Data-augmentation with synthetic identities for robust facial recognition

Richard Marriott


N^o d'ordre NNT : 2020LYSEC48

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de l'Ecole Centrale de Lyon

Ecole Doctorale INFOMATHS ED N^o 512

Spécialité: Informatique

Soutenue publiquement le 14/12/2020, par:

Richard T. MARRIOTT

Data-augmentation with Synthetic Identities for Robust Facial Recognition

---

Devant le jury composé de :

M. Liming Chen	Professeur de l'Ecole Centrale de Lyon	Directeur de thèse
M. Sami Romdhani	Docteur, IDEMIA	Encadrant
M. Stéphane Gentric	Docteur, IDEMIA	Encadrant
Mme. Bernadette Dorizzi	Professeur de Télécom SudParis	Présidente du jury
M. Dimitris Samaras	Professeur du Stony Brook University	Rapporteur
M. Boulbaba Ben Amor	Professeur de l'Institut Mines-Télécom Lille Douai	Rapporteur
M. Ioannis Kakadiaris	Professeur de l'University of Houston	Examinateur
Mme. Séverine Dubuisson	Professeur de l'Aix-Marseille Université	Examinateur

Acknowledgements

I would like to briefly try to thank all of the people that have contributed to this work, whether it be directly or indirectly, in practical ways or simply by way of encouragement, motivation or welcome distraction.

First and foremost I would like to thank my supervisors, Liming, Sami and Stéphane. I count myself fortunate to have benefited from your guidance and example. Thank you to Sami for day-to-day support/discussions, novel ideas and your critical eye; thank you to Liming for regularly allowing me take a step back to see my work from a different perspective, for initially introducing me to deep learning, and for maintaining a stimulating environment at ECL; and thank you to Stéphane for sharing your biometric expertise and, in particular, for advice that allowed me to find additional value in my work and ultimately enabled me to pull my thesis together into its final form on time.

I would also like to thank each of the members of the jury, in particular Dimitris and Boulbaba for acting as rapporteurs and providing valuable feedback including encouraging remarks, constructive criticism, and interesting ideas for further analysis of the problem. Dimitris also acted as the external member of my "comité de suivi de thèse" for which I am additionally grateful. Thank you to Ioannis and Séverine for taking the time to read through the manuscript and for posing interesting and insightful questions. Similarly, thank you to Bernadette who also kindly agreed to chair the jury.

Of course, completing a PhD in informatics would not be possible without the technical savoir faire of a great many people. For this support, I thank all of my colleagues at IDEMIA, who are too many to name individually. I will thank individually, however, my immediate team-mates, both present and past - Rahat, Damien, Pierre, Baptiste and Iana, and also Safa for her direct contributions to the work of Chapter 4.

If you haven't already defended your theses, bon courage to my fellow thésards in the lab at ECL, and congratulations to the rest. Thank you to postdocs and doctorants alike for your camaraderie over the years. I apologise for my unceremonious, Covid-related departure from Lyon and for depriving you of a final pot de départ. I'm sure, however, that with all my comings and goings, you must feel like I've had enough leaving celebrations by now. Congratulations also to Robin, my "camarade de CIFRE" at IDEMIA. I wish you well in each of the three pillars of life.

To my friends, Bhaskar, Didi, Lauren and Miguel, thank you for making Paris a home from home, et à ma chère amie, Sophie Cerre, je te remercie pour plein de beaux souvenirs de Lyon.

Last but not least, thank you to my parents and relatives, without whose support my initial move to France and subsequent career in computer vision would not have been possible.

Contents

List of Tables

List of Tables

List of Figures

Abstract

In 2014, use of deep neural networks (DNNs) revolutionised facial recognition (FR). DNNs are capable of learning to extract feature-based representations from images that are discriminative and robust to extraneous detail. Arguably, one of the most important factors now limiting the performance of FR algorithms is the data used to train them. High-quality image datasets that are representative of real-world test conditions can be difficult to collect. One potential solution is to augment datasets with synthetic images. This option recently became increasingly viable following the development of generative adversarial networks (GANs) which allow generation of highly realistic, synthetic data samples. This thesis investigates the use of GANs for augmentation of FR datasets. It looks at the ability of GANs to generate new identities, and their ability to disentangle identity from other forms of variation in images. Ultimately, a GAN integrating a 3D model is proposed in order to fully disentangle pose from identity. Images synthesised using the 3D GAN are shown to improve large-pose FR and a state-of-the-art accuracy is demonstrated for the challenging Cross-Pose LFW evaluation dataset.

The final chapter of the thesis evaluates one of the more nefarious uses of synthetic images: the face-morphing attack. Such attacks exploit imprecision in FR systems by manipulating images such that they might be falsely verified as belonging to more than one person. An evaluation of GAN-based face-morphing attacks is provided. Also introduced is a novel, GAN-based morphing method that minimises the distance of the morphed image from the original identities in a biometric feature-space. A potential counter measure to such morphing attacks is to train FR networks using additional, synthetic identities. In this vein, the effect of training using synthetic, 3D GAN data on the success of simulated face-morphing attacks is evaluated.

Keywords: Facial recognition, data-augmentation, generative adversarial network, GAN, disentanglement, face-morphing attack

Résumé

En 2014, l'utilisation des réseaux neuronaux profonds (RNP) a révolutionné la reconnaissance faciale (RF). Les RNP sont capables d'apprendre à extraire des images des représentations basées sur des caractéristiques qui sont discriminantes et robustes aux détails non pertinents. On peut dire que l'un des facteurs les plus importants qui limitent aujourd'hui les performances des algorithmes de RF sont les données utilisées pour les entraîner. Les ensembles de données d'images de haute qualité qui sont représentatives des conditions de test du monde réel peuvent être difficiles à collecter. Une solution possible est d'augmenter les ensembles de données avec des images synthétiques. Cette option est récemment devenue plus viable suite au développement des "generative adversarial networks" (GAN) qui permettent de générer des échantillons de données synthétiques très réalistes. Cette thèse étudie l'utilisation des GAN pour augmenter les ensembles de données FR. Elle examine la capacité des GAN à générer de nouvelles identités, et leur capacité à démêler l'identité des autres formes de variation des images. Enfin, un GAN intégrant un modèle 3D est proposé afin de démêler complètement la pose de l'identité. Il est démontré que les images synthétisées à l'aide du GAN 3D améliorent la reconnaissance des visages aux poses larges et une précision état de l'art est démontrée pour l'ensemble de données d'évaluation "Cross-Pose LFW".

Le dernier chapitre de la thèse évalue l'une des utilisations plus néfastes des images synthétiques : l'attaque par morphing du visage. Ces attaques exploitent l'imprécision des systèmes de RF en manipulant les images de manière à ce qu'il puisse être faussement vérifié qu'elles appartiennent à plus d'une personne. Une évaluation des attaques par morphing de visage basées sur le GAN est fournie. Une nouvelle méthode de morphing basée sur le GAN est également présentée, qui minimise la distance entre l'image transformée et les identités originales dans un espace de caractéristiques biométriques. Une contre-mesure potentielle à ces

attaques par morphing consiste à entraîner les réseaux FR en utilisant des identités synthétiques supplémentaires. Dans cette veine, l'effet de l'entraînement utilisant des données synthétiques GAN 3D sur le succès des attaques simulées de morphing facial est évalué.

Mots clés: Reconnaissance faciale, enrichissement des données, generative adversarial networks, GAN, disentanglement, attaque par morphing du visage

Publications

"Taking Control of Intra-class Variation in Conditional GANs Under Weak Supervision". R. T. Marriott, S. Romdhani, and L. Chen.
In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). [Marriott *et al.* 2020a]

"An Assessment of GANs for Identity-related Applications".
R. T. Marriott, S. Madiouni, S. Romdhani, S. Gentric and L. Chen.
In 2020 IEEE International Joint Conference on Biometrics (IJCB 2020). [Marriott *et al.* 2020b]

"A 3D GAN for Improved Large-pose Facial Recognition".
R. T. Marriott, S. Romdhani, and L. Chen.
Accepted to CVPR 2021. [Marriott *et al.* 2020c]

"Robustness of Facial Recognition to GAN-based Face-morphing Attacks".
R. T. Marriott, S. Romdhani, S. Gentric and L. Chen.
In submission. [Marriott *et al.* 2020d]

Chapter 1

Introduction

The fundamental challenge in facial recognition (FR) is to be able to extract robust representations of identity from images. Ideally, these representations should be unique, easily separable in feature-space, and invariant to confounding factors such as pose, illumination, age and expression. Until *circa* 2010, the most successful representations for FR were formed by combining hand-crafted features such as Gabor [Liu & Wechsler 2002] or local binary pattern (LBP) filter activations [Ahonen *et al.* 2006]. These local appearance features demonstrated a degree of invariance to illumination and expression in face-images. However, in 2010 with the introduction of learned filters [Cao *et al.* 2010], and particularly in 2011/2012 with the popularisation of deep learning [Ciresan *et al.* 2011, Krizhevsky *et al.* 2012], hand-crafted features quickly became obsolete. In 2014, Facebook released DeepFace [Taigman *et al.* 2014] - a deep neural network with an architecture similar to that of AlexNet [Krizhevsky *et al.* 2012]. AlexNet is well-known for having significantly outperformed all other methods in the ImageNet challenge in 2012 [Russakovsky *et al.* 2015]. Inheriting from this success, DeepFace was the first FR system to achieve equivalent to human-level performance on the Labelled Faces in the Wild (LFW) benchmark. It's success derives from it's ability to learn robust, discriminative features from Facebook's dataset of 4.4 million images of four thousand different identities.

This shift from hand-crafted to learned features also shifted many researchers' focus towards the effective use of data. In order to learn robust, discriminative features, deep neural networks must be "shown" in what consists an identity; i.e. the dataset should contain many examples of different identities (in order to learn to discriminate between them), and it should contain a variety of different ex-

amples of each of those identities (in order to learn robustness to the differences within those image sets). On both of these fronts, it can be problematic to obtain such datasets in practice. Data-protection regulations inhibit the collection and use of images of most subjects with the exception of celebrities, and even for celebrities, automatic image-scraping algorithms can introduce large amounts of labelling noise [Wang *et al.* 2018]. The domain-shift present due to differences in the image-capture conditions of training and test data also poses a major problem. The very fact of being permitted to use an image for training implies some degree of cooperation with the subject. Images released publicly tend to be high-quality, frontal images of well-lit, smiling subjects. This contrasts with the non-cooperative capture conditions involved in many applications of FR. For example, images captured by CCTV cameras may be poorly lit and may contain any range of pose and expression. Bridging this gap between conditions at training and test time is an important challenge in FR.

Facial recognition is a somewhat special case of image-classification since faces display *a priori* form, with most faces being describable by points on a continuum of common features. This makes it possible to accurately model and manipulate facial appearance and provides an avenue by which one might tackle the problem of limited training data. Two possible approaches present themselves:

1. Normalisation of test data

2. Augmentation of training data

Both of these options have been evaluated by the FR community with varying degrees of success. Normalisation of test data was extensively used prior to the adoption of deep learning methods. Given frontalised face images, for example, facial appearance can be reasonably well described by simple linear combinations of features, as was done explicitly by the well-known EigenFaces method [Turk & Pentland 1991]. Even DeepFace found that it was important to perform face frontalisation using a 3D shape model prior to encoding. Despite its name, however, DeepFace was a relatively shallow network by today's standards, employing

only two convolutional layers followed by three fully connected layers. More recent networks have been able to achieve higher accuracy than DeepFace without performing frontalisation [Schroff *et al.* 2015, Hasnat *et al.* 2017, Deng *et al.* 2019a]. These deeper, more nonlinear networks are able to infer relatively stable representations of identity from images with less sensitivity to the context. Given this ability of modern networks to ignore nonlinear phenomena, we aim to improve robustness via augmentation of training data rather than normalisation of test data. This has the additional benefit of involving one less step in the recognition pipeline at test time, which is potentially important given the often time-critical applications of FR.

1.1 Data-augmentation and synthetic identities

The main sources of non-identity variation in face images are pose, illumination and expression (sometimes known as the PIE attributes). The effect on an image of manipulating these attributes is (or at least should be) directly related to the 3D shape of the face. Most attempts at face data-augmentation therefore involve the use of some 3D face model to facilitate image-manipulation. Commonly used models are 3D morphable models (3DMMs) [Blanz & Vetter 1999], which represent 3D face-shape, and sometimes texture, as linear combinations of basis vectors. These linear texture models can be useful during fitting of the 3D shape model to images. However, images rendered using the linear texture models tend to be unrealistic and smooth, lacking in high-frequency detail. For this reason, face data-augmentation methods have tended to extract textures directly from input images prior to manipulation of the PIE attributes. This means that synthesised images used in data-augmentation tend to belong to identities that are already present in the training dataset.

In 2014, [Goodfellow *et al.* 2014] introduced the Generative Adversarial Network (GAN) as a method of generating synthetic data samples. (A thorough introduction to the GAN will be given in section 2.1.) The GAN was enthusiastically adopted by the deep learning community and has since been the subject of a vast amount of development. During the time taken to compile this manuscript, the ci-

tation count for [Goodfellow *et al.* 2014] increased from 18829 to 28647 (and counting). By 2017 it was possible to generate highly realistic synthetic face images at a resolution of 1024×1024 pixels [Karras *et al.* 2018]. Based on tests of the perceptual continuity of images upon linear interpolation in the latent spaces of GANs, and also upon identification of "nearest neighbour" images in training datasets, face images randomly synthesised by GANs are believed to represent novel identities [Karras *et al.* 2018]. In Section 3, further evidence will be provided that supports this to indeed be the case. GANs therefore provide a potential avenue for augmentation of FR training datasets with realistic images of synthetic identities. This is an area that has not been well explored in the literature and is the primary subject of this thesis.

Augmentation of FR datasets with synthetic identities has several potential benefits:

1. Training of FR networks with additional identity samples is likely to lead to a more discriminative feature space.

2. Image sets generated for a synthetic identity will not suffer from labelling problems (provided the synthetic identity is well maintained).

3. It may be possible to generate useful synthetic FR training sets from face-image datasets containing little useful identity information, e.g. datasets containing only one photo per identity.

4. Image quality and identity consistency within synthetically augmented image sets may be greater since no reconstruction of existing identities need be performed.

These potential benefits are re-visited in Chapter 5 where datasets of synthetic identities, generated by our 3D GAN, are evaluated for FR.

Whilst GANs may potentially provide the aforementioned benefits to FR, the ability to generate realistic synthetic identities may be a double-edged sword: GANs also constitute a tool that could be used to attack FR systems. The following section explains why this is the case and introduces the face-morphing attack.

1.2 Face-morphing attacks

Whilst FR must be robust to non-identity variation, it must also be robust to imposters, i.e. two subjects sharing similar facial characteristics should not be confused with one-another. In 2017, [Ferrara *et al.* 2014] demonstrated how imprecision in identity classification, by both humans and FR systems, could be exploited in order for an imposter to be granted unauthorised access. An attack on a passport-controlled frontier might proceed as follows:

1. An accomplice is identified who is willing to share their passport with an imposter. (Ideally the chosen accomplice will resemble the imposter.)

2. A morphed image is produced of a synthetic identity that resembles both the accomplice and the imposter.

3. The accomplice presents the morphed image at the time of application for a new passport. (The image is found to be plausible and so is accepted.)

4. The resulting passport is then shared with and used by the imposter.

Work in the literature has shown that detection of morphed images is not sufficiently effective [Scherhag *et al.* 2017b, Makrushin & Wolf 2018]. Rather than trying to detect morphed images, Chapter 6 presents work showing that a significant reduction in the number of successful simulated morphing attacks can be achieved via improvements in fidelity of the FR system. The impact of training an FR system with additional, synthetic identities is also assessed. Results currently show only subtle improvements. However, we believe training with synthetic identities to be a promising direction for improving robustness to morphing attacks.

1.3 Problem definition

Facial recognition systems need to be robust to intra-class variation (i.e. non-identity variation), and also to imposters, whether those imposters be coincidental or deliberate face-morphing attacks. It is assumed by the FR community that the

nonlinear feature encodings of deep neural networks are capable of learning such robustness given data constituting a suitably dense and wide sampling of both identity and non-identity variation. For various reasons, such high-quality datasets are difficult to collect. The recent development of GANs, however, provides a potential solution for generation of high-quality synthetic datasets.

Four potential advantages of data-augmentation using synthetic identities (as opposed to existing ones) were identified in Section 1.1. Data-augmentation using synthetic identities is an approach that has not yet been widely evaluated in the literature. In this thesis, the feasibility and effectiveness of performing such data-augmentation is investigated. The specific research problems to be addressed are as follows:

1. Do GANs actually generate new identities? Published results suggest that this is the case. However, there have been no systematic, quantitative studies to confirm this.

2. How can identity be successfully disentangled from other attributes in images generated by GANs?

3. Can synthetic identities be used successfully to improve the accuracy of facial recognition?

4. Do images generated by GANs pose a threat to the security of FR systems?

5. Do synthetic identities help to improve robustness of FR systems to face-morphing attacks?

1.4 Contributions

The contributions made in each chapter of this thesis are listed here:

- Chapter 3

    1. Explicit demonstration that GANs do not overfit to the training dataset and do, in fact, generate new identities (thereby introducing an indirect method of assessing overfitting, and also mode-collapse, in GANs).

- Chapter 4

    1. Introduction of the intra-class variation isolation mechanism for training conditional GANs to learn disentangled, multi-variate models of variation with only the weak supervision of binary labels.
    2. Introduction of a "GAN triplet loss" for improved disentanglement of identity from other image factors in SD-GANs.
    3. Demonstration that identity-constrained, 2D GAN methods do not adequately disentangle identity from other kinds of variation in generated images.

- Chapter 5

    1. Development of the 3D GAN which integrates a 3DMM into the generator of a GAN for identity-preserving disentanglement of pose.
    2. Demonstration that synthetic, 3D GAN data can be used to improve robustness of FR algorithms to large poses giving a state-of-the-art accuracy on the challenging CPLFW dataset.

- Chapter 6

    1. An assessment of "style-based" GAN face-morphing attacks (concurrent with the evaluation in [Venkatesh *et al.* 2020]).
    2. Introduction and evaluation of the "dual biometric face-morphing method".
    3. Demonstration that improvements to the fidelity of FR systems lead to increased robustness to face-morphing attacks provided morphed images are considered when setting acceptance thresholds.
    4. An assessment of the effect of training with synthetic identities on the success of simulated face-morphing attacks.

1.5 Outline of the thesis

The thesis has been organised into the following chapters:

- Chapter 2 – "Literature review"

  A synthesis of the literature covering the fundamentals of GANs and data-augmentation for FR.

- Chapter 3 – "Do GANs actually generate new identities?"

  This chapter investigates over-fitting in GANs and provides explicit, quantitative evidence that new identities, not present in the training dataset, are generated.

- Chapter 4 – "Disentanglement of identity in GANs"

  This chapter introduces a new mechanism for disentangling various forms of labelled variation, from identity and from one another, in the latent space of a conditional GAN. Also introduced is a novel triplet loss function for training "SD-GANs" which is demonstrated to improve the disentanglement of identity. The characteristics of the resulting synthetic data are assessed and compared with those of data generated by other, similar methods. Despite the demonstrated improvements, full disentanglement of identity in typical GANs is shown to remain an unsolved problem.

- Chapter 5 – "A 3D GAN for identity-preserving disentanglement of pose"

  Motivated by the disappointing conclusions of the previous chapter, this work integrates a 3D morphable model into a GAN as a fool-proof way of modifying pose whilst preserving identity. The method inherits the stability of a 3DMM and the realism of GAN-generated images. Results show that image sets generated using this method can be successfully used to improve the accuracy of large-pose FR.

- Chapter 6 – "Robustness of facial recognition algorithms to morphing attacks"

  This chapter investigates the threat posed by style-based GAN face-morphing methods to FR systems. Also assessed is the effect on the success of simulated attacks of training using synthetic 3D GAN data.

- Chapter 7 – "Conclusions and Future Work"

  Finally we conclude and discuss potential directions for future research.

Chapter 2

Literature Review

2.1 Development of Generative Adversarial Networks

The work in this thesis builds upon the Generative Adversarial Network (GAN) [Goodfellow *et al.* 2014] and some of the recent developments thereto. This part of the literature review aims to provide a solid theoretical background of GANs, drawing from various important works in the literature and providing additional intuitive explanations where possible. It will cover GANs, Wasserstein GANs, conditional GANs, and various improvements to training techniques. We first begin, however, with a brief look at the form of generative neural networks prior to GANs.

