temmingh, Eustasius Musenge, Megan Campbell, Ezra Susser, and Dan J Stein. The relationship between childhood trauma and schizophrenia in the genomics of schizophrenia in the xhosa people (sax) study in south africa. *Psychological medicine*, 50(9):1570, 2020.

[152] SL Matheson, Alana M Shepherd, RM Pinchbeck, KR Laurens, and Vaughn J Carr. Childhood adversity in schizophrenia: a systematic meta-analysis. *Psychological medicine*, 43(2):225, 2013.

[153] T. Matsubara, T. Tashiro, and K. Uehara. Deep neural generative model of functional mri images for psychiatric disorder diagnosis. *IEEE Transactions on Biomedical Engineering*, 66(10):2768–2779, 2019.

[154] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[155] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020.

[156] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[157] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[158] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[159] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[160] Francisco J Moreno-Barea, Fiammetta Strazzera, José M Jerez, Daniel Urda, and Leonardo Franco. Forward noise adjustment scheme for data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*, pages 728–734. IEEE, 2018.

[161] Tiago S Nazaré, Gabriel B Paranhos da Costa, Welinton A Contato, and Moacir Ponti. Deep convolutional neural networks and noisy images. In *Iberoamerican Congress on Pattern Recognition*, pages 416–424. Springer, 2017.

[162] Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1127–1131. IEEE, 2021.

[163] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr : a convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30, 2016.

[164] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[165] Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the future - big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216–1219, Sep 2016. 27682033[pmid].

[166] Jihoon Oh, Baek-Lok Oh, Kyong-Uk Lee, Jeong-Ho Chae, and Kyongsik Yun. Identifying schizophrenia using structural mri with a deep learning algorithm. *Frontiers in Psychiatry*, 11:16, 2020.

[167] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

[168] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[169] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Bier - boosting independent embeddings robustly. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[170] Roberto Opromolla, Giuseppe Inchingolo, and Giancarmine Fasano. Airborne visual detection and tracking of cooperative uavs exploiting deep learning. *Sensors*, 19(19), 2019.

[171] Daniel Organisciak, Edmond S. L. Ho, and Hubert P. H. Shum. Makeup style transfer on low-quality images with weighted multi-scale attention. In *Proceedings of the 2020 International Conference on Pattern Recognition*, ICPR '20, 2020.

[172] Daniel Organisciak, Brian K. S. Isaac-Medina, Matthew Poyser, Shanfeng Hu, Toby P. Breckon, and Hubert P. H. Shum. UAV-ReID: A benchmark on unmanned aerial vehicle re-identification, 2021.

[173] Daniel Organisciak, Chirine Riachy, Nauman Aslam, and Hubert Shum. Triplet loss with channel attention for person re-identification. *Journal of WSCG*, 27, 01 2019.

[174] Daniel Organisciak, Dimitrios Sakkos, Edmond SL Ho, Nauman Aslam, and Hubert P H Shum. Unifying person and vehicle re-identification. *IEEE Access*, 8:115673–115684, 2020.

[175] Daniel Organisciak, Hubert PH Shum, Ephraim Nwoye, and Wai Lok Woo. Robin: A robust interpretable deep network for schizophrenia diagnosis. *Expert Systems with Applications*, 201:117158, 2022.

[176] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

[177] Lena Palaniyappan, Gopikrishna Deshpande, Pradyumna Lanka, D Rangaprakash, Sarina Iwabuchi, Susan Francis, and Peter F Liddle. Effective connectivity within a triple network brain system discriminates schizophrenia spectrum disorders from psychotic bipolar disorder at the single-subject level. *Schizophrenia research*, 214:24–33, 2019.

[178] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[179] Magdalini Paschali, Muhammad Ferjad Naeem, Walter Simson, Katja Steiger, Martin Mollenhauer, and Nassir Navab. Deep learning under the microscope: Improving the interpretability of medical imaging neural networks. *arXiv preprint arXiv:1904.03127*, 2019.

[180] Meenal J Patel, Alexander Khalaf, and Howard J Aizenstein. Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*, 10:115–123, 2016.

[181] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13893–13894. AAAI Press, 2020.

[182] Judea Pearl. *Causality*. Cambridge university press, 2009.