2.1.1 From Boltzmann Machines to GANs

Most work on generative neural networks prior to GANs was based on some version of the Boltzmann Machine [Hinton & Sejnowski 1983]. The Boltzmann Machine or its deep equivalent [Salakhutdinov & Hinton 2009] (an example of which is pictured in Figure 2.1) is an undirected graph of binary nodes **h** and **v** connected by edges with weights **W**. The **h** are "hidden nodes" and the **v** are "visible nodes". The aim is to train the weights such that, by following certain node-update rules, the visible nodes activate with the same distribution from which a set of training data were sampled (for example, a set of training images that have been binarised and vectorised). Unlike typical causal networks such as multi-layer perceptrons, there is no feed-forward mechanism to generate visible outputs deterministically from a hidden state. Boltzmann machines are energy-based models and generate stochastic output where the probability of occurrence of a particular global state (hidden plus visible nodes) is inversely proportional to the global energy assigned to that state

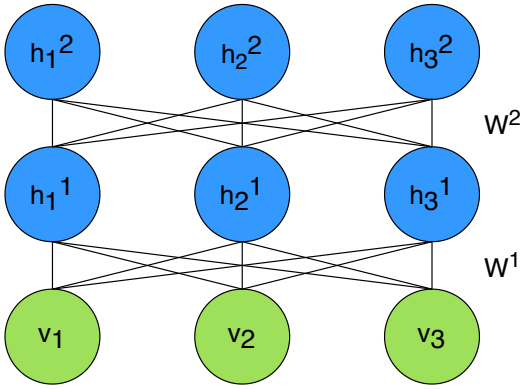


Figure 2.1: An example of the structure of a Deep Boltzmann Machine (DBM) with two hidden layers and a visible layer of nodes connected by weights $\mathbf{W}^1$ and $\mathbf{W}^2$. DBMs have no within-layer connections to enable efficient updates of layers in parallel.

by the following equation.

$$E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \mathbf{W}^1, \mathbf{W}^2) = -\mathbf{v}^T \mathbf{W}^1 \mathbf{h}^1 - \mathbf{h}^{1T} \mathbf{W}^2 \mathbf{h}^2 \tag{2.1}$$

If we assume the edge-weights have been trained, a sample may then be generated as follows. Initially, both the hidden and visible nodes are set randomly. Nodes are then selected in a random order (with hidden and visible nodes being treated equally) and are assigned a value of either 0 or 1 to locally minimise equation (2.1). Eventually, the network will reach "thermal equilibrium" at which point the visible units should represent high probability (and therefore low energy), coherent images. (The weights were trained such that this would be the case.) Note that this thermal equilibrium is not a static state. Since each node-update (or layer-update for deep Boltzmann machines) is performed in isolation, the network does not reach a global minimum. Instead, oscillations occur and the network "fizzes" with each update causing the configuration of visible nodes to gradually cycle through samples of the learned model distribution corresponding to the current energy minimum. This gradual sampling procedure is often described as "running a Markov chain" since each global configuration is dependent only on the previous state. To sample from isolated regions of the distribution, the network may need to be re-initialised to a different random state before following a different chain to thermal equilibrium.

The training procedure for the weights aims to maximise the probability that randomly occurring visible states correspond to the training data vectors when the system is at thermal equilibrium. The probability of occurrence of a particular vector, $\mathbf{v}$, is given by

$$p(\mathbf{v}; \mathbf{W}^1, \mathbf{W}^2) = \frac{\sum_{\mathbf{h}^1, \mathbf{h}^2} e^{-E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \mathbf{W}^1, \mathbf{W}^2)}}{\sum_{\mathbf{u}} \sum_{\mathbf{h}^1, \mathbf{h}^2} e^{-E(\mathbf{u}, \mathbf{h}^1, \mathbf{h}^2; \mathbf{W}^1, \mathbf{W}^2)}} \tag{2.2}$$

and the chosen training procedure minimises

$$G(\mathbf{W}^1, \mathbf{W}^2) = -\mathbb{E}_{\mathbf{v} \sim p_{data}}[log(p(\mathbf{v}; \mathbf{W}^1, \mathbf{W}^2))] \tag{2.3}$$

where p_{data} is the real data-generating distribution.

The principal issue with Boltzmann machines is the difficulty of estimating the values of the sums in equation (2.2). The denominator of equation (2.2) is a sum over all possible configurations of the Boltzmann machine, which quickly becomes intractable as the size of the network increases. Instead, an approximation is made by taking a few samples of the more important modes of the distribution by running a set of Markov chains to their corresponding thermal equilibria. However, even this sampling procedure is computationally expensive.

As a solution to some of the problems with Boltzmann Machines, [Bengio *et al.* 2014] proposed the Generative Stochastic Network (GSN). The method transforms the problem of unsupervised distribution learning into something akin to supervised function approximation, and in doing so is able to capitalise on some of the recent successes in deep learning; in particular denoising autoencoders [Vincent *et al.* 2008]. Rather than training a neural network to parametrise full, multi-modal probability distributions, GSN parametrise only the transition function of the Markov chain between global states; i.e. given a state $\tilde{\mathbf{x}}_t$ (where the tilde denotes that noise has been added to the state $\mathbf{x}_t$) the GSN learns to generate $\mathbf{x}_{t+1}$. Note that the weights of a Boltzmann machine explicitly represented correlations between elements of the state vector whereas the weights of a GSN

Figure 2.2: Synthetic samples generated by running the Markov chain of a GSN. The image from the training set closest to the final synthetic image in each row is shown in the final column. Figure taken from [Bengio *et al.* 2014].

belong to a function *applied* to a state vector. Since the weights of the network are trained to approximate the conditional distribution $p(\mathbf{x}_{t+1}|\tilde{\mathbf{x}}_t)$ rather than the entire distribution $p(\mathbf{x})$, the training need only be concerned with maximising the probability of "realistic states" $\mathbf{x}_{t+1}$ (i.e. states corresponding to the training data) occurring in the vicinity of $\tilde{\mathbf{x}}_t$. This local, conditional distribution is likely to have far fewer modes than the full probability distribution of the data and so presents an easier learning task to the network. During training of the transition function, the states $\tilde{\mathbf{x}}_t$ are created by applying a stochastic "corruption function", $\tilde{\mathbf{x}}_t = C(\mathbf{x}_t)$, to training images. The GSN is then trained via backpropagation as a denoising autoencoder to minimise the reconstruction error between $\mathbf{x}_{t+1} = GSN(\tilde{\mathbf{x}}_t)$ and $\mathbf{x}_t$, and therefore maximise the likelihood that $\mathbf{x}_{t+1}$ approximates a sample from the data distribution. Running of the Markov chain to generate samples then involves alternately injecting noise to locally perturb the model state (the image) followed by applying the GSN to walk the perturbed state back to the manifold of probable images. The effect of running a GSN Markov chain is shown in Figure 2.2 together with nearest neighbour samples from the training dataset.

Although GSNs simplify the training procedure of generative networks by facilitating the use of backpropagation, sampling of a diverse set of images can be slow due to low mixing rates, i.e. the rate at which the model distribution is explored

by the Markov chain. The generative adversarial network (GAN) overcomes this issue by avoiding the use of Markov chains altogether.

Whereas GSNs transformed the unsupervised problem of generating samples from some data distribution into the supervised problem of image-denoising, Generative Adversarial Networks (GANs) [Goodfellow *et al.* 2014] transform the problem into one of supervised binary classification. GANs learn a deterministic mapping between some known, input distribution - typically a standard, multi-variate Gaussian distribution - and some complex, multi-dimensional distribution from which data can be sampled, in our case the distribution of images of faces. As with GSNs, this mapping is trained via back-propagation. However, unlike GSNs, the training signal is not derived from a maximum likelihood loss function that aims to reconstruct training data samples. Indeed, since the form of the GAN's mapping between the input distribution and the data distribution is not known *a priori*, it is not clear to which data samples the random model output would be compared. Instead, a network known as the discriminator, is trained to judge whether images belong to the model distribution or the training distribution. The GAN's generator and discriminator networks are trained alternately in a mini-max game in which the generator's goal is to produce images that are incorrectly judged as being real by the discriminator. This is equivalent to minimising the distance between the model distribution and the training distribution. The precise distance being minimised depends on the form of the loss function used to train the discriminator. Two loss functions will be described in the following sections: the original "non-saturating" loss, and the Wasserstein loss which is used preferentially in the work of this thesis.

2.1.2 The "non-saturating" GAN

The loss function originally recommended for training GANs, proposed by [Goodfellow *et al.* 2014], is now commonly known as the "non-saturating loss" due to the modification made to generator's loss function to avoid weak gradients. The discriminator is trained using the standard cross-entropy loss for binary classifica-

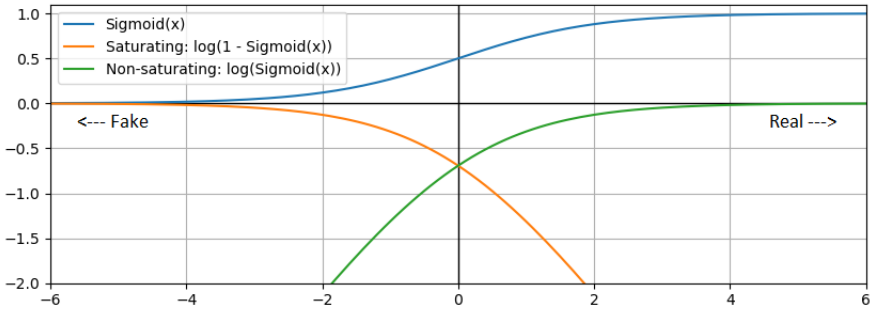


Figure 2.3: Visualisation of generator loss functions of the saturating and non-saturating GAN. The output of the GAN's discriminator is a sigmoid function trained to output 1 for real images and 0 for fake images. Where the sigmoid output is 0, however, training of the generator struggles to minimise the saturating loss function due to weak gradients. Instead, the non-saturating loss is maximised.

tion:

$$\max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}}[logD(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z}[log(1 - D(G(\mathbf{z})))] \tag{2.4}$$

where $\mathbf{z}$ is a vector of random values drawn from some known distribution p_z, $G(\mathbf{z})$ is the generator network, and $D(\mathbf{x})$ is the discriminator network that outputs a scalar value in the range $[0, 1]$, typically squashed by a sigmoid function. Where the input to the discriminator is real, the relevant term of equation (2.4) is the cross-entropy term corresponding to labels of 1, and where the input to the discriminator is generated, it is the cross-entropy term corresponding to labels of 0, i.e. the discriminator is trained to output the probability that it's input is drawn from p_{data} rather than $G(p_z)$.

The generator could then be trained to minimise the same loss function, i.e.

$$\min_G V(D,G) = \mathbb{E}_{\mathbf{z}\sim p_z}[log(1 - D(G(\mathbf{z})))] \tag{2.5}$$

However, since the output of D is limited to the interval $[0, 1]$, training a good discriminator can quickly cause output to saturate, resulting in weak gradients with respect to the parameters of the generator. This can be seen in Figure 2.3 where the gradient of the "Saturating" loss function ("log(1 - Sigmoid(x))") tends to zero as the output of the discriminator ("Sigmoid(x)") tends to zero. Instead, it

is recommended that the generator be trained to *maximise* the function

$$\max_G V(D,G) = \mathbb{E}_{\mathbf{z}\sim p_z}[logD(G(\mathbf{z}))] \tag{2.6}$$

which results in the same fixed point of the dynamics of G and D but ensures stronger gradients with respect to the raw output of the discriminator, and to the parameters of the generator via back-propagation. (See the green curve in Figure 2.3.)

Assuming the discriminator has been trained to give a good estimate of $p_{data}(\mathbf{x})$, training the generator to minimise (2.4) (or alternatively equation (2.5)) is equivalent to minimising the Jensen-Shannon divergence between the model distribution, p_g, and the training distribution, p_{data}. This can be shown by substituting $D(\mathbf{x}) = p_{data}(\mathbf{x}) = p_{data}(\mathbf{x})/(p_{data}(\mathbf{x}) + p_g(\mathbf{x}))$ (since $p_{data}(\mathbf{x}) + p_g(\mathbf{x}) = 1$) into the following definition of the Jensen-Shannon divergence for a finite set of samples:

$$JSD(p_{data}|p_g) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{data}}\left[log\left(\frac{p_{data}(\mathbf{x})}{(p_{data}(\mathbf{x}) + p_g(\mathbf{x}))/2}\right)\right] + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_g}\left[log\left(\frac{p_g(\mathbf{x})}{(p_{data}(\mathbf{x}) + p_g(\mathbf{x}))/2}\right)\right] \tag{2.7}$$

Performing the aforementioned substitution, and also that of $p_g(\mathbf{x})/(p_{data}(\mathbf{x}) + p_g(\mathbf{x})) = 1 - (p_{data}(\mathbf{x})/(p_{data}(\mathbf{x}) + p_g(\mathbf{x}))) = 1 - D(\mathbf{x})$ gives

$$JSD(p_{data}|p_g) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{data}}\left[log\left(2D(\mathbf{x})\right)\right] + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_g}\left[log\left(2(1 - D(\mathbf{x}))\right)\right] \tag{2.8}$$

$$2\,JSD(p_{data}|p_g) = \mathbb{E}_{\mathbf{x}\sim p_{data}}\left[log\left(D(\mathbf{x})\right)\right] + \mathbb{E}_{\mathbf{x}\sim p_g}\left[log\left(1 - D(\mathbf{x})\right)\right] + 2log(2) \tag{2.9}$$

Finally, substituting $G(\mathbf{z})$ where $\mathbf{z} \sim p_z$ for $\mathbf{x} \sim p_g$ and rearranging gives

$$2\,JSD(p_{data}|p_g) - log(4) = \mathbb{E}_{\mathbf{x}\sim p_{data}}\left[log\left(D(\mathbf{x})\right)\right] + \mathbb{E}_{\mathbf{z}\sim p_z}\left[log\left(1 - D(G(\mathbf{z}))\right)\right] \tag{2.10}$$

which is identical to equation (2.4) and has a minimum at the same point in the generator's parameter space as the Jensen-Shannon divergence.

As previously noted, if the discriminator becomes too effective at distinguish-

ing between generated images and training images, the loss saturates leading to weak gradients. This is a fundamental problem of the Jensen-Shannon divergence: if the supports of distributions p_{data} and p_g do not overlap, i.e. if $p_{data}(\mathbf{x}) = 0 \,\forall\, \mathbf{x}\; s.t.\; p_g(\mathbf{x}) > 0$, then the divergence assumes its maximum value which remains constant regardless of distance between the two disjoint distributions. The introduction of the "non-saturating" loss of equation (2.6) partially remedies this issue. However, whereas the Jensen-Shannon divergence has finite values everywhere, for perfect classification of disjoint distributions, the non-saturating generator loss becomes undefined; i.e. if the discriminator is able to classify all generated samples with no uncertainty, equation (2.6) collapses to $log(0) = NaN$. For this reason, training of non-saturating GANs is notoriously tricky and it is necessary to find a good balance between the generator and discriminator, either by limiting the capacity of the discriminator or by limiting the number of training iterations prior to updating the generator. The following section introduces the Wasserstein loss which greatly simplifies the training of GANs.

2.1.3 The Wasserstein GAN

The previous section discussed how the original, non-saturating GAN approximates minimisation of the difference between the generated distribution and the training distribution as measured by the Jensen-Shannon divergence. It was noted, however, that the Jensen-Shannon divergence saturates for disjoint distributions, which are in fact common when learning mappings from a low-dimensional latent space onto manifolds in a high-dimensional image space. To avoid this problem, the Wasserstein GAN [Arjovsky *et al.* 2017] aims to estimate the Wasserstein-1 distance, commonly known as the "earth mover's distance", which is defined as follows:

$$W(p_{data}, p_g) = \inf_{\gamma \in \Pi(p_{data}, p_g)} \mathbb{E}_{(x,y)\sim\gamma} \left[\, ||x - y|| \, \right] \tag{2.11}$$

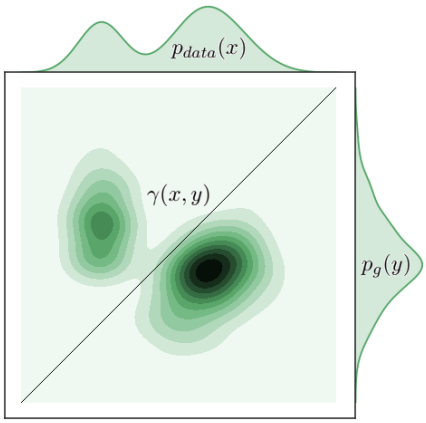


Figure 2.4: An example of a joint probability distribution, $\gamma(x, y)$, representing one possible "transport plan" between its marginal distributions, $p_{data}(x)$ and $p_g(y)$.

where $\Pi(p_{data}, p_g)$ is the set of all joint distributions, $\gamma(x, y)$, with marginals equal to p_{data} and p_g, i.e.

$$\int \gamma(x, y)\delta y = p_{data}(x), \quad \int \gamma(x, y)\delta x = p_g(y), \tag{2.12}$$

See Figure 2.4 for an example with one-dimensional p_{data} and p_g. The Wasserstein distance quantifies the solution to an optimal transport problem in which the aim is to move the "mass" from one distribution to the other whilst exerting the least amount of "work". The mass can be thought of as being moved in infinitesimal units of consistent size meaning that the work is a function only of the transportation distance, as can be seen from equation (2.11). The amount of mass being transported, let's say from $p_{data}(x)$ to $p_g(y)$ (although the distance is symmetric), depends on the frequency with which coordinates (x, y) are sampled from $\gamma(x, y)$. In essence, $\gamma(x, y)$ acts like a transportation plan dictating how often the "earth mover" should visit x to move mass to y. Optimal plans will tend to lie closer to the diagonal where $x \sim y$ and so less space need be traversed.

Calculating the Wasserstein distance, and therefore the optimal "transport plan", is intractable since the number of potential solutions, $\gamma(x, y)$, to equations (2.12) is infinite. Instead, [Arjovsky *et al.* 2017] makes use of the Kantorovich-

Rubinstein duality which states that

$$W(p_{data}, p_g) = \sup_{||f||_L \leq 1} \mathbb{E}_{x \sim p_{data}} [f(x)] - \mathbb{E}_{x \sim p_g} [f(x)] \tag{2.13}$$

i.e. the Wasserstein distance can be calculated by finding the 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ that maximises the above difference of expectations. In practice, $f(x)$ is parametrised by a neural network, $D(\mathbf{x})$, with parameters θ_d, and so equation (2.13) becomes

$$W(p_{data}, p_g, \theta_d) = \max_{\theta_d} \mathbb{E}_{x \sim p_{data}} [D(x; \theta_d)] - \mathbb{E}_{x \sim p_g} [D(x; \theta_d)] \tag{2.14}$$

where the network plays a similar role to the discriminator in equation (2.4). Since the output of $D(\mathbf{x}; \theta_d)$ belongs to $\mathbb{R}$ rather than being a probability, [Arjovsky *et al.* 2017] refers to the network as the "critic". Here, however, we will continue to refer to it as the discriminator and so keep the notation $D(\mathbf{x})$. Finally, we introduce the generator network that provides the mapping between the known distribution, p_z, and the distribution of generated images, p_g. The goal is to train the parameters of the generator, θ_g, to minimise the estimated Wasserstein distance, and so we arrive at our final mini-max objective function.

$$W(p_{data}, p_z, \theta_d, \theta_g) = \min_{\theta_g} \max_{\theta_d} \mathbb{E}_{x \sim p_{data}} [D(x; \theta_d)] - \mathbb{E}_{z \sim p_z} [D(G(\mathbf{z}; \theta_g); \theta_d)] \tag{2.15}$$

Figure 2.5 is taken from [Arjovsky *et al.* 2017] and shows the response of a "WGAN Critic" trained to distinguish between two widely separated, Gaussian distributions. The Lipschitz constraint on the discriminator - implemented in this example by clipping all elements of θ_d to the interval $[-0.01, 0.01]$ - limits the growth of the function to be at most linear with clean gradients at all points. This fact means that Wasserstein GANs train quickly and stably.

2.1.4 Regularising GANs

Enforcing the Lipschitz constraint of the Wasserstein GAN by clipping weights is problematic if the width of the clipping interval is not well tuned. An interval that

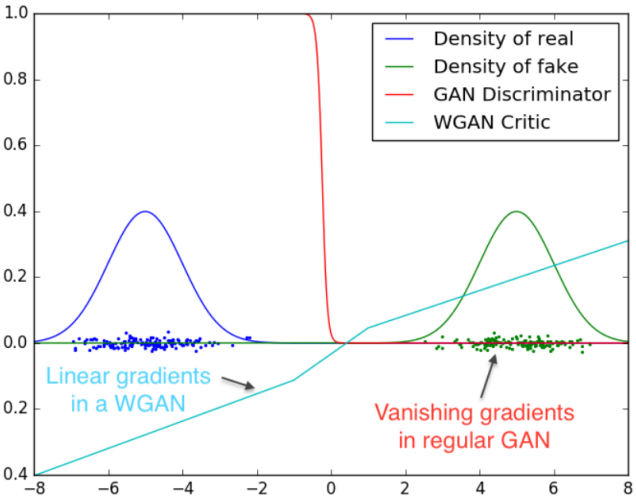


Figure 2.5: Diagram depicting the form of the output of a weight-clipped critic trained using the Wasserstein loss in comparison to a discriminator with sigmoid output. Gradients for the critic are linear despite the real and fake distributions being disjoint. Figure take from [Arjovsky *et al.* 2017].

is too wide or too narrow can lead to exploding or vanishing gradients respectively. Even if the clipping interval is well tuned, it may still be overly restrictive since interventions are performed everywhere that weights grow beyond a certain size irrespective of whether or not Lipschitz continuity is obeyed by the learned function as a whole. It was found by [Gulrajani *et al.* 2017] that weight clipping leads to simplistic functions being learned by the discriminator. Instead, they propose enforcing Lipschitz continuity via penalisation of the norm of gradients.

A function is K-Lipschitz continuous if there exists a $K \geq 0$ such that

$$||f(\mathbf{x}_1) - f(\mathbf{x}_2)|| \leq K||\mathbf{x}_1 - \mathbf{x}_2|| \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbf{X} \tag{2.16}$$

K is the Lipschitz constant that we wish to be equal to 1. If the function is differentiable everywhere then this is equivalent to saying

$$||\nabla_{\mathbf{x}} f(\mathbf{x})|| \leq K \quad \forall \mathbf{x} \in \mathbf{X} \tag{2.17}$$

The regularisation proposed by [Gulrajani *et al.* 2017] is then

$$\mathcal{L}_d = \mathop{\mathbb{E}}_{\mathbf{z}\sim p_z}[D(G(\mathbf{z}))] - \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{data}}[D(\mathbf{x})] + \lambda \mathop{\mathbb{E}}_{\hat{\mathbf{x}}\sim p_{\hat{\mathbf{x}}}}\left[(||\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})||_2 - 1)^2\right] \tag{2.18}$$

where λ is some positive constant and $\hat{\mathbf{x}} = \alpha G(\mathbf{x}) + (1-\alpha)\mathbf{x}$ with α sampled from the uniform distribution in the range $[0, 1]$. Lipschitz continuity should be enforced everywhere in the space. However, for computational efficiency gradients are evaluated only at a sample of random points intermediate to the real and generated samples. The motivation for evaluating at these points is to have well-behaved gradients as the generated distribution approaches the real distribution. Note that, to be consistent with [Gulrajani *et al.* 2017], we have swapped the order of the generated and real terms of the discriminator loss. The order is irrelevant, however, since the Wasserstein distance is symmetric. The loss in equation (2.18) is to be minimised meaning the negation of the Wasserstein distance is estimated. In doing so, the difference of the L2 norm of gradients from a value of 1 is also minimised. The method is found to perform well in practice resulting in stable training and the learning of more appropriate functions than those learned using a weight-clipped network.

It was identified in [Mescheder *et al.* 2018] that, although training using WGAN-GP is stable, it does not converge. This is true even locally for simple, toy data distributions. Upon convergence of GAN training, one would expect the generator weights to reach a stable point and for the discriminator to cease producing corrective gradient directions, i.e. the Wasserstein distance would evaluate to zero. The gradient penalty term of [Gulrajani *et al.* 2017], however, is quadratic about the point at which gradients have a norm on 1, i.e. the regularisation does not just dampen gradients but also encourages them to maintain non-zero values. This leads the GAN training procedure to continuously update the generator such that it never achieves convergence.

In parallel to the development of Wasserstein GANs, [Roth *et al.* 2017] aimed to solve the previously identified problem of non-overlapping supports in standard and non-saturating GANs by use of regularisation similar to that of the Wasserstein

GAN's gradient penalty. The problem had previously been avoided, for example in [Li *et al.* 2017a], by adding instance noise to images. [Roth *et al.* 2017] showed that adding a zero-centred gradient penalty to the discriminator loss is locally equivalent to adding instance noise. [Mescheder *et al.* 2018] propose the following simplified zero-centred gradient penalty:

$$R_1(\theta_d) = \frac{\gamma}{2}\mathbb{E}_{\mathbf{x}\sim p_{data}}\left[\,||\nabla_{\mathbf{x}} D(\mathbf{x};\theta_d)||^2\,\right] \tag{2.19}$$

As well as bridging gaps between disjoint supports, this R_1 regularisation acts to penalise the discriminator for deviating from the Nash-equilibrium (at which point the generated distribution should coincide with the real data distribution). [Mescheder *et al.* 2018] showed training of non-saturating GANs with R_1 regularisation to be locally convergent. A second form of regularisation ("R_2") with gradients evaluated at points in the space corresponding to generated data, was also proposed and found to perform similarly.

In the work of this thesis, we preferentially use the Wasserstein loss due to its stability and ability to approach realistic distributions more quickly than the non-saturating GAN. Despite the gradient penalty of [Gulrajani *et al.* 2017] being known to impede convergence, we do not switch to a zero-centred penalty. Such strong regularisation is not required by Wasserstein GANs and we instead opt for the one-sided gradient penalty proposed in [Chen *et al.* 2018].

$$R_{max}(\theta_d) = \mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{x}}}\left[max(0, ||\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}};\theta_d)||_2 - 1)\right] \tag{2.20}$$

The effect of this form of regularisation is to dampen gradients breaking the 1-Lipschitz condition but to switch off regularisation where gradients are within the allowed limit. We have not seen work, however, that explicitly demonstrates convergence of Wasserstein GANs under such regularisation.

2.1.5 Conditional GANs

A virtue of the fact that GANs learn a deterministic mapping between the latent distribution and the data distribution is that generated images are directly controllable via manipulation of the latent vector. Control over specific semantics can be given to elements of the input distribution via conditional training using labelled training images. Rather than learning to judge membership of the training image distribution alone, the discriminator of a conditional GAN (cGAN) [Mirza & Osindero 2014] learns to judge membership of the joint distribution of images and associated labels, i.e. if a generated image is to be judged as real, its characteristics must be in agreement with the accompanying label. The following equation gives the conditional version of the discriminator loss of the Wasserstein GAN in equation (2.14):

$$\mathcal{L}_{\theta_D} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{data}}[D(\mathbf{x},\mathbf{y};\theta_D)] - \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\rho}\sim p_\rho}[D(G(\mathbf{z},\boldsymbol{\rho};\theta_G),\boldsymbol{\rho};\theta_D)] \tag{2.21}$$

where $\mathbf{x}$ is an image and $\mathbf{y} \in \mathbb{R}^n$ the associated vector of labels drawn from the distribution of real data; $\boldsymbol{\rho} \in \mathbb{R}^n$ is a vector of conditioning parameters of the same form as the label vector, $\mathbf{y}$, selected from the distribution p_ρ. Typically p_ρ will be the distribution of labels of the real data. Notice that $\mathbf{y}$ and $\boldsymbol{\rho}$ share the same pathway into the discriminator. $G(\mathbf{z},\boldsymbol{\rho};\theta_G)$ must therefore learn to generate images with the same relationship to $\boldsymbol{\rho}$ as $\mathbf{x}$ has to $\mathbf{y}$. Once the cGAN's discriminator is trained, the following loss is minimised with respect to the parameters of the generator:

$$\mathcal{L}_{\theta_G} = \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\rho}\sim p_\rho}[D(G(\mathbf{z},\boldsymbol{\rho};\theta_G),\boldsymbol{\rho};\theta_D)] \tag{2.22}$$

Conditional GANs are used extensively in the literature to control both categorical and continuous characteristics of images, for example, pose and expression. This thesis primarily investigates the use of conditional GANs for the purpose of data-augmentation of facial recognition datasets. The following section therefore gives an overview of the recent work on data-augmentation in the facial recognition literature.

2.2 Data-augmentation for Facial Recognition

The primary goal of this thesis is to improve facial recognition (FR) via augmentation of image datasets. We aim to do so by exploiting GANs and their ability to generate new identities. Alternative approaches might be to normalise or augment images of existing identities, or to use non-GAN-based methods such as auto-encoders or standard 3D-based methods. There are many works in the literature demonstrating image-manipulation techniques. Here, we focus on those that were evaluated on FR tasks. Classical, non-face-specific data-augmentation methods such as image-jitter and photometric modifications will not be covered.

Table 2.1 gives a reasonably comprehensive overview of recent augmentation and normalisation methods in the literature that evaluate on FR tasks. The methods are split roughly 55% / 45% between training-data-augmentation and test-data-normalisation respectively. (See the "Augmentation" and "Norm" columns.) This proportion may be biased given the topic of this thesis. The majority of augmentation methods manipulate existing identities. Only four of the methods choose to generate synthetic identities despite the various potential advantages of doing so. We have also indicated whether the work proposes a 2D generative method (column "2D Gen"). By this we mean, for example, a method that uses a CNN to generate face-images directly, often relying on a biometric identity constraint in order to generate the desired identity. We believe that this is an important distinction and have doubts as to the usefulness of such methods for improving FR. This will be discussed in more detail below. Other methods either make use of 3D models or manipulate input images directly without risk of altering the identity.

It can be seen that the variety of attributes manipulated by the augmentation methods is wider than that of the normalisation methods. This is the case for the simple reason that it is easier to add plausible nuisance factors to images than it is to correctly model and remove existing ones. For example, a model of a pair of glasses can easily be superimposed on an image, whereas removal of glasses would require regeneration of occluded facial detail. Table 2.1 gives examples of both a 2D and a 3D method of normalising expression. However, it is typically

Method	2D Gen	Augmentation	Norm	Pose	Illum.	Exp.	Glasses	Hair	Makeup
[Masi *et al.* 2016]$^{+}$	-	Existing IDs	-	Yes	-	Yes	-	-	-
[Crispell *et al.* 2017]	-	Existing IDs	-	Yes	Yes	-	-	-	-
[Lv *et al.* 2017]	-	Existing IDs	-	Yes	Yes	-	Yes	Yes	-
[Peng *et al.* 2017]	-	Existing IDs	-	Yes	-	-	-	-	-
[Guo *et al.* 2018]	-	Existing IDs	-	-	-	-	Yes	-	-
[Kortylewski *et al.* 2018]	-	Synth IDs	-	Yes	Yes	Yes	-	-	-
[Gecer *et al.* 2020]	-	Synth IDs	-	Yes	Yes	Yes	-	-	-
[Gecer *et al.* 2018]	Part	Synth IDs	-	Yes	Yes	Yes	-	-	-
[Sáez Trigueros *et al.* 2021]*	Yes	Both	-	Yes	Yes	Yes	Yes	Yes	Yes
[Zhao *et al.* 2017]	Part	Existing IDs	-	Yes	-	-	-	-	-
[Sajid *et al.* 2018]	?	Existing IDs	-	-	-	-	-	-	Yes
[Deng *et al.* 2018]	Part	Existing IDs	Yes	Yes	-	-	-	-	-
[Shen *et al.* 2018]	Part	-	Yes	Yes	-	-	-	-	-
[Yin *et al.* 2017]	Part	-	Yes	Yes	-	-	-	-	-
[Cao *et al.* 2018a]	Yes	-	Yes	Yes	-	-	-	-	-
[Hu *et al.* 2018]	Yes	-	Yes	Yes	-	-	-	-	-
[Huang *et al.* 2017]	Yes	-	Yes	Yes	-	-	-	-	-
[Song *et al.* 2018]	Yes	-	Yes	-	-	Yes	-	-	-
[Hassner *et al.* 2015]	-	-	Yes	Yes	-	-	-	-	-
[Zhu *et al.* 2015]	-	-	Yes	Yes	-	Yes	-	-	-
[Tran *et al.* 2019]$^{\circ}$	Yes	-	-	Yes	-	-	-	-	-

Table 2.1: An overview of recent data-augmentation and data-normalisation methods in the literature that evaluated on FR. The "2D Gen" column indicates methods that use CNNs to directly generate images. "Part" indicates that the method is only partially 2D and that some 3D information has been used in the generation process. The method of [Sajid *et al.* 2018] was not exposed in the paper.
$^{+}$ Face-shape is also augmented.
* This method modifies all image properties, limited not only to the categories of the right-hand part of the table.
$^{\circ}$ [Tran *et al.* 2019] performs disentanglement via generation and not augmentation or normalisation.

only normalisation of pose that is tackled, which is the primary source of error in FR. Notice that the methods performing augmentation with synthetic identities are each able to manipulate the three most important nuisance factors of pose, illumination and expression. Being able to do so is a natural consequence of having generated a full model of the synthetic identity. Methods that augment or normalise existing identities tend to limit themselves to the manipulation of certain forms of variation, probably due to the difficulty of reconstructing existing conditions. The method ultimately proposed by this thesis avoids the difficulty of reconstruction by augmenting FR datasets using 3D-modelled, *synthetic* identities and varies pose, illumination and expression. In the rest of this section, each of the works in Table 2.1 is described in more detail.

2.2.1 Classical 3D methods

We begin by discussing some of the purely 3D normalisation and augmentation methods that were popular before GANs took the limelight. In the category of normalisation, these methods are [Hassner *et al.* 2015] and [Zhu *et al.* 2015], and in the category of augmentation, [Masi *et al.* 2016], [Crispell *et al.* 2017], [Lv *et al.* 2017] and [Peng *et al.* 2017]. The basic method of each begins by fitting a 3D model to an input image, usually by ensuring that the projection of a set of fiducial points corresponding to particular vertices of the 3D model's mesh, align with those detected in some input image. Texture is then extracted from the image onto the model's surface before being rotated, either to reduce or augment the pose variation, and projected back into the image.

The work of [Hassner *et al.* 2015] observes that estimation of 3D shape from a 2D image is not always robust when dealing with images in the wild. Due to errors in detected fiducial points and approximations in the 3D model representation, useful identifying features of the shape are not always well captured. They argue that use of a prototype head-shape is an adequate approximation and that adjusting only pose and projection parameters is a more robust and efficient solution. [Masi *et al.* 2016] extends this idea to fitting a selection of prototype head-shapes to each input image before augmenting pose. (Expression is also augmented via

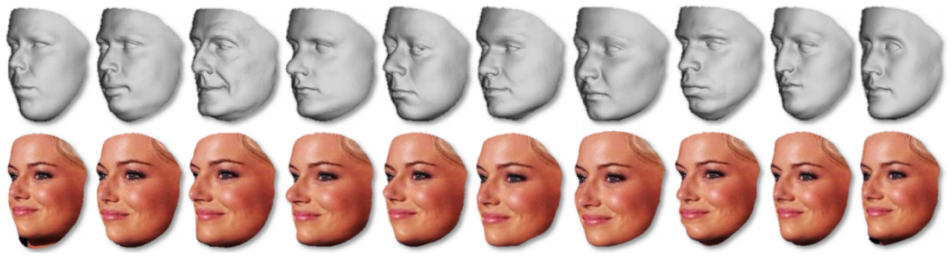


Figure 2.6: An example of the effect of augmenting face-shape. Figure taken from [Masi *et al.* 2016].

the adjustment of expression blendshapes during fitting and the addition of neutralised expressions to the training dataset.) Note that [Masi *et al.* 2016] argues use of multiple face-shapes for each image to be a useful augmentation and claims that identity is not affected. Figure 2.6 shows the effect of fitting the ten prototype heads to an image. Indeed, the textured faces in Figure 2.6 appear similar at first glance. However, use of such shape-augmentation probably prohibits use of a 3D lighting model for augmentation since the underlying shapes of the models would be more clearly revealed. The fact that the relatively crude methods of [Hassner *et al.* 2015] and [Masi *et al.* 2016] are able to improve FR corroborates the work of [Geirhos *et al.* 2019] which found that CNNs largely depend on texture rather than shape for classification. Although able to exploit this weakness of current CNNs, we have doubts as to the usefulness of such augmentation of shape in the long-term.

The methods of [Zhu *et al.* 2015], [Peng *et al.* 2017], [Crispell *et al.* 2017] and [Lv *et al.* 2017] each adjust both the pose and the shape of 3D models to fit fiducial points detected in input images before extracting the texture. The problem of estimating 3D shape from a 2D image is ill-posed and adjusting what is typically thousands of mesh vertices to capture the face shape is not feasible without prior information, i.e. without having a prior, statistical model of human head-shapes. Each of these methods simplifies the problem through use of a linear 3D morphable model (3DMM) originally proposed by [Blanz & Vetter 1999]. [Zhu *et al.* 2015] and [Peng *et al.* 2017] each use the Basel 3DMM of [Blanz & Vetter 1999] (augmented with Facewarehouse expression blendshapes [Cao *et al.* 2014] in the case of

[Zhu *et al.* 2015]), whereas [Lv *et al.* 2017] builds a model from the USF Human ID 3-D database of 100 laser-scanned heads, and [Crispell *et al.* 2017] builds one from samples from [Inc. 2016]. Each of these models reduces the problem of estimating thousands of vertex locations to one of estimating relatively few shape-basis coefficients, $\beta = [b_1, ..., b_{N_s}]$, and optionally a set of expression-basis coefficients, $\psi = [c_1, ..., c_{N_e}]$. The 3D shape, $\mathbf{S} \in \mathbb{R}^{N_v \times 3}$, described by these coefficients (where N_v is the number of vertices) is given by the following equation

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{n=1}^{N_s} b_n \mathbf{s}_n + \sum_{n=1}^{N_e} c_n \mathbf{e}_n \tag{2.23}$$

where $\bar{\mathbf{S}}$ is the mean model shape, $\mathcal{S} = [\mathbf{s}_1, ..., \mathbf{s}_{N_s}]$ are the principal components of non-expression shape variation, and $\varepsilon = [\mathbf{e}_1, ..., \mathbf{e}_{N_e}]$ are the principal components of expression variation. It was shown in [Booth *et al.* 2016] that as few as 40 principal component vectors are enough to recover 90% of the variance in the training dataset of 3D face scans, with 98% being recovered by 80. Even though 3D shape can be captured reasonably well by using 3DMMs, 3D methods that extract texture from an image each suffer from the problem of self-occlusion. Methods for tackling this problem are discussed in the following sections.

2.2.1.1 Symmetric in-filling

Despite it being possible to plausibly rotate textures extracted from images in three dimensions using 3D models, cases where there is self-occlusion in the input image remain problematic. For example, when the subject of the image is at pose, regions occluded by the nose and by the head itself leave gaps in the texture when projected to different angles. Another, more subtle effect is that surfaces at large angles to the camera are represented by fewer pixels in the original image. 3D rotation can then cause striping effects as these few pixels are projected to a larger area in the modified image. (See Figure 2.9.) Various in-filling and blending techniques that exploit facial symmetry are employed to combat these issues. The method of [Crispell *et al.* 2017], which performs augmentation of pose (and also of illumination), fills gaps caused by occlusion by sampling of pixels from symmetric regions

Figure 2.7: An example taken from [Hassner *et al.* 2015] of application of their symmetric in-filling technique. a) Original image; b) Blending weights calculated as a function of sampling frequency of pixels in the original image, super-imposed on the incomplete, frontalised texture; c) The result of symmetric in-filling and blending.

of the face. To avoid obvious discontinuities in the reconstructed texture, and also to avoid striping effects, symmetrically sampled pixels are blended with the incomplete texture using weights based on the angle of the estimated 3D surface to the viewing direction. In [Hassner *et al.* 2015], a similar method is used. However, the map of blending weights is formed by observing the number of times pixels in the original image are sampled in order to form the manipulated image. As previously described, surfaces at a large angle to the camera direction are represented by fewer pixels in the original image and so those pixels are likely to be frequently sampled to construct the manipulated image. This high sampling rate is used to indicate that regions constructed using these pixels may benefit from symmetric blending. An example of a frontalised face taken from [Hassner *et al.* 2015] is shown in Figure 2.7.

The method of [Zhu *et al.* 2015] uses a more sophisticated method of in-filling. It was recognised that uneven lighting effects break the assumption of symmetry made in the previously discussed in-filling methods. Reflecting lighting conditions into previously occluded regions where they may be out of context can look unnatural, particularly at occlusion boundaries. To help avoid this problem, [Zhu *et al.* 2015] applies the mean model-texture of the Basel 3DMM to their estimated shape and estimates the parameters of the spherical harmonic lighting model of [Ramamoorthi & Hanrahan 2001]. A complete image of "facial detail" is then constructed by taking the difference of the incomplete, frontalised image from the projection of the illuminated model and reflecting detail into occluded regions. Pois-

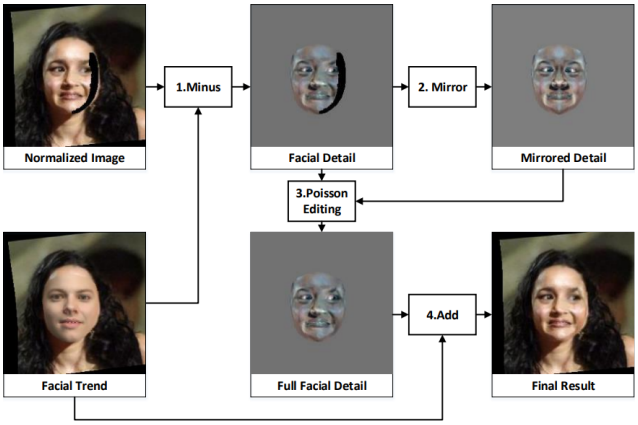


Figure 2.8: Diagram taken from [Zhu *et al.* 2015] showing the process of 3DMM + illumination model fitting followed by symmetric in-filling via reconstruction of facial detail. The approximate lighting conditions are removed from the detail map meaning the assumption of symmetry is more valid.

son editing is performed to ensure that there are no discontinuities at the occlusion boundaries in the detail image. The full, reconstructed detail is then added to the illuminated model to form the final image. This process is shown in Figure 2.8.

Symmetric in-filling is far from an ideal solution for completion of occluded textures. It can be seen in Figures 2.7 and 2.8 that the resulting images contain problematic artefacts that are likely to contribute to a domain gap between real and synthetic data. For this reason, the methods of [Lv *et al.* 2017] and [Peng *et al.* 2017] choose to avoid in-filling by only augmenting the pose of *frontal* images. This means that the rear of the head is still occluded but that the more important, central facial regions are intact. The problem of striping on the sides of the head remains, however, as can be seen in Figure 2.9. The "classical" (by which we mean non-GAN) methods discussed so far are able to manipulate pose whilst seemingly having little detrimental effect on the identity. In the next subsection we describe works that combine 3D methods with adversarial losses in attempts to preserve identity whilst avoiding artefacts.

2.2.2 Adversarial refinement methods

The works of [Zhao *et al.* 2017], [Gecer *et al.* 2018] and [Deng *et al.* 2018] each make use of 3D models for pose manipulation and are able to generate realistic

Figure 2.9: Artificial, "at-pose" images created by manipulating a frontal image in 3D. Notice the striping effects on the sides of the head in the top-left and bottom-right images due to low resolution of these regions in the original image. Figure taken from [Lv *et al.* 2017].

textures by applying adversarial losses. The methods used are significantly different from one another and are worth comparing.