[183] Irene Petersen, Catherine A Welch, Irwin Nazareth, Kate Walters, Louise Marston, Richard W Morris, James R Carpenter, Tim P Morris, and Tra My Pham. Health indicator recording in uk primary care electronic health records: key implications for handling missing data. *Clinical epidemiology*, 11:157, 2019.

[184] Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.

[185] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and SS Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.

[186] J. Qi and J. Tejedor. Deep multi-view representation learning for multi-modal features of the schizophrenia and schizo-affective disorder. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 952–956, March 2016.

[187] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:3–20, 2016.

[188] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[189] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[190] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[191] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

[192] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, volume 32, pages 68–80. Curran Associates, Inc., 2019.

[193] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2020.

[194] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Yang. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, Jan 2017.

[195] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paix˜ao, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Mac-

farlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1–9, 2020.

[196] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[197] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[198] Peter M. Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof. *Mahalanobis Distance Learning for Person Re-identification*, pages 247–267. Springer London, London, 2014.

[199] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

[200] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021.

[201] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[202] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[203] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Mkadry. How does batch normalization help optimization? In *Proceedings of the 32nd international conference on neural information processing systems*, pages 2488–2498, 2018.

[204] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[205] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[206] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[207] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016.

[208] Shiguang Shan, Yizheng Chang, Wen Gao, Bo Cao, and Peng Yang. Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 314–320. IEEE, 2004.

[209] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.

[210] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey, 2021.

[211] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, 22(1):173–183, Jan 2018.

[212] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[213] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[214] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[215] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[216] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

[217] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[218] Brandon M Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491, 2013.

[219] Ian Soboroff. Computing confidence intervals for common ir measures. In *EVIA@ NTCIR*. Citeseer, 2014.

[220] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[221] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[222] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[223] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)*, September 2018.

[224] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[225] Christian Szegedy, V. Vanhoucke, S. Ioffe, Jonathon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[226] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[227] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[228] Neeraj Tandon and Rajiv Tandon. Machine learning in psychiatry-standards and guidelines. *Asian journal of psychiatry*, 44:A1, 2019.

[229] Neeraj Tandon and Rajiv Tandon. Using machine learning to explain the heterogeneity of schizophrenia. realizing the promise and avoiding the hype. *Schizophrenia Research*, 214:70–75, 2019.

[230] Hao Tang, Hong Liu, Dan Xu, Philip H. S. Torr, and Nicu Sebe. AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks, 2019.

[231] Shangzhi Teng, Shiliang Zhang, Qingming Huang, and Nicu Sebe. Viewpoint and scale consistency reinforcement for UAV vehicle re-identification. *International Journal of Computer Vision*, 129(3):719–735, 2021.

[232] Hui Tian, Xiang Zhang, Long Lan, and Zhigang Luo. Person re-identification via adaptive verification loss. *Neurocomputing*, 359:93 – 101, 2019.

[233] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

[234] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[235] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. accepted as poster.

[236] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, pages 1–15, 2019.

[237] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[238] Michael Wainberg, Daniele Merico, Andrew Delong, and Brendan J Frey. Deep learning in biomedicine. *Nature biotechnology*, 36(9):829–838, 2018.

[239] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[240] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 274–282, New York, NY, USA, 2018. ACM.

[241] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. *arXiv preprint arXiv:2003.07853*, 2020.

[242] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. *Pro-*

ceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1386–1393, 2014.

[243] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 2020.

[244] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[245] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. Resource aware person re-identification across mutliple grane resolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[246] Ziyang Wang, Dan Wei, Xiaoqiang Hu, and Yiping Luo. Human skeleton mutual learning for person re-identification. *Neurocomputing*, 388:309–323, 2020.

[247] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[248] Longhui Wei, Xiaobin Liu, Jianing Li, and Shiliang Zhang. VP-ReID: Vehicle and person re-identification system. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, page 501–504, New York, NY, USA, 2018. Association for Computing Machinery.

[249] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[250] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

[251] Steven Euijong Whang and Jae-Gil Lee. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, 13(12):3429–3432, 2020.

[252] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[253] Julie L Winterburn, Aristotle N Voineskos, Gabriel A Devenyi, Eric Plitman, Camilo de la Fuente-Sandoval, Nikhil Bhagwat, Ariel Graff-Guerrero, Jo Knight, and M Mallar Chakravarty. Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? a multi-method and multi-dataset study. *Schizophrenia Research*, 214:3–10, 2019.