Similar to the classic 3D methods described so far, [Zhao *et al.* 2017] fits a 3DMM to fiducial points detected in the image, extracts texture from the image, and re-projects at augmented angles. Occluded regions are then filled, and striped regions refined, by applying a fully convolutional, image-translation network trained using an adversarial loss. Since augmented, large-pose images do not belong to the training distribution, the adversarial loss is likely to encourage the image-refinement network to warp images back into something resembling the frontally biased images of the training dataset. To avoid this problem, the network is trained with additional reconstruction and identity-preserving losses: the reconstruction loss minimises the L1 distance of the translated image from the original, artefact-containing, pose-augmented image - a formulation that is obviously not ideal; identity is preserved by applying a cross-entropy loss to intermediate features of the discriminator. The use of identity preserving losses when generating images for augmentation of FR datasets is a questionable practice that is best avoided. As will be seen below, however, it is a fairly common practice that tarnishes many 2D normalisation and augmentation methods.

The work of [Gecer *et al.* 2018] aims to generate balanced distributions of pose, expression and lighting conditions, not by augmenting existing images, but by sampling synthetic identities from the space of the Basel 3DMM. Balanced sets of

images are generated for each synthetic identity before being passed through an adversarially trained, "synthetic-to-real" translation network to improve realism and reduce the domain gap between real and synthetic data. The method suffers from the same issues as [Zhao *et al.* 2017] and, again, requires reconstruction and identity-preserving losses. Rather than application of a direct reconstruction loss between the 3DMM image and the translated version, a cycle-consistency loss is used [Zhu *et al.* 2017]. This cycle-consistent formulation is less problematic than the direct reconstruction used in [Zhao *et al.* 2017]. However, the incestuous use of an identity-preserving loss for data-augmentation for FR remains.

The method of [Deng *et al.* 2018] benefits from high quality training data and represents the state-of-the-art in pose normalisation. Following 3DMM fitting and extraction of an incomplete texture, the method applies a convolutional U-Net which completes the texture in texture space, rather than first projecting and performing texture-completion in image space. This is made possible by a high-quality set of ground-truth textures captured by a 3D scanner and also built from multi-view datasets (Multi-PIE and the UMDFaces video dataset [Bansal *et al.* 2017]) using Poisson editing techniques. The U-Net is trained as an auto-encoder to re-generate the ground-truth dataset of textures from incomplete textures with simulated occlusions. Quality of the re-generated texture is ensured by application of an adversarial loss. An identity-preserving loss is also applied. However, this seems only to be applied for good measure and does not represent an integral part of the method. Identity-consistent texture-completion is already ensured by reconstruction of the complete ground-truth texture. The work of [Deng *et al.* 2018] is particularly interesting since their model was evaluated for both pose-normalisation and augmentation of poses for existing identities. A selection of their results are shown in Table 2.2. Evaluation on the CFP dataset [Sengupta *et al.* 2016] shows that normalisation of pose to an intermediate angle (15°), between the large-pose probe image and the frontal reference image, gave the best mean accuracies. Augmentation of the training dataset (CASIA) with additional poses gave a similar but slightly lower mean within the same margin of error as the normalisation results. N.B. since Multi-PIE and UMDFaces were used for training of the texture-

Test	CFP Frontal-Profile verification accuracy (%)
Baseline	87.74 ± 1.07
Augmentation	93.09 ± 1.72
Normalisation to Frontal	93.55 ± 1.67
Normalisation to Profile	93.72 ± 1.59
Normalisation to 15°	94.05 ± 1.73

Table 2.2: Verification accuracy (%) comparison on the CFP dataset. Results taken from [Deng *et al.* 2018].

completion model, a strict evaluation should include these datasets when training the baseline FR network. Attributing improvements of 6% to these methods would therefore be an exaggeration given the additional training data used.

Rather than using adversarial losses to refine incomplete textures, a whole host of methods aim to directly generate images of subjects at new poses. These methods are discussed in the following section.

2.2.3 Direct generative (2D) methods

Direct generative methods could be divided into two categories: auto-encoder-type methods that require paired ground-truth images (e.g. a frontal image and a corresponding image taken from a different angle), and disentangled GAN-type methods that are trained without paired images, but which require attribute labels and a biometric loss to maintain the identity. Works that fall into the auto-encoder-type category are [Hu *et al.* 2018], [Huang *et al.* 2017] and [Yin *et al.* 2017], which are each trained to translate at-pose images to resemble their frontal counterparts, and also [Cao *et al.* 2018a], which operates in two stages - first to translate to frontal, and then to translate from frontal to another pose (although the method is evaluated for FR by frontalising images). Works that fall into the disentangled GAN category are [Tran *et al.* 2019] which generates pose-conditioned images, [Song *et al.* 2018] which generates images conditioned on feature-point configurations to give varying expressions (but was evaluated for FR by normalising expressions), and also [Sáez Trigueros *et al.* 2021] which generates images containing arbitrary conditions for specified identities (although those conditions are limited to the conditions of

Method	Type	Base	Architecture	15°	30°	45°	60°	75°	90°
[Tran *et al.* 2019]	GAN	Multi-PIE	DR-GAN	95.0	91.3	88.0	85.8	-	-
Baseline	-	Multi-PIE	CASIA-Net	95.0	92.6	89.8	84.3	75.9	58.2
[Yin *et al.* 2017]	Auto-enc	Multi-PIE	CASIA-Net	94.6	92.5	89.7	85.2	77.2	61.2
Baseline	-	CASIA + MS-Celeb-1M	Light CNN	99.78	99.80	97.45	73.30	32.35	9.00
[Huang *et al.* 2017]	Auto-enc	+ Multi-PIE	Light CNN	99.78	99.85	98.58	92.93	84.10	64.03
[Hu *et al.* 2018]	Auto-enc	+ Multi-PIE	Light CNN	99.95	99.37	98.28	93.74	87.40	77.10

Table 2.3: Rank-1 identification rate (%) across poses for the Multi-PIE dataset under setting 1.

Method	Type	Base	Architecture	15°	30°	45°	60°	75°	90°
Baseline	-	CASIA + MS-Celeb-1M	Light CNN	98.59	97.38	92.13	62.09	24.18	5.51
[Huang *et al.* 2017]	Auto-enc	+ Multi-PIE	Light CNN	98.68	98.06	95.38	87.72	77.43	64.64
[Cao *et al.* 2018a]	Auto-enc	+ Multi-PIE	Light CNN	99.1	98.9	96.7	91.0	80.3	65.4
[Hu *et al.* 2018]	Auto-enc	+ Multi-PIE	Light CNN	99.82	99.56	97.33	90.63	83.05	66.05

Table 2.4: Rank-1 identification rate (%) across poses for the Multi-PIE dataset under setting 2.

the training set).

Multi-PIE is the most common dataset used for evaluation of the above works. Where possible, we have provided published results in Tables 2.3 and 2.4. (Note that this does not include evaluation of [Song *et al.* 2018] and [Sáez Trigueros *et al.* 2021].) The majority of these methods are auto-encoder-type methods that are necessarily trained on controlled data-sets such Multi-PIE that have corresponding frontal-pose pairs. This evaluation on Multi-PIE therefore masks any problems of generalisation to other datasets that are inevitable for this type of method. [Huang *et al.* 2017], [Cao *et al.* 2018a] and [Hu *et al.* 2018] each frontalise the poses in probe images before comparing the feature encodings of Light CNN [Wu *et al.* 2018] with those for frontal reference images. Light CNN is used as the baseline for these studies. For poses of 45° and above, frontalisation using a network trained on Multi-PIE is shown to consistently improve the rank-1 identification rate, by a significant amount at larger poses. Note, however, that Light CNN has only been trained using the CASIA-Webface and MS-Celeb-1M datasets. Since the training partition of Multi-PIE is available, strictly, it should be used to train the Light CNN baseline. The only work that provides an appropriate baseline with which to compare is [Yin *et al.* 2017]. As part of the frontalisation method of [Yin *et al.* 2017], a biometric network with the CASIA-Net architecture

Augmentation type	Num synth images per ID / Num synth IDs				
	0	250	500	1000	1500
Existing IDs	67.58%	66.65%	69.02%	67.74%	-
Synth IDs (500 imgs each)	67.58%	-	65.76%	66.32%	68.77%

Table 2.5: TAR@FAR=0.01 evaluated on IJB-A for a non-specified ResNet architecture trained on VGGFace augmented with either various numbers of synthetic images for existing identities or various numbers of synthetic identities. Results taken from [Sáez Trigueros *et al.* 2021].

is trained as a constraint on the generated identity. This network is then used to provide feature encodings for the identification task. The properly trained CASIA-Net recovers most of the improvement due to frontalisation when other methods are compared with the Light CNN baseline. The improvements of [Yin *et al.* 2017] over its own baseline are far smaller than those assumed for the other methods. Given that frontalisation requires additional resources and time at test time, a fairer comparison would perhaps be to augment the Light CNN and CASIA-Net baselines to have the same number of parameters as the GAN and biometric networks combined.

DR-GAN [Tran *et al.* 2019] is compared only against other works in the literature, and so it is difficult to judge whether the method provides any benefit. This is especially so as DR-GAN creates templates by combining the feature-encodings of six images of each subject, a practice not used by other methods.

The method of [Sáez Trigueros *et al.* 2021] disentangles identity from other image properties in the latent space of a GAN using an auxiliary classifier of the identity. The generator is not conditioned directly on identity labels. Instead, an encoder is trained to transform identity labels into latent vectors obeying a Gaussian distribution (according to a second discriminator). This makes it possible to straightforwardly sample new identities from the ID latent space. Table 2.5 shows values of TAR@FAR=0.01 evaluated on IJB-A for a non-specified ResNet architecture trained on VGGFace and augmented with their synthetic data (also generated from VGGFace). Augmentation of existing identities and also with new identities was performed. The results are somewhat noisy with the synthetic data causing performance to drop in half of the cases. Their best result was achieved by aug-

menting *existing* identities with 500 synthetic images each. Similar to many of the previously described methods, the generation of useful training data by this method depends on the performance of the identity constraint. If the identity constraint is more powerful than the biometric network to be augmented, e.g. if the network implementing the GAN's ID-constraint uses a larger CNN, then the synthetic data may indeed be useful. The entire GAN + biometric network system may benefit from a form of semi-supervised learning [Salimans *et al.* 2016]. Whereas for frontalisation methods we might dispute the usefulness of the GAN due to the additional resources required at test time, data-augmentation can be performed entirely offline. Therefore, even though this purely 2D method does not exploit additional information (such as that which might be provided by a 3D model), it may be able to make better use of training data. Specifically designed teacher student methods, however, would probably outperform semi-supervised methods such as this. It is more plausible that augmentation and normalisation methods making use of additional information, such as 3D models, would be more beneficial to FR. In the following section we present a selection of these methods.

2.2.4 Augmentation of information

As discussed in the previous section, it is not clear that direct generative 2D methods can help improve facial recognition. Improvements shown for 2D pose normalisation methods all but disappear when compared against an appropriate baseline, and results for 2D augmentation are noisy at best. It is more plausible that FR might be improved by injecting new forms of information into the training rather than just recycling existing training data. Plausible methods include those described in Sections 2.2.1 and 2.2.2, although final 2D translation steps may jeopardise the integrity of the modified images. In [Kortylewski *et al.* 2018] data-augmentation using synthetic identities sampled from a 3DMM is used. Unlike [Gecer *et al.* 2018], however, a GAN is not used to retrospectively increase the realism of the data. Instead, FR networks are pre-trained using the synthetic data and then fine-tuned on real data. Such pre-training is shown to consistently improve FR performance as can be seen from the results in Table 2.6. It is possible, however, that some 3D

Datasets	Multi-PIE Accuracy	LFW Accuracy	IJB-A TAR@FAR=0.1
CASIA	91.2%	94.1%	86.8%
SYN-1M + CASIA	93.3%	95.8%	90.6%
SYN-2M + CASIA	95.4%	96.0%	92.4%

Table 2.6: Performances on various metrics achieved by a FaceNet-NN4 network [Schroff *et al.* 2015] trained using CASIA and synthetic data sampled from the Basel 3DMM. Synthetic data is used for pre-training only followed by fine-tuning on CASIA. Information taken from [Kortylewski *et al.* 2018].

information is being lost during fine-tuning. Ideally the 3D model would be realistic enough to be able to forgo the fine-tuning stage.

Another method providing potential value is the partially 3D method of [Shen *et al.* 2018]. The method is a disentangled GAN, conditioned on pose and is evaluated for FR via frontalisation of images in LFW and IJB-A. In addition to a constraint on the identity, a pre-trained estimator of 3DMM shape coefficients is applied. This estimator network introduces new information into training of the GAN and can be expected to improve the preservation of identity upon frontalisation of images. Unfortunately quantitative comparisons are, again, only made with other methods in the literature and so it is not clear whether the method is beneficial to FR or not. The largely 2D method of [Yin *et al.* 2017] also includes an estimator of 3DMM coefficients. However, unlike [Shen *et al.* 2018], there is no constraint in the loss function to ensure that this information is used correctly. In [Guo *et al.* 2018], 3D model fitting is performed but not for the purpose of adjusting pose. Instead, the model is used to correctly position and render one of four pairs of 3D-modelled pairs of glasses. Evaluation of synthetically augmented training shows improvement on their "MeGlass" evaluation dataset derived from the MegaFace dataset [Kemelmacher-Shlizerman *et al.* 2016]. Faces are also augmented with glasses in [Lv *et al.* 2017] but using 2D shapes rather than 3D models. Nevertheless, information is still being injected into the training as opposed to [Sáez Trigueros *et al.* 2021] that learns to generate glasses from patterns in existing training data.

Method	Base	Network	ID vec	Crop	Loss	Method	CFP-FP
Human	-	Brain	-	-	-	-	94.57%
[Tran *et al.* 2019]	CASIA	DR-GAN	320	96x96	-	Feature disentanglement (n=1)	90.82%
[Tran *et al.* 2019]	CASIA	DR-GAN	320	96x96	-	Feature disentanglement (n=6)	93.41%
[Peng *et al.* 2017]	CASIA, (300WLP, Multi-PIE)	CASIA-net	512	??x??	Softmax +CRL	Augmentation of existing IDs	93.76%
Baseline	CASIA	ResNet-27	512	112x112	Softmax	-	87.74%
[Deng *et al.* 2018]	CASIA, (Multi-PIE, UMDFaces)	ResNet-27	512	112x112	Softmax	Augmentation of existing IDs	93.09%
[Deng *et al.* 2018]	CASIA (Multi-PIE, UMDFaces)	ResNet-27	512	112x112	Softmax	Normalisation to 15°	94.05%
Baseline	CASIA	ResNet-50	512	112x112	ArcFace	-	95.56%
[Gecer *et al.* 2020]	CASIA	ResNet-50	512	112x112	ArcFace	Augmentation with synth IDs	97.12%

Table 2.7: A selection of state-of-the-art results evaluated on the frontal-profile protocol of the CFP dataset. Where available, baseline experiments from the respective papers have been included. In the cases of [Peng *et al.* 2017] and [Deng *et al.* 2018], pose-manipulation networks were trained using additional 2D FR datasets (shown in parentheses) that should strictly have been included in the baseline experiments. CRL refers to the additional "Cross-pose reconstruction loss" of [Peng *et al.* 2017]. We also note the number of images used to form templates by the method of [Tran *et al.* 2019] whose best results were achieved for $n = 6$ probe images.

As previously mentioned, an ideal method for augmentation of pose would involve being able to generate training images directly by rendering a more realistic 3D model. The method of [Gecer *et al.* 2020] comes very close to this by training a GAN to generate synthetic samples from the distribution of a high-quality training set of scanned 3D face-shapes and textures. The textures and shapes for specific, synthetic identities are then combined with generic UV maps for sub-surface scattering, translucency, specular intensity, roughness, detail normals and their weights, and are rendered using the *Marmoset Toolbag* [Marmoset 2019] using illumination parameters from the Gaussian distribution of the 300W-LP dataset [Zhu *et al.* 2016]. A dataset of 10,000 synthetic identities with 50 images each is generated and added to CASIA-Webface in order to improve the performance of large-pose FR. Evaluation on the frontal-profile protocol of the CFP dataset is shown in Table 2.7. The method shows good improvement over the baseline (trained on CASIA only) but cannot be easily compared with other methods in the literature due to use of different FR networks and training losses. Advantages of the method are that is makes use of additional 3D information and it does not intro-

duce unwanted variation in identities since no reconstruction of existing identities is performed. The main disadvantage of the method is it's dependence on 3D-scanned training data, which is costly to collect. Also, the model covers the facial region only, rather than the full head. The 3D GAN method proposed in Chapter 5 allows generation of realistic, synthetic identities whilst avoiding both of these issues.

2.3 Summary

The work of this thesis makes extensive use of GANs; in particular the Wasserstein GAN. In Section 2.1 we described the development of GANs, beginning with their evolution from Botlzmann machines via the Generative Stochastic Network (GSN). GANs transformed the problem of unsupervised distribution-learning into one of supervised classification of real and synthetic images. By taking advantage of deep classification networks to judge whether arbitrary images belong to the real distribution or the model distribution, it became possible to train deterministic generators to map random latent space configurations to arbitrary images from the data-generating distribution, thus avoiding the need for maximum likelihood training and stochastic sample-generation. It was shown that training generator networks to minimise the standard cross-entropy loss for binary classification is equivalent to minimising the Jensen-Shannon divergence (JSD) between the model distribution and the data distribution. It was also shown, however, that the JSD saturates for disjoint distributions, which tend to be common in very high-dimensional spaces such as images, leading to weak training gradients. The alternative, non-saturating loss helps to avoid weak gradients but can lead to instabilities if training of the generator and discriminator is not well-balanced. The Wasserstein GAN was introduced to stabilise training but requires regularisation of gradients. We described how the widely adopted quadratic gradient penalty of [Gulrajani *et al.* 2017] impedes convergence of the Wasserstein GAN and argued the case for the one-sided penalty of [Chen *et al.* 2018], which is the form of regularisation used in the work of this thesis. Finally, we described the conditional GAN which allows disentanglement of the latent space and has many practical applications, for example in

data-augmentation for FR.

In Section 2.2 we gave an overview of recent works in the literature that attempt to improve the performance of FR by either augmenting training data or by normalising probe data at test time. We observed that data-augmentation methods target a multitude of nuisance factors whereas normalisation methods tend to specialise, on pose or expression in our examples. Works employing "classical 3D" methods of manipulating pose were then described and were shown to suffer from problems due to self-occlusion. Many works propose the use of adversarial losses as a potential solution, allowing removal of artefacts in synthetic images. Some methods attempted to bypass the 3D manipulation stage altogether, generating new poses directly. These methods, however, rely on identity constraints which raises the question of how useful they could actually be for improving FR. We showed that many works in the literature employing direct, 2D generative methods for normalisation trained their normalisation networks using additional data that was excluded from their baseline experiments. It is therefore not clear that these methods add value, especially given the additional resources required to perform normalisation at test time. We argued that FR is more likely to be improved via injection of new information into training rather than just re-cycling existing training data via 2D generative neural networks. The methods of [Deng *et al.* 2018] and [Gecer *et al.* 2020] represent the state of the art in pose-normalisation and augmentation of pose for both existing and synthetic identities. Better performance on the CFP dataset is achieved by augmentation using synthetic identities generated by the method of [Gecer *et al.* 2020]. However, comparison is not clear due to different experimental settings. In chapter 5, our 3D GAN method is evaluated against both.

Chapter 3

Do GANs actually generate new identities?

3.1 Introduction

It is widely accepted that Generative Adversarial Networks (GANs) are able to generate images of new identities when trained on datasets of face-images; so much so that websites such as https://thispersondoesnotexist.com/ were created to publicise the work of [Karras *et al.* 2018], [Karras *et al.* 2019] and later [Karras *et al.* 2020]. Can we be sure, however, that generated images of faces do in fact depict imagined subjects? This is an important question for applications such as data-augmentation and also data-anonymisation. Typically, qualitative evidence is presented in the form of a handful of generated images accompanied by their nearest neighbours from the training dataset. The visual differences between the identities are usually enough to satisfy the reader that new identities are indeed being generated. However, in the same way that we find doppelgangers in the real world, it is inevitable that, occasionally, images of subjects will be generated that resemble subjects in the training dataset. The question is how often does this occur?

Since there is no reason for standard GANs to differentiate facial features, important for facial recognition (FR), from any other image feature, assessing the degree to which synthetic identities resemble those of the training dataset is essentially equivalent to assessing the degree of overfitting of the generator to the training dataset. Measuring the degree of overfitting is an aspect of GAN training that is typically neglected, with most metrics assessing only the quality and vari-

ety of images. In this chapter we apply a state-of-the-art FR network to compare images synthesised by GANs with those of the training dataset. By comparing distributions of matching scores we are able to conclude that GANs do indeed generate images of imagined identities. We also show that the same method can be used as a proxy for measuring the amount of mode-collapse suffered by GANs. We begin by identifying the shortcomings of existing GAN metrics in Section 3.2.1 before elaborating on the importance of the ability of GANs to generate new identities in Sections 3.2.2 and 3.2.3; in Section 3.3 we describe our assessment of various real and synthetic datasets and present results; finally, in Section 4.5 we make concluding remarks.

3.2 Related work

3.2.1 GAN metrics

The *de facto* standards for assessing GANs are currently the Inception Score (IS) [Salimans *et al.* 2016] and the Fréchet Inception Distance (FID) [Heusel *et al.* 2017]. Both of these metrics, however, suffer from two main issues:

1. They make strong assumptions about the probability distributions of data. For example, IS makes the assumption that the distribution of images can be well represented by the distribution across ImageNet categories, whereas FID assumes that the distribution can be approximated by a multi-variate Gaussian, which can lead to ambiguity between the correct distribution, and a distribution with dropped modes but additional, spurious modes that act to give the same mean and variance.

2. They summarise measures of both the quality of generated images and the variety of images in a single score. This can lead to ambiguity between, for example, distributions containing high quality images from a single mode compared to distributions containing multi-modal but low quality images.

The more recent methods of [Sajjadi *et al.* 2018] and [Kynkäänniemi *et al.* 2019] avoid the second issue above by providing two separate measures: Precision as a

measure of image-quality, and Recall as a measure of image variety. Values of Precision and Recall (P&R) are derived from the overlap between estimates of the forms of the real and synthetic data distributions in a discriminative feature space. The models used to represent data distributions are more general than those used by IS and FID, and so go some way to solving the first issue above. None of these metrics, however, is sensitive to the problem of overfitting. A "Memory GAN" that simply memorises and reproduces exact copies of the training data will produce optimal IS, FID and P&R scores. In this chapter we wish to assess the ability of GANs to generate new identities, i.e. to ensure that GANs' generators do not overfit training datasets with respect to identity. None of the aforementioned metrics is able to provide this information.

The lack of a suitable metric for assessing overfitting was identified in [Lucic *et al.* 2018] although, despite being stated as one of the motivations for proposing their measure of P&R calculated for toy datasets, the paper does not elaborate on how overfitting might be measured. In [Webster *et al.* 2019], overfitting is measured by analysing discrepancies in the distributions of reconstruction error for training and validation images upon inversion of the generator. The method appears to be informative. However, in practice it is expensive to invert the GAN's generator for all training images and, on a more fundamental level, the invertibility of the generator for a particular image does not necessarily reveal the proclivity of the generator to generate that image. In the work of this chapter, we efficiently project datasets of facial images into a biometric feature-space where comparisons of the generated distribution and the training distribution can be made based on subtle yet robust features. Doing so is primarily motivated by our interest in identity. However, biometric datasets may serve as a useful standard for evaluating GANs in general.

3.2.2 Data-anonymisation

With the recent implementation of the General Data Protection Regulation (GDPR) in Europe, and similar regulations elsewhere, the anonymisation of data is increasingly becoming a topic of interest. Unrestricted use of images and datasets

of images containing faces depends on the ability to remove information that allows identification of the original subjects. A crude method of doing this might be simply to detect faces and to set those pixels to zero. Ideally, however, we would like to remove identity information without affecting other semantic properties of the image or of the dataset as a whole.

Most methods in the literature tackle the problem of data-anonymisation by modifying individual images, leaving all properties untouched except for the identity. Some of the more successful early attempts at doing so involved reconstructing the subject of the original image using a 3D morphable model (3DMM) [Blanz & Vetter 1999] before modifying the model's shape parameters to change the identity [Gross *et al.* 2008, Samarzija & Ribaric 2014]. In [Meden *et al.* 2017], a CNN conditioned on biometric vectors is used to generate a face-image that is a mixture of nearby identities before blending it back into the original image. More recent methods treat the problem as image-to-image translation or as image in-painting. In [Wu *et al.* 2019] an auto-encoder is trained to optimise a reconstruction loss, but also a biometric loss to enforce distance between the original and translated images in an identity feature-space. In [Sun *et al.* 2018a] and [Hukkelas *et al.* 2019] faces are first obscured and then auto-encoders trained to in-paint the obscured regions, harmonising them as best they can with the rest of the image in order to fool an adversarial loss. In [Sun *et al.* 2018b] a hybrid 3DMM plus in-painting method is used.

Due to the constraint of having to maintain the specific contexts of individual images, the aforementioned problem is difficult and generated images tend to be of low quality, either lacking in high-frequency detail or containing obvious artefacts. A simpler task which generally results in generated images of higher quality is to train a GAN whose sole objective is to accurately approximate the distribution of real images as a whole. In this case, synthetic images sampled from the learned distribution retain much of the character of the original dataset but the specifics of images generally differ. As mentioned in Section 3.2.1, although much work has been done to assess the quality and variety of GAN-generated images, there is no work explicitly demonstrating such GANs to consistently generate new identities.

3.2.3 Augmentation using synthetic identities

In the work of this thesis we aim to make use of synthetic identities for data-augmentation. Synthetic data-augmentation might take one of two forms:

1. Supervised: generation of sets of ID-labelled images where each set contains a "unique" identity;

2. Semi-supervised: generation of examples of arbitrary identity (the goal being to avoid classifying them as the labelled identities of the training dataset).

We are aware of only a handful of examples of "supervised" data-augmentation using synthetic identities. These were identified in Table 2.1 of the Literature Review. Both [Gecer *et al.* 2018] and [Gecer *et al.* 2020] train FR networks using both real and synthetic identities simultaneously. This makes it important to be sure that new, synthetic identities are being generated to ensure that collisions with existing identities are rare. If this were not the case, there would be undesirable consequences for the compactness of learned class-representations in the feature-space. In [Kortylewski *et al.* 2018] synthetic identities were used for pre-training only and so generation of new identities is less important. As noted in Section 2.2.4, however, pre-training followed by fine-tuning on real data is probably not ideal due to forgetting of information from the synthetic dataset.

To our knowledge, there are no evaluations of semi-supervised learning for facial recognition in the literature using synthetic data generated by GANs. However, the technique has been evaluated for the similar problem of person re-identification [Zheng *et al.* 2017]. Although not pursued in this thesis, the concept of semi-supervised learning is discussed in the following subsection to help emphasise the importance of being able to generate new identities.

3.2.3.1 Semi-supervised, synthetic augmentation

Semi-supervised learning using GANs was originally proposed in [Salimans *et al.* 2016] (the same paper to introduce the Inception Score), and concurrently in [Odena 2016]. These works proposed what we will term a

"$k+1$" method in which the classifier that they are trying to improve (which has k classes) is also trained to be the discriminator of a GAN by assigning synthetic images to an additional, $k+1^{th}$ class. Why this was found to improve classification of real images is not entirely clear. One might imagine that it has something to do with multi-task learning, or alternatively that the larger generator+classifier network acts as teacher of the smaller, student network that is the classifier alone.

It was later shown in [Dai *et al.* 2017] that classification using a semi-supervised, $k+1$ training method is optimised when using a "bad" GAN, i.e. a GAN with a "complement generator", trained to generate data samples that lie *outside* the regions in feature space that describe real members of the classes. The reason for this is that, when using a complement generator, the classifier learns decision boundaries at the edge of, rather than within, the feature manifold of each class. With a typical generator, decision boundaries might stray within the manifold if real examples happen to look too much like synthetic samples. For the case of augmenting facial recognition datasets, generated samples should therefore represent novel identities that lie in distinct regions of an image feature space.

The findings of [Dai *et al.* 2017] suggest that it might not be straightforward to train a classifier with a $k+1^{th}$ class using realistic face images generated by a standard GAN. The assumption that the GAN generates *only* new identities might be too strong. An alternative semi-supervised training method that makes a weaker assumption is Label Smoothing Regularisation for Outliers (LSRO) [Zheng *et al.* 2017]. LSRO takes a randomly generated set of synthetic images and assigns each a label vector with uniformly distributed values, as opposed to a one-hot vector indicating a specific class. This label-smoothing for the synthetic images is thought to have a regularising effect on the classifier, reducing the classifier's confidence when seeing identities that are similar to those in the training set. Even for the LSRO method, however, it is still important that the synthetic training set does not overfit to the identities of the training dataset. In the following section we show results that explicitly confirm this to be the case.

3.3 Results

As previously discussed, existing GAN metrics (e.g. IS, FID and P&R) measure the quality and variety of generated images but do not indicate whether overfitting to the training dataset has occurred. Here, we assess the overfitting of the official versions of the Progressive GAN [Karras *et al.* 2018], and of the more recent StyleGAN [Karras *et al.* 2019], by generating sets of synthetic face-images and comparing them against the training dataset using a state-of-the-art biometric network based on [Deng *et al.* 2019a]. We show results for two versions of StyleGAN: the default, validation version, as is used to generate the statistics presented in the paper, and a version with StyleGAN's style-mixing enabled during generation, labelled as "mix" in the Figures and in Table 3.1. Style-mixing is typically only enabled during training of StyleGAN to help disentangle different scales. However, we note that reducing the variability of input to StyleGAN by disabling style-mixing causes increased mode-collapse. Each GAN was trained on the CelebA-HQ dataset containing $30,000$ images at 1024×1024 resolution of $6,217$ labelled identities. Labelling in CelebA (and therefore in CelebA-HQ) is noisy, with many subjects belonging to more than one ID category. This causes the level of similarity between identities in the real data to be overestimated. For this reason, we have also included results for an additional, proprietary dataset of mugshot images with clean labels. This dataset was not used to train the GANs being assessed here and is only included for comparison. Datasets of $30,000$ images of random, synthetic identities were generated for each GAN.

The matching scores used in our analysis are analogous to cosine similarities in the feature space of a biometric network. Feature vectors were generated and then converted to scores such that higher scores represent stronger similarity between identities. Two images might be considered to contain the same identity if the matching score is larger than, for example, 3614.5. This threshold corresponds to a false acceptance rate (FAR) of 1×10^{-4} for the CelebA-HQ dataset, as can be seen from Figure 3.1. In Section 3.3.1 we discuss results showing that new identities are indeed generated by both GANs; then in Section 3.3.2 we describe how the same

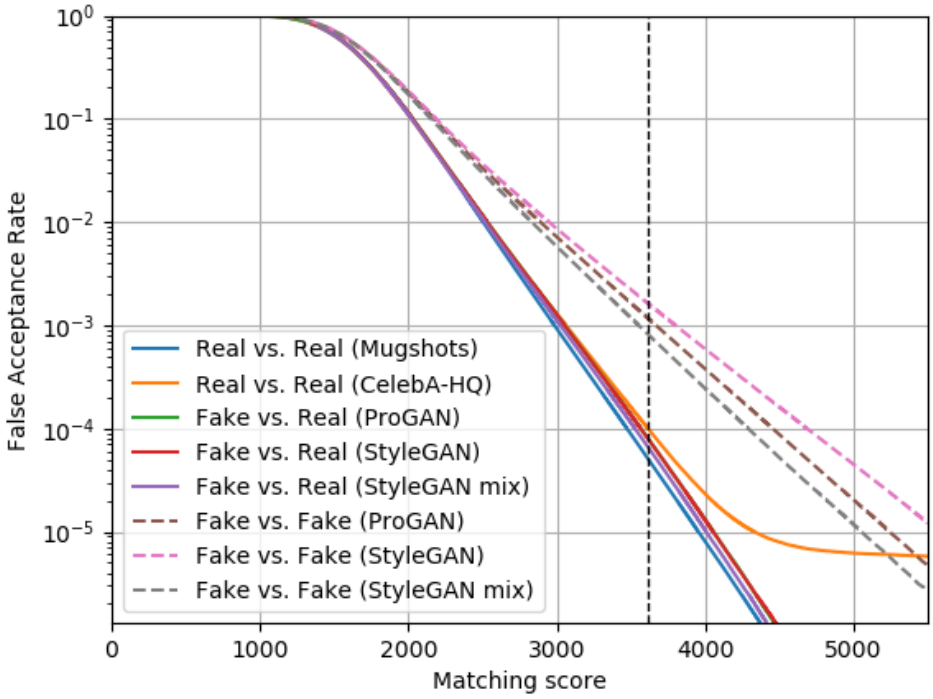


Figure 3.1: False acceptance rates across biometric matching score thresholds for all pairs of non-mated images either within or between real and synthetic datasets as indicated in the legend.

matching scores can be used as a measure of mode-collapse.

3.3.1 Generation of new identities

To assess overfitting, we compare each synthetic image with all images from the training dataset (CelebA-HQ). Inevitably, some of the images match more closely than others. There is no score threshold that we can define, however, that tells us when any individual image has "overfit" to the training dataset. We wish for the GAN to approximate the real-world distribution of images of faces and, as in the real world, sometimes look-alikes do occur. Instead, we must ensure that the frequency at which look-alikes occur is not significantly greater than within the training dataset itself.

Figure 3.1 shows false acceptance rates as a function of matching score threshold, i.e. the curves are the normalised cumulative distributions of all non-mated

matching scores within and between various combinations of dataset. The curves labelled "Real vs. Real" show the distributions within the proprietary "Mugshots" dataset, and within CelebA-HQ. By definition, if we were to choose a matching score threshold of 3614.5, at an FAR of 1×10^{-4} for CelebA-HQ, one in every $10,000$ randomly selected pairs of identities would be found to be look-alikes. It can be seen that, within CelebA-HQ, the proportion of identities found to match remains relatively high even for large matching score thresholds of $5,000$ and above. This is due to the previously mentioned labelling problems within CelebA. Despite these problems, we have left the curve as an example of the dynamics that can be expected if duplicate identities are present. The three curves labelled as "Fake vs. Real" show the distributions of matching scores between the dataset generated by the indicated GAN and CelebA-HQ. (Note that the green curve of the Progressive GAN is hidden beneath the red curve of the standard StyleGAN.) Each of these curves demonstrates consistently lower matching frequencies than for the images of CelebA-HQ. What is more, the matching frequencies are in close agreement with those of the clean, Mugshots dataset indicating that the synthetic datasets do not resemble CelebA-HQ significantly more than you would expect non-mated identities to resemble one-another in a dataset of real images. The dashed "Fake vs. Fake" curves will be discussed in Section 3.3.2.

Figure 3.1 shows distributions of matching scores for all image combinations, even those that map to distant parts of the identity feature-space. An alternative way to assess overfitting is to look only at the largest score for each image, i.e. to observe the distributions of nearest neighbour scores. These are plotted in Figure 3.2. Upon interpreting these results, we notice that the story is essentially the same: each of the three "Fake vs. Real" curves displays lower nearest neighbour matching scores than are found for CelebA-HQ, and they are in close agreement with the scores for the clean, Mugshots dataset. In fact, the distribution for "StyleGAN mix" has shifted to consistently lower scores than for the Mugshots dataset. This change in the relative positions of the curves is difficult to interpret. Examples of strongly matching image pairs are shown in Figure 3.3 accompanied by the corresponding matching score. Images with blue borders in the first three rows are images taken

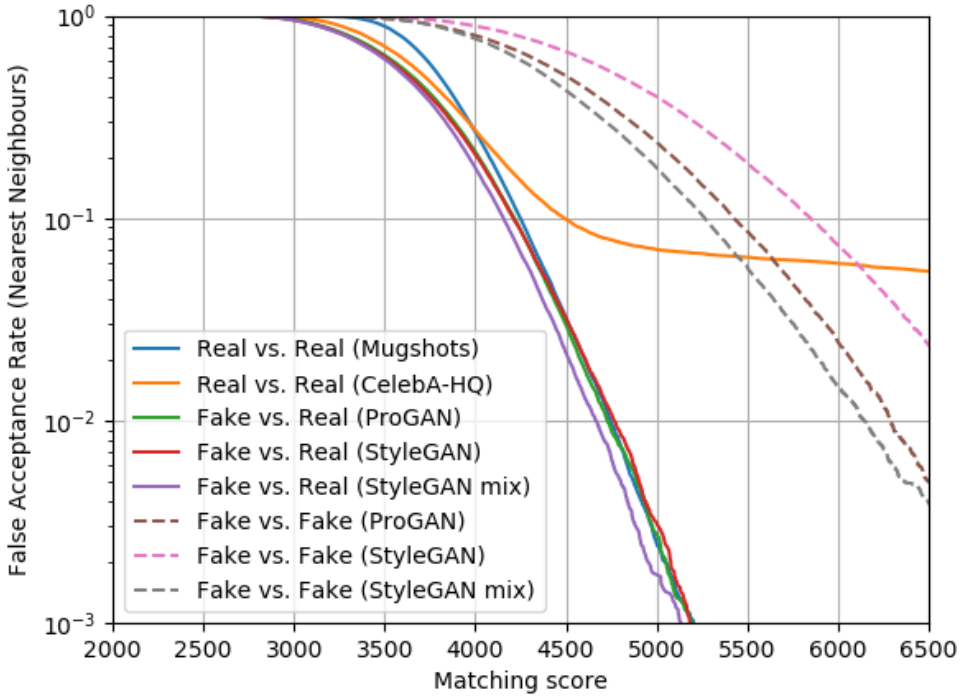


Figure 3.2: False acceptance rates of nearest neighbour images across across biometric matching score thresholds for non-mated image-pairs either within or between real and synthetic datasets as indicated in the legend. Examples of the synthetic images are given in Figure 3.3.

from CelebA-HQ, and images with green borders were generated by StyleGAN with style-mixing enabled. We include these images for completeness and have no need to draw further conclusions from them with regard to the frequencies of occurrence of look-alikes. It should be noted that, although we attempted to show the strongest non-mated matches within the CelebA-HQ dataset, due to labelling issues some stronger matches may have been missed due to doubts as to whether the images depicted the same subject or not. The selection may therefore be biased towards weaker matches that more obviously depict different subjects. One thing that can be noted from Figure 3.3 is that stronger matches tend to be found for female subjects. CelebA-HQ contains roughly two-thirds female subjects, which is sure to contribute to this effect. Lack of robustness of the FR algorithm to make-up may also contribute.

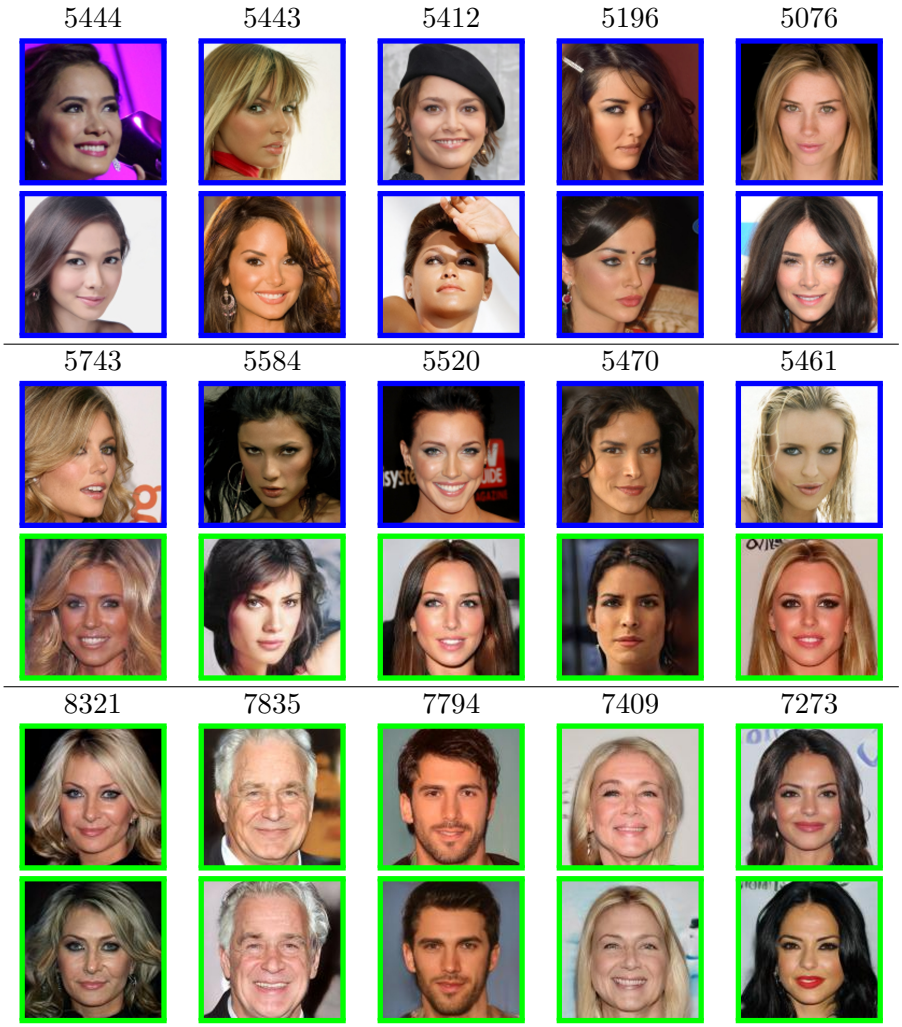


Figure 3.3: A selection of non-mated image-pairs displaying strong matching scores. Images with blue borders (first three rows) were taken from CelebA-HQ; images with green borders (final three rows) were generated by StyleGAN with style-mixing enabled during generation.

Comparison	False Acceptance Rate @3000	False Acceptance Rate @3614.5
Real vs. Real (CelebA-HQ)	1.29×10^{-3}	1.00×10^{-4}
Fake vs. Real (ProGAN)	1.26×10^{-3}	0.80×10^{-4}
Fake vs. Real (StyleGAN)	1.25×10^{-3}	0.80×10^{-4}
Fake vs. Real (StyleGAN mix)	1.11×10^{-3}	0.66×10^{-4}
Fake vs. Fake (ProGAN)	7.19×10^{-3}	1.16×10^{-3}
Fake vs. Fake (StyleGAN)	8.64×10^{-3}	1.62×10^{-3}
Fake vs. Fake (StyleGAN mix)	5.80×10^{-3}	0.82×10^{-3}

Table 3.1: FARs read from Figure 3.1 at two thresholds.

3.3.2 Mode-collapse of identity

In this section we discuss interpretation of the "Fake vs. Fake" curves (dashed lines) of Figures 3.1 and 3.2. These curves represent the distributions of matching scores *within* the synthetic datasets and show much higher frequencies of strong matches than are found within or between the other datasets (with the exception of the noisily labelled CelebA-HQ at certain thresholds). This indicates that, although synthetic subjects do not strongly resemble those of the training dataset, they do strongly resemble one-another. This is a symptom of the well-known problem of mode-collapse in GANs in which well-separated, random input vectors are mapped to similar points in image-space. Examples of strongly matching, non-mated synthetic images are presented in the final two rows of Figure 3.3.