[254] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017.

[255] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[256] World Health Organization. WHO methods and data sources for global burden of disease estimates 2000-2011. 2011.

[257] Di Wu, Si-Jia Zheng, Wen-Zheng Bao, Xiao-Ping Zhang, Chang-An Yuan, and De-Shuang Huang. A novel deep model with multi-loss and efficient training for person re-identification. *Neurocomputing*, 324:69 – 75, 2019. Deep Learning for Biological/Clinical Data.

[258] Xifang Wu, Songlin Sun, and Meixia Fu. Person re-identification based on semantic segmentation. In Yue Wang, Meixia Fu, Lexi Xu, and Jiaqi Zou, editors, *Signal and Information Processing, Networking and Computers*, pages 903–909, Singapore, 2020. Springer Singapore.

[259] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda,

Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[260] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[261] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-Aware Compositional Network for Person Re-identification. pages 2119–2128, 2018.

[262] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[263] L. Xu, Y. Du, and Y. Zhang. An automatic framework for example-based virtual makeup. In *2013 IEEE International Conference on Image Processing*, pages 3206–3210, Sep. 2013.

[264] Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[265] W. Yan, S. Plis, V. D. Calhoun, S. Liu, R. Jiang, T. Jiang, and J. Sui. Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sep. 2017.

[266] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognition*, 86:143–155, 2019.

[267] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[268] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Deep Representation Learning with Part Loss for Person Re-Identification. *IEEE Transactions on Image Processing*, PP(c):1, 2017.

[269] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[270] Qian Yu, Xiabin Ching, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching. (3).

[271] Rui Yu, Zhichao Zhou, Song Bai, and Xiang Bai. Divide and Fuse: A Re-ranking Approach for Person Re-identification. pages 1–13, 2017.

[272] Mingyong Zeng, Chang Tian, and Zemin Wu. Person Re-identification with Hierarchical Deep Learning Feature and efficient XQDA Metric. pages 1838–1846, 2018.

[273] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[274] Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 289–301. Curran Associates, Inc., 2020.

[275] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

[276] Honglun Zhang, Wenqing Chen, Hao He, and Yaohui Jin. Disentangled makeup transfer with generative adversarial network. *arXiv preprint arXiv:1907.01144*, 2019.

[277] Li Zhang, Mingliang Wang, Mingxia Liu, and Daoqiang Zhang. A survey on deep learning for neuroimaging-based brain disorder analysis. *Frontiers in neuroscience*, 14, 2020.

[278] Lining Zhang, Hubert P. H. Shum, Li Liu, Guodong Guo, and Ling Shao. Multiview discriminative marginal metric learning for makeup face verification. *Neurocomputing*, 333:339–350, 2019.

[279] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2021.

[280] Yiheng Zhang, Dong Liu, and Zheng Jun Zha. Improving triplet-wise training of convolutional neural network for vehicle re-identification. *Proceedings - IEEE International Conference on Multimedia and Expo*, (July):1386–1391, 2017.

[281] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[282] Jian Zhao, Lin Xiong, Jayashree Karlekar, Jianshu Li, Fang Zhao, Zhecan Wang, Sugiri Pranata, Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, volume 2, page 3, 2017.

[283] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.

[284] Wei-Shi Zheng, Gong Shaogang, and Xiang Tao. Person re-identification by probabilistic relative distance comparison. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 649–656, 2011.

[285] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2018.

[286] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[287] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[288] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019.

[289] Erjin Zhou, Zhimin Cao, and Jian Sun. Gridface: Face rectification via learning local homography transformations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[290] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019.

[291] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.

[292] Yi Zhou, Li Liu, and Ling Shao. Vehicle Re-Identification by Deep Hidden Multi-View Inference. 27(7):3275–3287, 2018.

[293] Yi Zhou and Ling Shao. Cross-View GAN Based Vehicle Generation for Re-identification. *BMVC*, 1:1–12, 2017.

[294] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[295] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng. Deep hybrid similarity learning for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3183–3193, Nov 2018.

[296] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[297] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.

[298] Suguo Zhu, Xiaowei Gong, Zhenzhong Kuang, and Junping Du. Partial person re-identification with two-stream network and reconstruction. *Neurocomputing*, 2019.