Similar to the case for overfitting, since the Progressive GAN and StyleGAN have no reason to treat identity features differently from any other image-feature, we can use the increase in "Fake vs. Fake" matching frequency as a general measure of mode-collapse in GANs. One might try to quantify the degree of mode-collapse as a single value by measuring the number of distinct identities being generated and representing this as a fraction of that for the dataset of real images. For example, the threshold of 3614.5 was chosen to give an FAR of 1×10^{-4} for CelebA-HQ. This means that the algorithm, in conjunction with this threshold, is capable of distinguishing $1/1 \times 10^{-4} = 10,000$ real identities. When applied to a synthetic dataset (say "StyleGAN mix"), the biometric network finds only $1/0.82 \times 10^{-3} = 1220$ dis-

tinguishable identities, implying that the synthetic identities have collapsed to span a region of only $1220/10,000 = 12.2\%$ of that of the real distribution. However, by performing the equivalent calculation using values from Table 3.1 for the cruder threshold of 3000, we find that the estimate of the level of mode-collapse improves to $1.29/5.80 = 22.2\%$; i.e. the estimation of the degree of mode-collapse is dependent on the precision with which data-points are represented in the feature space. To avoid this complication, we therefore recommend making comparisons between datasets by analysing the dynamics of the FAR at all thresholds, as presented in Figure 3.1.

3.4 Conclusion

We performed an analysis of the ability of GANs to generate new identities. In doing so, we introduced a technique of analysing both the degree of overfitting of generators to the training dataset, and the degree of mode-collapse. We used this technique to show that overfitting is minimal and that GANs trained on face-images are therefore capable of generating new identities. This validates the use of GANs for data anonymisation and also validates the assumptions made when performing data-augmentation with synthetic identities, both supervised and semi-supervised, for example, the assumption that synthetic images can indeed be safely assigned to a $k + 1^{th}$ class without detrimentally affecting existing classes.

Chapter 4

Disentanglement of identity in GANs

4.1 Introduction

In order to be able to perform supervised data-augmentation for facial recognition (FR) using GAN-generated, synthetic identities, it must be possible to control factors of variation in images without affecting the identity. In Section 4.4 of this chapter we present an analysis of the ability of various formulations of GAN to disentangle identity from other image characteristics in their latent spaces. Ultimately we conclude that the level of disentanglement of identity in 2D GANs is not sufficient for the data to be of use for supervised data-augmentation for FR. Before discussing the results of this study, we introduce two contributions in the field of disentanglement in GANs. Although perhaps not useful for data-augmentation, these developments may find suitable applications in other areas such as generating anonymised evaluation sets of mated images or image-editing. The first contribution is the Intra-class Variation Isolation (IVI) mechanism [Marriott *et al.* 2020a] that allows the learning of disentangled, multivariate representations of variation in images using only the weak supervision of simple, binary labels. IVI is introduced in Section 4.2. The second contribution is the integration of a triplet loss term into the loss function of an SD-GAN [Donahue *et al.* 2018] to improve the disentanglement of identity [Marriott *et al.* 2020b]. Both of these methods are assessed in the final disentanglement study alongside a third method proposed by [Shen *et al.* 2020].

4.2 Taking control of intra-class variation under weak supervision

While standard GANs can generate realistic images, the precise form of these images cannot be easily controlled. The subject of the images is dependent in some way on the values of the random input vector, but *a priori* we do not know in what way. An obvious solution to this problem is the Conditional GAN (cGAN), in which the GAN's discriminator is trained to distinguish *real image+label* pairs from *fake image+label* pairs, thereby encouraging the generator to produce images that correctly correspond to the label upon which it is conditioned. In their typical form, cGANs are trained under the strong supervision of these labels: given binary category labels, images belonging to certain categories can be generated (but whose specific form is still governed by the entangled, real-valued random vector); and given continuous, real-valued labels, cGANs can be forced to obey more specific semantics. Previous works, for example, have conditioned on binary hair-colour labels [Choi *et al.* 2018], and on continuous expression action unit magnitudes [Pumarola *et al.* 2018]. However, the annotation of training images with descriptions of multi-dimensional phenomena such as expression and lighting, is notoriously difficult. In this section we introduce a method that enables GANs to learn continuous, multi-variate representations of variation without the need for precise annotation of training images. Training data need only be annotated with binary labels indicating the presence or absence of a particular form of variation, for example, "ambient lighting / non-ambient lighting". We coin the method Intra-class variation isolation (IVI) and the resulting network the *IVI-GAN*. A video demonstrating the continuous variation of various isolated parameter sets can be viewed at https://youtu.be/hoWOFeADwdY.

4.2.1 Related work

Most recent works aiming to manipulate the semantics of synthesised images take the form of image-translation networks or auto-encoders with adversarial losses added to help ensure that images look realistic. In [Lai & Lai 2018] face images are

translated from being "at pose" to frontal while a discriminator ensures realism and that expression remains constant. The identity and other image properties, however, are only preserved via a pixel-wise comparison of the generated image with a ground-truth frontal image. Since such corresponding pairs of images are rarely available, other researchers have instead proposed auto-encoding methods such as CycleGAN [Zhu *et al.* 2017], whose image-shaped latent space is trained to resemble the distribution of a different class of images, not necessarily paired with the input image. For example, in both [Choi *et al.* 2018] and [Bozorgtabar *et al.* 2019], a discriminator is used to encourage generated images to belong to different expression categories, while the cycle-consistency loss ensures that the general structure of the image remains unaffected, thus implicitly preserving the identity. In [Pumarola *et al.* 2018], continuous expression action unit labels are used to provide continuous control over the transformed images, rather than just control over the expression category. The downside, however, is that precise, real-valued labels are difficult to obtain. In addition, methods relying on cycle-consistency losses cannot be easily used to manipulate pose since this involves more significant alterations to image structure, thereby destroying the implicit constraint on identity.

Other works consider identity preservation explicitly by adding biometric losses. In [Lindt *et al.* 2019] a pre-trained biometric network is added as a constraint during manipulation of expression. In [Bao *et al.* 2018] a biometric network is trained in parallel with the generative network whereas in [Tran *et al.* 2019] the GAN's discriminator itself is trained to classify the identity as a secondary task. Both [Bao *et al.* 2018] and [Tran *et al.* 2019] are able to convincingly modify the pose of input images. All of these methods, however, require strong supervision to control image properties: in [Tran *et al.* 2019] fine-grained pose-category labels are needed and in [Lindt *et al.* 2019], similar to [Pumarola *et al.* 2018], continuous expression labels are needed.

InfoGAN [Chen *et al.* 2016] is the only work of which we are aware that attempts to disentangle control of image properties in an entirely unsupervised manner. By maximising the mutual information between a small subset of input parameters and the generated images, those parameters are attributed control over the most signif-

icant forms of variation in the image dataset. In practice these forms of variation tend to resemble semantics such as pose and lighting. However, there is no guarantee of which semantics will be learned, nor that identity will be preserved. Rather than an entirely unsupervised approach, in this work we propose a weakly supervised alternative, compromising between the unpredictable semantics of InfoGAN and the strong requirement for precise labels of typical conditional generation methods. Our weakly supervised technique allows control of specific forms of variation but does not require a prior model of that variation.

Concurrent with our own work, [Shen *et al.* 2020] demonstrates an alternative method for achieving continuous control over image properties using only binary labels. Rather than training a conditional GAN on labelled training images, labels are used retrospectively to identify the principal directions of variation of various attributes in the latent space of a pre-trained GAN. In their work, no special attention is paid to preserving identity. Also, in the absence of multiple labels, their method is not capable of identifying multi-variate forms of variation such as illumination or the configuration of the background.

4.2.2 Method

Intra-class Variation Isolation (IVI) can be implemented in any conditional GAN. Our best results are achieved using the Wasserstein loss and so we present this version. The original loss functions for a conditional Wasserstein GAN were given in Chapter 2. We restate them here for convenience.

$$\mathcal{L}_{\theta_D} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{data}}[D(\mathbf{x},\mathbf{y};\theta_D)] - \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\rho}\sim p_\rho}[D(G(\mathbf{z},\boldsymbol{\rho};\theta_G),\boldsymbol{\rho};\theta_D)] \tag{4.1}$$

$$\mathcal{L}_{\theta_G} = \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\rho}\sim p_\rho}[D(G(\mathbf{z},\boldsymbol{\rho};\theta_G),\boldsymbol{\rho};\theta_D)] \tag{4.2}$$

In order to control some attribute of a generated image in a continuous fashion, for example the pose, a typical cGAN requires each training image to be labelled with a precise pose-estimation, $\mathbf{y} \in \mathbb{R}^n$. Intra-class variation isolation, on the other hand, requires only the weak supervision of binary category labels indicating whether or not a particular form of variation is present. The IVI-GAN is then allowed to learn

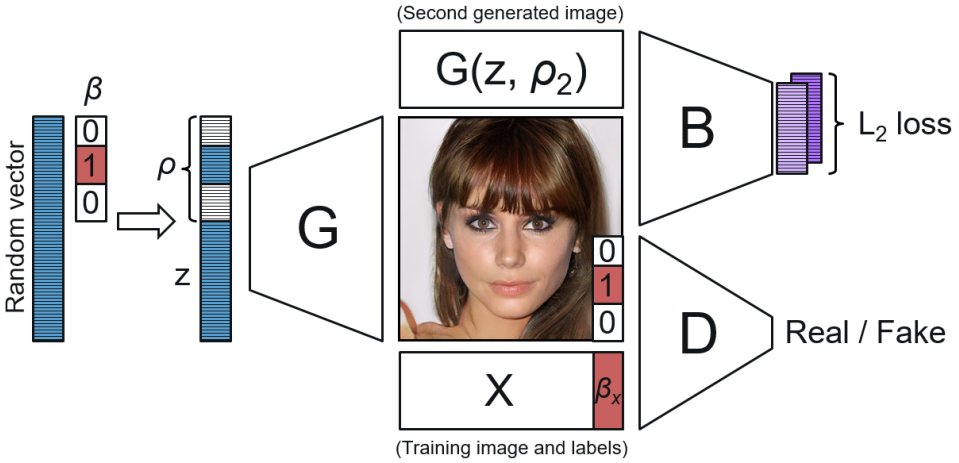


Figure 4.1: An illustration of IVI-GAN. The real-valued parameter-vector, ρ, is formed by masking sections of an extended, random vector using the randomly selected, binary label-vector, β. It is β (*not* ρ) that is then fed to the discriminator with the generated image.

its own multivariate model of that form of variation. Technically, the changes to the standard cGAN are very simple and involve modifying only the form of the labels provided to the cGAN. The loss functions to be minimised are

$$\mathcal{L}_{\theta_D} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{data}}[D(\mathbf{x},\mathbf{y};\theta_D)] - \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\beta}\sim p_\beta}[D(G(\mathbf{z},\boldsymbol{\rho};\theta_G),\boldsymbol{\beta};\theta_D)] \tag{4.3}$$

$$\mathcal{L}_{\theta_G} = \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\beta}\sim p_\beta}[D(G(\mathbf{z},\boldsymbol{\rho};\theta_G),\boldsymbol{\beta};\theta_D)] \tag{4.4}$$

where $\mathbf{y} \in \{0,1\}^n$ are now binary labels for n categories, and $\boldsymbol{\beta} \in \{0,1\}^n$ are binary labels sampled from the same distribution as $\mathbf{y}$ (but, as before, do not necessarily have to be the same). The novel aspect of the loss is the way in which the random parameters, $\boldsymbol{\rho}$, are chosen.

$$\boldsymbol{\rho} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_n \end{bmatrix} \text{ where } \mathbf{p_i} \in \begin{cases} \mathcal{N}^{q_i}, & \text{if } \beta_i = 1 \\ \mathbf{0}^{q_i}, & \text{if } \beta_i = 0 \end{cases} \tag{4.5}$$

Here, $\mathcal{N}^{q_i}$ is a random Gaussian vector of length q_i, and $\mathbf{0}^{q_i}$ is a vector of zeros.

N.B. The mechanism used to form $\boldsymbol{\rho}$ in (4.5) could be interpreted as a function

$f(\mathbf{z}', \boldsymbol{\beta}) = (\mathbf{z}, \boldsymbol{\rho})$, where $\mathbf{z}'$ is an extended random vector providing the additional random values used to form $\boldsymbol{\rho}$. Since this function could be implemented as part of a modified generator, G' (which is permitted to have an arbitrary architecture), we have $G(\mathbf{z}, \boldsymbol{\rho}) = G(f(\mathbf{z}', \boldsymbol{\beta})) = G'(\mathbf{z}', \boldsymbol{\beta})$; i.e. equations (4.3) and (4.4) are mathematically equivalent to (4.1) and (4.2) but for binary labels. Figure 4.1 depicts this intra-class variation isolation mechanism as part of our IVI-GAN.

All variation in images generated by the IVI-GAN must be derived from the combination of $\mathbf{z}$ and $\boldsymbol{\rho}$. The values of $\mathbf{z}$ are independent of $\boldsymbol{\beta}$ and so can always be used freely by the generator, irrespective of the labels. The parameter sets forming $\boldsymbol{\rho}$, however, are only available when certain forms of variation are labelled as being present. The idea is that the generator will then only use those parameters to describe that form of variation since the presence of non-zero parameters cannot be relied upon to describe anything else.

When labelling training images, it is best that $\beta_i = 0$ be used to describe a unique image property. For example, in the case of expression, $\beta_{exp} = 0$ should correspond to a neutral expression and $\beta_{exp} = 1$ to all non-neutral expressions. If $\beta_{exp} = 0$ were chosen to represent a non-unique property, such as "not smiling", then all of the different ways of not smiling would necessarily end up being encoded by $\mathbf{z}$. In the case of lighting, we chose $\beta_{lighting} = 0$ to represent "ambient lighting". This means that the colour and intensity of ambient light in our images is encoded by $\mathbf{z}$ but that all non-ambient lighting phenomena are encoded by $\boldsymbol{\rho}_{lighting}$.

We find that the IVI mechanism does indeed encourage disentanglement of labelled variation. This is aided by the natural parsimony of the generator which must find efficient ways of representing the training image distribution despite having only limited capacity. To do this, common features are, of course, reused by the generator to form different images. When making only subtle modifications to images, e.g. adding lighting effects, the generator tends to leave the general structure of images unchanged, thus implicitly preserving the identity along with other image attributes. However, when making more significant changes to images, such as modifying the pose, this implicit identity preservation cannot be relied upon. We therefore introduce an explicit constraint on the identity which is described in

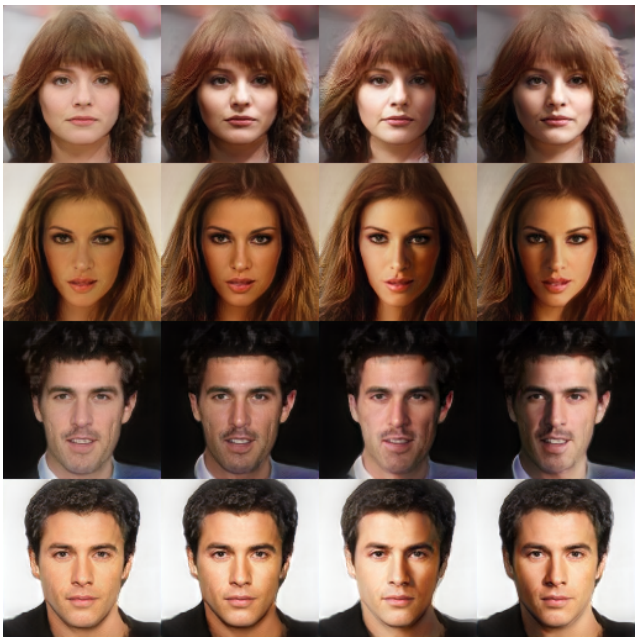


Figure 4.2: Lighting conditions manipulated by IVI-GAN for four synthetic identities. Left-hand column: $\boldsymbol{\rho}_{lighting} = \mathbf{0}^4$ (ambient); the other columns show the effect of assigning a value of 3.0 or -3.0 to individual elements of the lighting vector while keeping other parameters constant.

the following section.

4.2.2.1 The biometric identity-constraint

To ensure that identity remains constant upon adjusting other image-properties, we have added a term to the generator loss of IVI-GAN involving a pre-trained biometric network [Hasnat *et al.* 2017] that accepts a facial image as input and produces a 128-dimensional encoding of the identity.

$$\mathcal{L}_{ID} = \mathbb{E}_{\mathbf{z}\sim p_z}\left[\|B(G(\mathbf{z},\boldsymbol{\rho})) - B(G(\mathbf{z},\boldsymbol{\rho}_2))\|^2\right] \tag{4.6}$$

where B is the biometric network and $\boldsymbol{\rho}_2$ is a second set of random label-parameters. (N.B. we have dropped the θ for convenience of notation.) By running the generator twice for the same $\mathbf{z}$ but different $\boldsymbol{\rho}$, and constraining the identity encodings to remain close, we ensure that the identity is encoded as part of $\mathbf{z}$ and not affected

by changes to $\boldsymbol{\rho}$. The full generator loss is then

$$\mathcal{L}_{\theta_G} = \mathbb{E}_{\mathbf{z}\sim p_z, \boldsymbol{\beta}\sim p_\beta}[D(G(\mathbf{z}, \boldsymbol{\rho}), \boldsymbol{\beta})] + \lambda_{ID}\mathcal{L}_{ID} \tag{4.7}$$

where λ_{ID} is a hyper-parameter to be tuned. The biometric identity constraint is depicted on the right-hand side Figure 4.1.

4.2.2.2 An additional, structural constraint for lighting

Despite the natural tendency of the generator to leave the structure of images unaffected when modifying properties such as lighting, subtle unwanted changes can still be seen, e.g. small changes to pose. To avoid this, we introduce a constraint on the image structure to be used when modifying lighting conditions.

Lighting is an additive phenomenon. For example, an image of a scene with two light sources is equivalent to the sum of two images of the same scene with the two light sources acting on it separately. One way to generate an image of a face under a particular lighting condition, therefore, is to add together two constituent images of the same subject. By re-formulating our generator in this way, the constraint that the composite image must appear realistic to the discriminator ensures that features in the two constituent images are of the same general structure. If not, the composite image would appear blurry and unrealistic due to misalignment and other inconsistencies. We propose that the generator be replaced with the following:

$$G(\mathbf{z}, \boldsymbol{\rho})_{comp} = G(\mathbf{z}, \mathbf{0}) + G(\mathbf{z}, \boldsymbol{\rho}) \tag{4.8}$$

where $\boldsymbol{\rho}$ represents lighting parameters only and $\mathbf{0}$ is a vector of zeros the same length as $\boldsymbol{\rho}$ indicating that ambient lighting should be generated. In our IVI-GAN, $G(\mathbf{z}, \boldsymbol{\rho})_{comp}$ replaces G in both equation (4.3) and (4.4). By choosing $G(\mathbf{z}, \mathbf{0})$ as the second constituent image, we straightforwardly ensure that $G(\mathbf{z}, \mathbf{0})_{comp}$ generates ambiently lit images. We also find that this formulation can be used to help constrain facial structure when modifying the appearance of the background.

4.2.3 Implementation

Our implementation is built upon the stable and efficient progressive GAN of [Karras *et al.* 2018], a Tensorflow implementation of which was made publicly available by Nvidia. The progressive GAN begins by generating images of 4×4 resolution and then progressively fades in new convolutional upscaling layers until the desired resolution is reached. There has been much recent work on improving the quality of GAN-generated images published in the literature. We tested a selection of these enhancements and found that the best results were produced by a progressive Wasserstein GAN with the standard gradient penalty (GP) term of [Gulrajani *et al.* 2017] where the weight of the GP term was allowed to evolve throughout training based on an *adaptive lambda* scheme similar to that in [Chen *et al.* 2018]. In the generator we use orthogonal initialisation of weights and replace the pixel-wise feature normalisation used in [Karras *et al.* 2018] with the orthogonal regularisation of [Brock *et al.* 2017] using the suggested weight of 0.0001.

4.2.3.1 Conditioning the GAN

The way in which labels and label-parameters are used to condition GANs is an open area of research. For example, [Miyato & Koyama 2018] finds that, given certain assumptions about the form of the distribution of data, the optimal method of conditioning the discriminator should be to learn some inner-product of the label-vector with the channels at each pixel; in [Dumoulin *et al.* 2017b] the generator network is conditioned via instance-normalisation parameters. We have used the more straight-forward method of concatenating label-vectors with inputs but expect that these more sophisticated methods of conditioning may be used to improve results. We concatenate our IVI parameter vector, $\boldsymbol{\rho}$, with the random vector on input to the generator, and on input to the discriminator we concatenate the binary labels, $\mathbf{y}$ and $\boldsymbol{\beta}$, as additional channels repeated at each pixel of the real and generated images respectively.

An analysis of where best to introduce conditioning vectors to the discriminator was performed in [Perarnau *et al.* 2016]. They predicted that earlier in the network

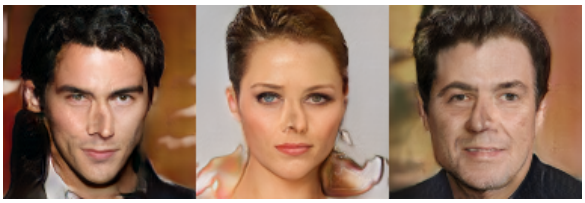


Figure 4.3: Examples of broken nose features generated during tests of an auxiliary classifier (not used in IVI-GAN).

should be better since the discriminator would be afforded more learning interactions with the information. In fact they conclude that it is best to concatenate the conditioning with the first hidden layer. However, in a progressive architecture, the first convolutional layer of the discriminator is not faded in until the final stages of training. In our case, it therefore makes more sense to concatenate the conditioning with the image where it is then scaled down and fed to each layer of the progressive network as they are being faded in. Ultimately, once all layers of the progressive discriminator are active, the conditioning information and image are only fed to the first layer of the discriminator.

Many applications make use of auxiliary classifiers (ACs) [Odena *et al.* 2017] as a way of ensuring that conditional parameters are not ignored during generation of images. We tested this method in conjunction with IVI using the auxiliary classifier already implemented in Nvidia's progressive GAN code. However, we found results to be unsatisfactory. As noted in [Miyato *et al.* 2018], auxiliary classifiers encourage the generator to produce images that are easy to classify; a goal which is not in alignment with the principal training objective of the GAN. We found that, given a large weight in the discriminator loss, the AC-term caused mode-collapse, squeezing variation into narrow, well-separated categories. For example, upon varying continuous, conditional pose parameters we observed a discrete jump in the generated pose between frontal and large poses. Giving less weight to the AC-term ameliorated the discrete jumps in pose. However, more subtle artefacts remained, such as broken noses pointing in one direction or the other; a feature obviously used by the discriminator to help classify slightly non-frontal poses. See Figure 4.3 for examples of this behaviour.

Figure 4.4: Results demonstrating drift in identity when varying a pitch-like parameter in IVI-GAN *without* biometric identity constraint. Each row of images was generated from the same $\mathbf{z}$ vector.

4.2.3.2 Tuning of the biometric constraint

As previously mentioned, modifications to images that significantly affect their general structure, such as changes to pose, can lead to identity shift and shifts in other image properties. Changes to the property of pitch seem to be particularly prone to this issue. (See Figure 4.4.) To counter these problems, IVI-GAN incorporates the explicit, biometric identity constraint described in Section 4.2.2.1. The biometric network [Hasnat *et al.* 2017] was pre-trained on images of resolution 96×96 and so we only activate the additional ID-loss during the final stabilisation period of the training of the progressive GAN. We performed experiments with $\lambda_{ID} = [1.0, 0.1, 0.01, 0.001, 0.0001]$ and finally use $\lambda_{ID} = 0.0001$. Higher values were found to inhibit pose-variation too much.

4.2.4 Preliminary Results

We evaluated IVI-GAN on the CelebA dataset [Liu *et al.* 2015a], and on a dataset of synthetic face images generated using the Basel 3D morphable model (3DMM) [Blanz & Vetter 1999]. In Section 4.2.4.1 we present a selection of results for CelebA, including a qualitative comparison with similar results taken from [Shen *et al.* 2020]. In Section 4.2.4.2 we quantitatively investigate pose changes in images generated from CelebA to give an idea of the form and consistency of multivariate models learnt via weak supervision. Finally, in Section 4.2.4.3, we show additional results for a balanced dataset of synthetic 3DMM images.

Figure 4.5: Left: A continuous eye-wear model learned by IVI-GAN. The style of eye-wear is controlled by the direction of a two-dimensional unit vector. Setting the length of the vector to zero removes the eye-wear (top row); Right: A colourful selection of images demonstrating the capability of IVI-GAN to generate a range of different backgrounds while preserving identity and other image attributes.

4.2.4.1 Taking control of variation in CelebA

We trained IVI-GAN on a 100k image subset of the CelebA dataset. Images were prepared in a similar way to the CelebA-HQ dataset of [Karras *et al.* 2018] but super-resolution was not used. Our network was trained progressively up to a resolution of 128×128 and was conditioned on a selection of the attribute labels available with CelebA. These attributes were complemented with binary labels for lighting, the background, and for pose. Lighting and background labels were found by hand labelling a set of 10k images as containing either ambient or non-ambient lighting, and having either plain or busy/coloured backgrounds. Two simple classifiers were then trained to label the remaining images. Pose labels were found by applying an off-the-shelf pose detector and categorising all images with yaw or pitch greater than three degrees as being non-frontal.

Figures 4.2, 4.5, 4.6 and 4.7 show the effect of varying the multi-dimensional parameter vectors learned for lighting (4 parameters), eye-wear (2 parameters normalised to unit length), background (10 parameters) and pose (2 parameters).

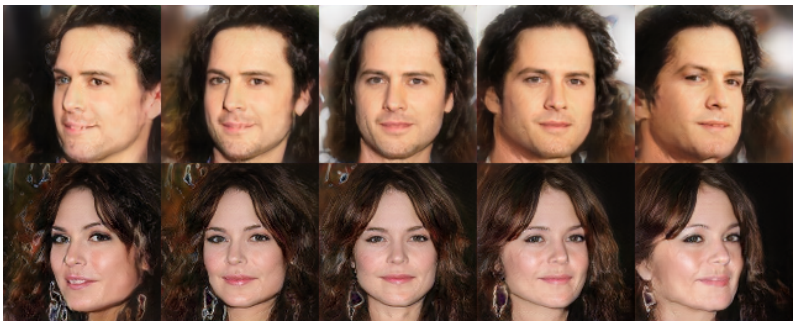


Figure 4.6: Images demonstrating the effect of varying one of the pose parameters, used by IVI-GAN to represent yaw-like variation. The parameter is varied between -3.0 and 3.0. In the middle column $\boldsymbol{\rho}_{pose} = (0, 0)$.

Figure 4.7: Images demonstrating the effect of varying the other pose parameter, used by IVI-GAN to represent pitch-like variation. The parameter is varied between -3.0 and 3.0. In the middle column $\boldsymbol{\rho}_{pose} = (0, 0)$.

Each row of images in Figure 4.5 (left) corresponds to a particular configuration of $\boldsymbol{\rho}_{glasses}$. We use a vector of two parameters normalised to unit length. Only four instances of variation are shown but the style of glasses can be varied continuously by rotating the unit vector, with each style morphing smoothly into the next. Glasses can be removed completely by setting $\boldsymbol{\rho}_{glasses} = (0, 0)$ (top row). We see that modifications to the style of glasses are well disentangled from the other image parameters and from the identity.

In Figure 4.5 (right), each row corresponds to a different random instantiation of $\boldsymbol{\rho}_{background}$. (Setting $\boldsymbol{\rho}_{\boldsymbol{background}} = \mathbf{0}^{\mathbf{10}}$ results in the same set of images as shown in the first row of Figure 4.5 (left).) Again, we see that modifications to the background leave other image properties and the identity largely unaffected. However, we notice that certain features of the background have a tendency to

Figure 4.8: Images taken from [Shen *et al.* 2020].

be present in most images of certain identities. We believe this effect is due to unwanted, spurious correlations in the training dataset.

Although unguided by precise labels, Figures 4.6 and 4.7 show that IVI-GAN has learned to use one pose parameter to represent yaw-like variation and the second to represent pitch-like variation. Other image properties such as lighting, the background and the identity remain consistent. These results (and also those of Figure 4.5 (left)) can be compared with those in Figure 4.8 which have been taken from [Shen *et al.* 2020]. Note that in [Shen *et al.* 2020], images were generated at higher resolution. Here, we have down-sampled them to 128×128 resolution for closer comparison with our own. The identities in Figure 4.8 seem to be reasonably well preserved despite [Shen *et al.* 2020] having taken no explicit steps to achieve this. We suspect that this would not be the case, however, if pitch were to be varied using the same method. With fewer images in the training set exhibiting pitch variations, correlation of large pitches with certain identities can lead to shifts towards those identities. In contrast to [Shen *et al.* 2020], IVI-GAN simultaneously learns a multivariate representation of both yaw *and* pitch. It also explicitly ensures that identity-drift is kept to a minimum via its biometric identity constraint.

4.2.4.2 Weak learning of multivariate models

IVI-GAN is able to learn its own models of variation given only binary labels that indicate the presence or absence of that form of variation. A desirable property of such a model is that the same parameter values result in the same semantic properties irrespective of the identity and other image properties. In Table 4.1 we

Table 4.1: Statistics of poses detected in images of 100 random identities for the given parameter-configurations.

GAN configuration		Mean pose		StdDev
Type	$\boldsymbol{\rho}$	**yaw**	**pitch**	**pose**
IVI-GAN	$(0.0, 1.0)$	$10.2°$	$-1.3°$	$7.5°$
	$(0.0, 2.0)$	$23.8°$	$-6.5°$	$8.8°$
	$(0.0, 3.0)$	$33.7°$	$-12.6°$	$8.8°$
cGAN	$(10.2°, -1.3°)$	$9.3°$	$0.8°$	$4.8°$
	$(23.8°, -6.5°)$	$23.5°$	$-5.2°$	$6.3°$
	$(33.7°, -12.6°)$	$32.5°$	$-11.4°$	$6.9°$

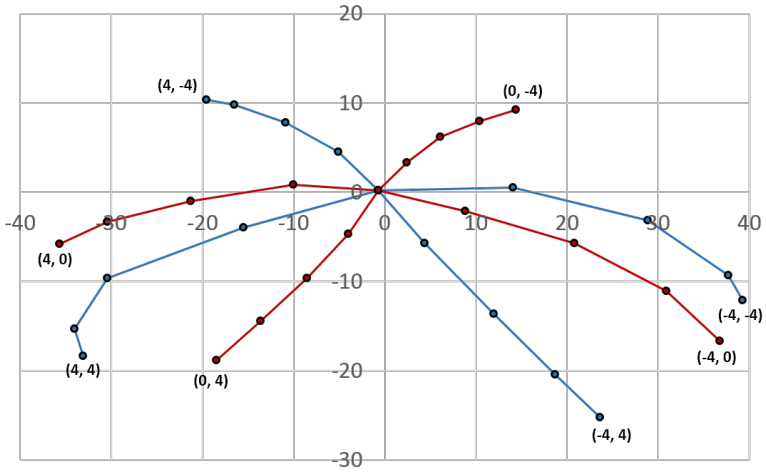


Figure 4.9: Detected poses in images generated by an IVI-GAN. Horizontal and vertical axes indicate detected yaw and pitch for a selection of parameter values (indicated in the plot) averaged over 100 identities.

compare the consistency of poses generated by IVI-GAN (without ID constraint) with those generated by a cGAN conditioned on precise, real-valued pose labels. The only difference between the cGAN and the IVI-GAN is that, for the cGAN, $\boldsymbol{\rho}_{pose}$ are real-valued pose labels instead of random parameters, and are fed to the discriminator in place of the binary labels, β_{pose}. We generated 100 random identities for each of the pose parameter configurations given in Table 4.1 and then used a pose-detector to find the mean generated pose and the standard deviation of angles from that direction. Note that, for a more straightforward comparison, the pose parameter configurations that were fed to the cGAN are the mean poses

Figure 4.10: Examples of images analysed in Table 4.1 generated by an IVI-GAN. Five random identities are shown in frontal poses (top row) and with pose parameters prescribed as $\rho_{pose} = (0.0, 2.0)$ (bottom row).

Figure 4.11: Examples of images analysed in Table 4.1 generated by the cGAN. Five random identities are shown in frontal poses (top row) and with pose parameters prescribed as $\rho_{pose} = (23.8°, -6.5°)$ (bottom row).

detected in the images generated by IVI-GAN. We see that, despite the absence of strong supervision, IVI-GAN is able to generate poses with a consistency close to that of a standard cGAN. (Note that the consistency of the cGAN statistics may be artificially high since the same detector was used to label the training images.) The form of the pose model learned by IVI-GAN (with ID constraint) is depicted in Figure 4.9 and examples of the images analysed in Table 4.1 are given in Figures 4.10 and 4.11. It can be seen that the visual quality of images generated by IVI-GAN is similar to that of the cGAN.

4.2.4.3 Learning from a balanced, synthetic dataset

The quality of generated images at large poses and containing other extreme conditions is limited by the availability of such images in training datasets. As a cleaner

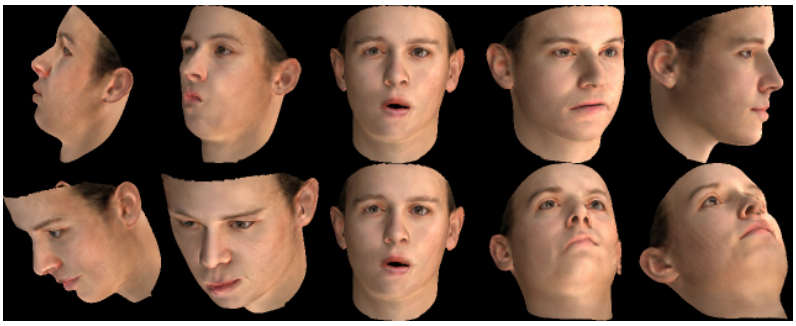


Figure 4.12: Results demonstrating pose-variation in images generated by IVI-GAN trained on a synthetic dataset. The two rows show the effect of varying the two, uniform pose parameters between -3.0 and 3.0. All other parameters were kept the same.

Figure 4.13: Results demonstrating expression and lighting variation in images generated by IVI-GAN trained on a synthetic dataset. The left-hand images show neutral expression ($\boldsymbol{\rho}_{exp} = \mathbf{0}^8$, top) and ambient lighting ($\boldsymbol{\rho}_{lighting} = \mathbf{0}^9$, bottom). The other images show the effects of activating individual, expression and lighting parameters with values of -3.0 or 3.0.

test of IVI, we trained IVI-GAN on synthetic face images generated by a 3D morphable model (3DMM) [Blanz & Vetter 1999] and lit using a spherical harmonic lighting model [Ramamoorthi & Hanrahan 2001]. Identities and expressions were sampled from random Gaussian distributions, and lighting and pose from uniform distributions. Figure 4.12 demonstrates that, given adequate data, IVI-GAN is able to generate high-quality results for the full range of poses, i.e. for yaws in the range $[-90°, 90°]$ (top) and pitches in the range $[-45°, 45°]$ (bottom). We note that, since there is no explicit constraint on expression, the expression we selected for the frontal image is lost during the disruptive pose changes. The desired expression can often be recovered, however, by readjusting expression parameters afterwards, as

has been done in the bottom row of Figure 4.13. Since adjusting the expression is a more subtle image-modification, it does not affect the pose.

With full control over the synthetic, 3DMM dataset, we were able to easily generate ground-truth labels of neutral/non-neutral expression. (Similar labels were not available during our tests on CelebA.) The top row of Figure 4.13 shows the range of expressions learnt by IVI-GAN, whilst the bottom row shows some of the more distinctive lighting modes that were learned. Note that the lighting condition remains consistent as expression is varied and vice-versa.

4.3 A Triplet Loss for GANs

We now introduce our second contribution to the field of disentanglement in GANs. The method builds upon the SD-GAN of [Donahue *et al.* 2018] which is a method designed to disentangle identity from other image characteristics by training the GAN's discriminator to judge whether *pairs* of generated images appear to be from a training data distribution of mated image pairs. To improve upon the method we integrate an "imposter" term into the SD-GAN's loss function to help limit intra-class identity variance. We first introduce the formulation of the SD-GAN in Section 4.3.1 before describing the novel GAN triplet loss in Section 4.3.2. In Section 4.3.3 we present a partial evaluation of the method leaving full evaluation of the ability of the method to disentangle identity until the disentanglement study of Section 4.4 where comparisons are made with other methods.

4.3.1 SD-GAN

The SD-GAN ("Semantic Decomposition" GAN) was proposed in [Donahue *et al.* 2018]. Like IVI-GAN, the training of SD-GAN causes semantic disentanglement of dimensions in the generator's latent space. A variety of formulations of SD-GAN is proposed in [Donahue *et al.* 2018] including an energy-based formulation making use of an auto-encoder as its discriminator. We opted for the simpler, DC-GAN version which we have depicted in Figure 4.14. Similar to the biometric constraint of IVI-GAN, a generator pass is performed

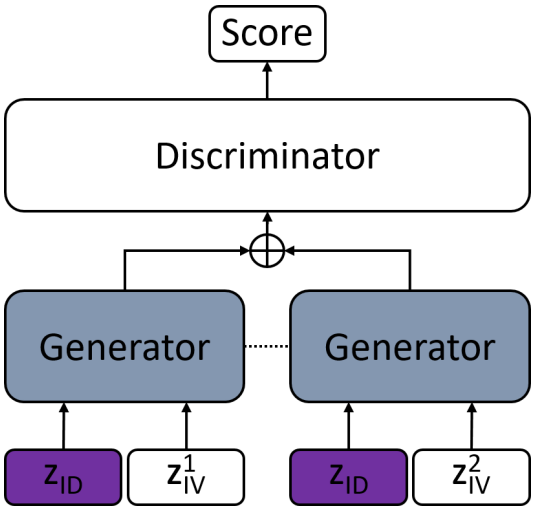


Figure 4.14: Diagram showing the specific SD-GAN architecture used in our evaluation. $\mathbf{z}_{ID}$, $\mathbf{z}^1_{IV}$ and $\mathbf{z}^2_{IV}$ are random vectors; the dotted line indicates shared network weights and the plus indicates channel-wise concatenation of generated images. (IV refers to "Intra-class Variation".)

for each of two latent vectors, $\mathbf{z}_1 = [\mathbf{z}_{ID}, \mathbf{z}^1_{IV}]$ and $\mathbf{z}_2 = [\mathbf{z}_{ID}, \mathbf{z}^2_{IV}]$, that share the common sub-vector $\mathbf{z}_{ID}$. The two generated images are then concatenated along the channel axis before being passed to the discriminator that judges the realism of the image pair as compared to pairs of distinct training images that share a common identity. The discriminator therefore judges based on two criteria: image pairs must appear to be realistic and must also appear to contain the same identity.

As in [Donahue *et al.* 2018], we train the SD-GAN via an adapted version of the Wasserstein loss functions. For convenience, we restate the original Wasserstein loss functions here.

$$\mathcal{L}_{\theta_D} = \mathbb{E}_{\mathbf{x}\sim p_{data}}[D(\mathbf{x};\theta_D)] - \mathbb{E}_{\mathbf{z}\sim p_z}[D(G(\mathbf{z};\theta_G);\theta_D)] \tag{4.9}$$

$$\mathcal{L}_{\theta_G} = \mathbb{E}_{\mathbf{z}\sim p_z}[D(G(\mathbf{z};\theta_G);\theta_D)] \tag{4.10}$$

where $\mathbf{x}$ are images selected at random from the real data distribution, p_{data}; $\mathbf{z}$ is a random vector whose values are selected randomly from some simple distribution, typically a standard Gaussian; and θ_D and θ_G parameterise the discriminator, D,

and the generator, G, respectively. For the SD-GAN, these losses are then adapted to

$$\mathcal{L}_{\theta_D} = D(\mathbf{x}) - D([G(\mathbf{z}_1), G(\mathbf{z}_2)]) \tag{4.11}$$

$$\mathcal{L}_{\theta_G} = D([G(\mathbf{z}_1), G(\mathbf{z}_2)]) \tag{4.12}$$

where we have dropped the expectations and network parameters for simplicity of notation. Here, $\mathbf{x}$ is a pair of images of matching identity. In the following section we describe how these loss functions are modified to form our GAN triplet loss.

4.3.2 Formulation of the GAN triplet loss

The triplet loss was first used for face recognition in [Schroff *et al.* 2015], derived from [Weinberger & Saul 2009], and is typically used to train classifier networks. It has two terms that each ensure desirable characteristics of embeddings in the discriminative feature space: a first term minimises the distance between pairs of images of the same class and ensures intra-class compactness, while a second term maximises the distance from one image of each pair to an image of a different class and ensures good inter-class separation. During training of an SD-GAN, in addition to learning to judge the realism of images, the discriminator performs a biometric function, learning to judge whether pairs of images belong to the same identity class or not. Learning of this function is based only on the consumption of pairs of mated images by the discriminator. This could be thought of as being analogous to the first term of the original triplet loss that ensures good intra-class compactness. Our results show that the biometric function of the SD-GAN's discriminator can also benefit from the integration of a term that acts upon non-mated pairs of images, i.e. pairs of "imposters".

To integrate the imposter term of our GAN triplet loss, the losses in equations

(4.11) and (4.12) are modified to

$$\mathcal{L}_{\theta_D} = D(\mathbf{x}) - \frac{1}{2}[D([G(\mathbf{z}_1), G(\mathbf{z}_2)]) + D(\overline{\mathbf{x}})] + \lambda(D([G(\mathbf{z}_1), G(\mathbf{z}_2)]) - D(\overline{\mathbf{x}}))^2 \quad (4.13)$$

$$\mathcal{L}_{\theta_G} = D([G(\mathbf{z}_1), G(\mathbf{z}_2)]) \quad (4.14)$$

where $\overline{\mathbf{x}}$ is a pair of images containing non-mated identities. To strictly follow the analogy with the original triplet loss, one of these images would be shared with $\mathbf{x}$ and the second would be an imposter. However, in practice we sample $\overline{\mathbf{x}}$ randomly from a set of pre-defined image pairs that demonstrate high matching scores as judged by a biometric network.

The core idea of the GAN triplet loss is encapsulated in the second term of equation (4.13): the discriminator is applied to the non-mated image pairs and the resulting scores averaged with those of the synthetic images to create a new "fake data" term. These two different ways of image pairs appearing to be "fake" (either being synthetic, or real but non-mated) are then contrasted with a real pair of matching images in the first term of the loss. Having added the term $D(\overline{\mathbf{x}})$, $\mathcal{L}_{\theta_D}$ could now be minimised by simply forcing apart the embeddings of $\mathbf{x}$ and $\overline{\mathbf{x}}$ in the feature space of the discriminator and ignoring the synthetic images. To avoid this, we add the third, quadratic term to ensure that the synthetic and non-mated terms retain roughly the same magnitude. This ensures that useful gradients are back-propagated to the generator. In our experiments we found that a weight of $\lambda = 0.001$ worked well and that a value of $\lambda = 0$ resulted in decreased image quality.

4.3.3 Preliminary Results

Official code is not available for SD-GAN and so we implemented our own version, again, building upon the architecture of NVidia's Progressive GAN. Since the discriminator performs an additional biometric function, we double the number of filters in each of its convolutional layers. We found that doing so improved the visual quality of generated images. Other than this modification, the architectures of the generator and discriminator are the same as those found in [Karras *et al.* 2018]. We trained the SD-GAN and our SD-GAN with triplet loss on a proprietary dataset of

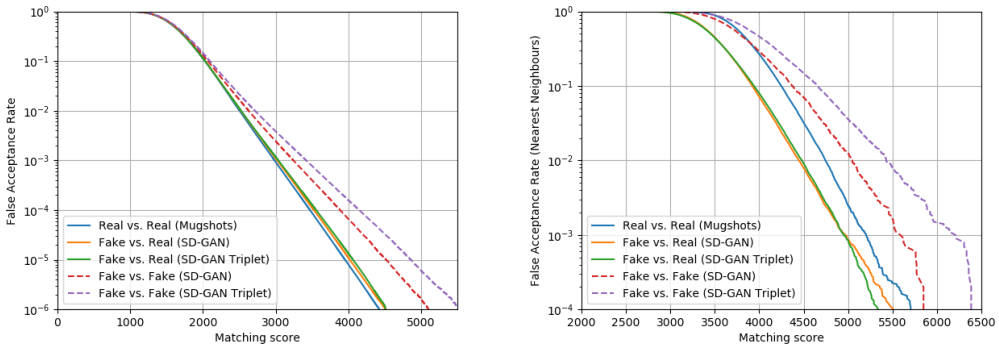


Figure 4.15: Distributions of biometric matching scores for non-mated pairs within and between the real and synthetic datasets. The right-hand plot shows the distributions of nearest neighbour matching scores only.

Figure 4.16: Synthetic samples generated by an SD-GAN trained on a proprietary dataset of mugshots.

96, 286 frontal mugshot images containing 10, 064 identities. To ensure that overfitting to the dataset has not occurred, we performed the analysis of matching scores recommended in Chapter 3. From Figure 4.15 we see that distributions of matching scores within the real, Mugshots dataset are similar to those between the real and synthetic datasets. Matching scores between nearest neighbours were found to be significantly weaker. We can conclude, therefore, that overfitting has not occurred and are able to show images in Figures 4.16 and 4.17 generated by the two GANs despite the dataset being proprietary.

Figure 4.17: Synthetic samples generated by an SD-GAN trained using our GAN triplet loss.

4.4 Results: Measuring the disentanglement of identity in GANs

This section presents a study of the level of disentanglement of identity from other image characteristics in various forms of disentangled GAN, including IVI-GAN and SD-GAN that were presented in the previous two sections. In the literature we find three general methods of disentanglement:

1. Training a conditional GAN (cGAN) and adding a biometric constraint on the identity. In this work we evaluate IVI-GAN. However, there are other works that fall into this category [Tran *et al.* 2019, Sáez Trigueros *et al.* 2021].

2. Training a standard GAN and retrospectively discovering semantically meaningful axes of variation in the latent space [Shen *et al.* 2020, Härkönen *et al.* 2020]. We assess the InterFaceGAN method of [Shen *et al.* 2020]. The method does not explicitly avoid changes to identity upon traversing the GAN's latent space and assumes that a robustly trained GAN will naturally learn to disentangle semantic factors.

3. Training the discriminator of the GAN to act upon pairs of images and to simultaneously penalise poor realism and poor identity consistency. The only work of which we are aware that has previously attempted this is the SD-GAN of [Donahue *et al.* 2018].

To assess the ability of the various GAN-types to disentangle identity from

other image properties, we analyse distributions of matching scores within mated image sets intended to depict the same subject. If disentanglement is successful, we should expect False Rejection Rates (FRR) (at say FAR=10^{-4}) to be comparable to those for sets of real images. We perform this analysis for datasets generated by examples of each type of GAN identified above. These methods are IVI-GAN [Marriott *et al.* 2020a], InterFaceGAN [Shen *et al.* 2020] and SD-GAN [Donahue *et al.* 2018]. We also evaluate the improved disentanglement of the SD-GAN resulting from integration of our GAN triplet loss term.

IVI-GAN and the Progressive GAN used by the InterFaceGAN method were trained using CelebA and CelebA-HQ respectively. The SD-GAN was trained on our proprietary dataset of mugshots since it's discriminator must also learn a reliable biometric function. As shown in the previous chapter, CelebA is not cleanly labelled and also contains fewer identities than are typically used to train state-of-the-art biometric networks such as that used by IVI-GAN. Each GAN was used to generate sets of ten images for 1000 identities as described in the subsections below. Since pose is generally the most disruptive factor to the identity and is disentangled by both IVI-GAN and InterFaceGAN, we ensure a fair comparison with the SD-GANs by selecting pose parameters such that the standard deviation of yaw angles detected in generated images matches that of images generated by SD-GAN.

4.4.1 Generation of the synthetic datasets

4.4.1.1 IVI-GAN

Two datasets of $10,000$ images were generated for IVI-GAN trained on the CelebA dataset. For each dataset, 1000 random $\mathbf{z}$ were selected as well as ten random parameter vectors, ρ, for each $\mathbf{z}$, corresponding to varying pose, expression, lighting and eyewear. All parameters were selected as during training of the IVI-GAN with the exception of the pose parameters which were selected from a standard Gaussian distribution but then, for one of the datasets, were scaled by 0.21 in order for the standard deviation of detected yaw angles to match those generated by SD-GAN trained on the Mugshots dataset. Two sets of samples generated by IVI-GAN with

Figure 4.18: Two sets of synthetic samples generated by IVI-GAN with random pose, expression, lighting and eyewear.

Figure 4.19: Two sets of synthetic samples generated using the InterFaceGAN method [Shen *et al.* 2020]. Pose, expression and eyewear were manipulated by random amounts.

scaled pose can be seen in Figure 4.18.

4.4.1.2 SD-GAN

A dataset of 10,000 images was generated for both the SD-GAN described in Section 4.3.1, and the SD-GAN with triplet loss described in Section 4.3.2. Each GAN was trained on the proprietary dataset of mugshot images. For each dataset 1000 random $\mathbf{z}_{ID}$ were selected as well as ten random $\mathbf{z}_{IV}$ for each $\mathbf{z}_{ID}$. Samples of images generated by the SD-GAN can be seen in Figures 4.16 and 4.17.

4.4.1.3 InterFaceGAN

InterFaceGAN [Shen *et al.* 2020] is not a GAN architecture but rather a method of controlling properties in images generated by existing, pre-trained GANs. By observing generated images $G(\mathbf{z};\theta)$ and associating binary attribute labels with

latent vectors, $\mathbf{z}$, e.g. "smiling/not smiling", classifiers can be trained to find the hyper-planes in the latent space separating these image characteristics. An image characteristic can then be controlled by traversing the latent space along the axis perpendicular to the associated hyper-plane. Image sets were generated using the publicly available version of the Progressive GAN trained on CelebA-HQ. Again, two datasets of $10,000$ were generated, each with scaled pose. A thousand random $\mathbf{z}$ were first selected and then ten different images generated for each by traversing the latent space by random distances in the pose, smile and eyewear directions. Final distances from each of the three hyper-planes were selected from standard Gaussian distributions. Pose distances were scaled by 0.9 and 0.15 to give the same standard deviation of yaw angles as for the unscaled IVI-GAN and for the SD-GAN. In [Shen *et al.* 2020] correlations were observed between eyewear and gender. To help avoid these unwanted changes to the identity, the eyewear boundary conditioned on gender, provided by the authors, was used. Two sets of samples generated by the Progressive GAN using the InterFaceGAN method (with pose scaled by 0.15) can be seen in Figure 4.19.

4.4.2 Comparison of matching-score distributions for disentangled, synthetic datasets

Figure 4.20 shows distributions of matching scores for all mated image pairings within the various datasets. Note that in Figure 4.20 (left), all curves are not strictly comparable since the variance in yaw angle generated by both IVI-GAN and InterFaceGAN was significantly larger than for the SD-GAN datasets. This has been indicated by plotting dashed curves. We include this figure to demonstrate the poor matching scores found for mated pairs in the InterFaceGAN dataset. In Figure 4.20 (right), in which the variance in yaw has been matched across synthetic datasets, the distribution of matching scores for InterFaceGAN is improved. However, perhaps unsurprisingly, InterFaceGAN remains to be the least successful method at disentangling identity. Ideally, score distributions should be similar to those calculated for the dataset of real images (blue curve). We note, however,

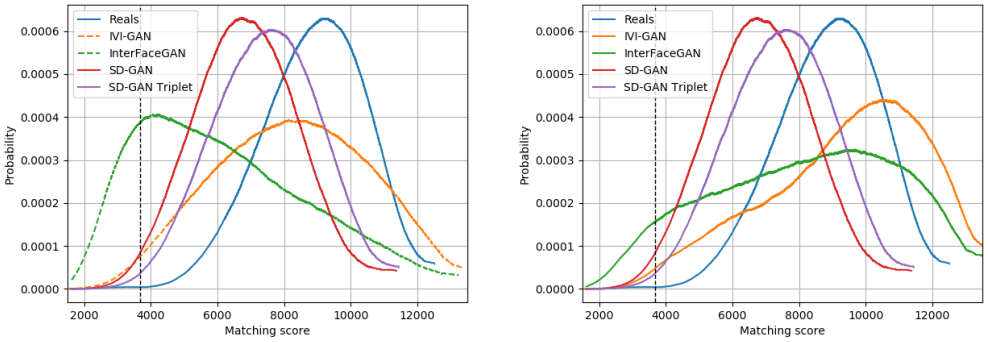


Figure 4.20: The probability distributions of biometric matching scores for all mated pairs of images within the dataset indicated in the legend. Left: the pose-parameters of InterFaceGAN were scaled down such that the standard deviation of yaw detected in images matches that of IVI-GAN (10.1°); Right: the pose-parameters of both InterFaceGAN and IVI-GAN were scaled down such that the standard deviation of yaw detected in images matches those of the SD-GAN (3.4°).

Dataset	FRR@3677.5	FRR@FAR=10^{-4}	StdDev(Yaw)	LPIPS-Intra	LPIPS-Inter
Reals (mugshots)	2.97×10^{-3}	2.97×10^{-3}	5.5°	0.402	0.506
IVI-GAN (CelebA)	4.14×10^{-2}	4.35×10^{-1}	10.1°	0.334	0.558
InterFacePGAN (CelebA-HQ)	2.75×10^{-1}	3.78×10^{-1}	10.1°	0.315	0.583
IVI-GAN (CelebA)	2.45×10^{-2}	2.95×10^{-1}	3.4°	0.225	0.527
InterFaceGAN (CelebA-HQ)	1.05×10^{-1}	1.71×10^{-1}	3.4°	0.176	0.555
SD-GAN (mugshots)	4.37×10^{-2}	8.94×10^{-2}	3.4°	0.320	0.419
SD-Triplet (mugshots)	1.86×10^{-2}	6.69×10^{-2}	3.4°	0.306	0.428

Table 4.2: A selection of statistics for mated image sets from various datasets. Also reported are mean, inter-class LPIPS distances. The grey rows show statistics for datasets with larger variance in pose.

that significant portions of each synthetic distribution lie below the threshold of 3677.5 which corresponds to FAR=10^{-4} for the dataset of real images. Table 4.2 reports FRRs at this threshold, and also at points corresponding to FAR=10^{-4} based on the matching scores between non-mated pairs for each synthetic dataset. To give an idea of the remaining discrepancies in non-identity variation, Table 4.2 also reports average perceptual distances between mated pairs ("LPIPS-Intra") and non-mated pairs ("LPIPS-Inter") as measured by the LPIPS perceptual similarity metric [Zhang *et al.* 2018]. We used version 0.1 of LPIPS with a VGGNet base and additional linear calibration layer. (Note that these distances are not insensitive to changes in identity.)

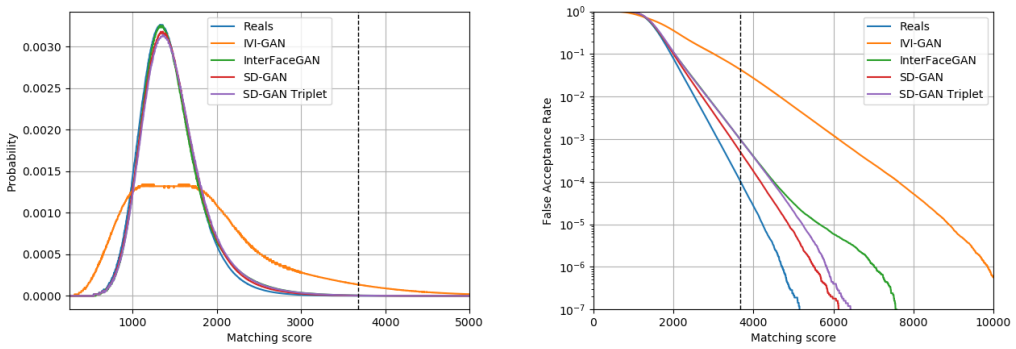


Figure 4.21: The probability distributions of biometric matching scores for all pairs of images *not* sharing the same identity within the real or synthetic dataset indicated in the legend.

As previously mentioned, in absolute terms (based on FRR@3677.5), InterFaceGAN proved to be the least effective method at maintaining identity with an FRR of 10.5% despite demonstrating lower intra-class LPIPS distances than both IVI-GAN and SD-GAN. It appears that, without applying an explicit biometric constraint, identity is not well disentangled from other properties in the latent space of the Progressive GAN. SD-GAN with triplet loss was found to be the most effective method at preserving identity, followed by IVI-GAN. Based on this metric (FRR@3677.5), IVI-GAN outperforms the standard SD-GAN, even in the case that we do not limit the range of generated poses. This is the case despite IVI-GAN demonstrating larger intra-class LPIPS distances. (See the grey rows of Table 4.2.)

While the identity constraints of SD-GAN Triplet and IVI-GAN are the most effective, using them appears to come at a cost. Figure 4.21 shows matching score distributions for non-mated pairs. We see that IVI-GAN suffers from significant mode-collapse in the biometric feature space and demonstrates strong matching scores for a high proportion of non-mated pairs. We also notice that integration of the triplet loss into SD-GAN increases mode-collapse in comparison to the standard SD-GAN. Interestingly, this collapse appears to be confined to the biometric feature space and does not strongly affect the inter-class LPIPS distances. SD-GAN Triplet, for example, demonstrates larger inter-class LPIPS distances than for the standard SD-GAN despite mode-collapse having been measured for identity. The

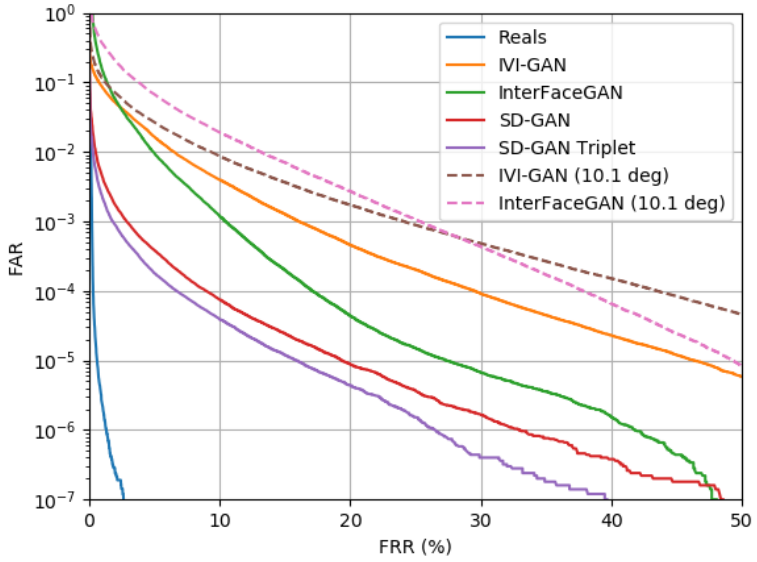


Figure 4.22: ROC curves for each of the identity-disentangled datasets.

overall level of disentanglement is probably best judged by statistics such as the FRR@FAR=10^{-4}, which takes account of both intra-class compactness and inter-class separation. By this metric we see that integration of the GAN triplet loss was beneficial, decreasing FRR@FAR=10^{-4} from 8.94×10^{-2} to 6.69×10^{-2}. In fact, when looking at the ROC curves in Figure 4.22 we see that the triplet loss improves FRR at all FAR thresholds. We also see that InterFaceGAN outperforms IVI-GAN at most thresholds (solid curves). However, this is not the case if pose is allowed to vary more freely (dashed curves). Note that none of the synthetic datasets comes close to the level of disentanglement seen in the dataset of real images.

4.5 Conclusion

Through implementing Intra-class Variation Isolation, we showed that it is possible to adapt a conditional GAN in order to gain continuous, disentangled control over image attributes without the need for extensive labelling. Only simple binary labels, indicating whether an attribute is present in one form or another, are required. IVI then allows a GAN to discover its own, multivariate way of modelling the variation

present within that attribute category in an unsupervised fashion. To the best of our knowledge, IVI-GAN is the first network to achieve this separation in a weakly supervised manner.

We also assessed the performance of methods designed to disentangle identity from other image properties. We evaluated InterFaceGAN, IVI-GAN and SD-GAN, and showed that our novel GAN triplet loss can be used to improve the disentanglement of identity. None of the algorithms, however, is able to disentangle identity to a satisfactory degree. Even the lowest value of FRR@FAR=10^{-4}, found for our SD-GAN with triplet loss, is more than an order of magnitude larger than that found for real data. Supervised data-augmentation involving the generation of sets of identity-disentangled images is therefore unlikely to be fruitful when using methods similar to those evaluated here, and more work needs to be done to in order to gain full control over image variation for applications such as generation of anonymised test sets and identity-robust image-editing.

Chapter 5

A 3D GAN for identity-preserving disentanglement of pose

5.1 Introduction

State-of-the-art facial recognition (FR) algorithms are trained using millions of images. With the internet as a resource, face-images are relatively easy to come by. However, the distribution of semantics throughout these images is usually highly unbalanced. For example, the majority of available photographs are frontal portraits of smiling subjects, with images containing large poses being relatively scarce. Robustness to pose is currently thought to be the largest challenge for face recognition. Some researchers have attempted to avoid the problem by first frontalising probe images [Zhu *et al.* 2015, Hassner *et al.* 2015, Shen *et al.* 2018], whilst others have attempted to learn additional robustness to pose by synthetically augmenting training datasets [Masi *et al.* 2016, Crispell *et al.* 2017, Zhao *et al.* 2017, Deng *et al.* 2018]. We advocate this second approach since it does not require additional resources at test time.

Synthetic augmentation of poses in training data has typically been achieved by fitting some 3D face model to input images, extracting textures, and then re-projecting those textures at modified poses [Crispell *et al.* 2017, Zhao *et al.* 2017]. With recent advances in the development of Generative Adversarial Networks (GANs), however, a viable alternative has emerged. In Chapter 3, GANs were shown to be capable of generating realistic images of new identities and so restrict-

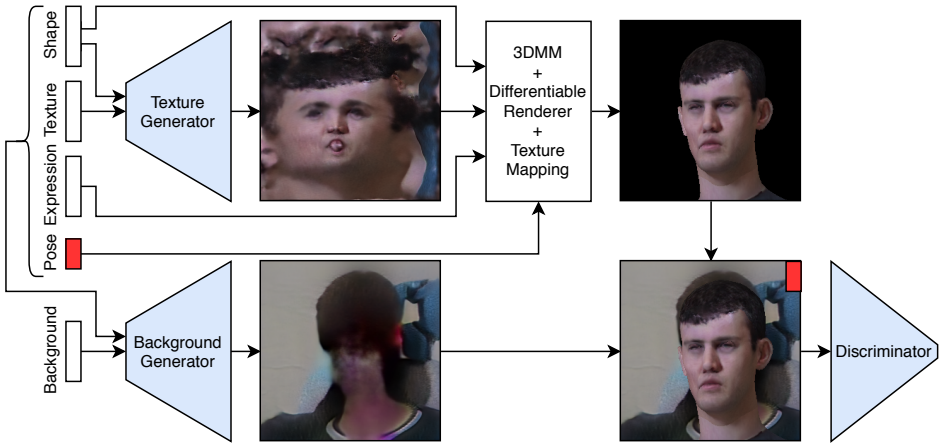


Figure 5.1: The 3D GAN's generator consists of two CNNs that generate facial texture and background. Facial texture is rendered into the background using some random sample of shape from the 3D model's distribution. The random pose and expression vectors are used only for rendering, not for generation of texture, and so remain disentangled from the identity. All parameters are passed to the background generator to allow harmonisation of the background conditions with the rendered subject. Note that all vectors are randomly sampled and that no direct comparison with training images is performed.

ing data-augmentation to existing identities is not necessary. We then showed in Chapter 4 that, even with explicit constraints on identity, 2D GANs are not capable of adequately disentangling identity from other characteristics. To remedy this situation, we incorporate a 3D morphable model (3DMM) [Li *et al.* 2017b] into a GAN so that images of new, synthetic identities can be generated, and the pose modified without identity being compromised. As pointed out in Chapter 2, 3D models were used to augment FR datasets with synthetic identities in [Kortylewski *et al.* 2018] and [Gecer *et al.* 2020]. The method presented here makes the contribution of allowing synthetic identities to be generated in 3D using only in-the-wild images. No specially captured scans of facial texture are required.

The rest of the chapter is organised as follows: in Section 5.2 we discuss work related to the use of 3D face models in image-generation and data-augmentation; in Section 5.3 we introduce our method; in Section 5.4 we present results justifying the formulation of our 3D GAN as well as an evaluation of data-augmentation using the synthesised data; and in Section 5.5 we conclude.

5.2 Related Work

5.2.1 Generative 3D networks

Prior to the recent explosion in the development of GAN-related methods, the best way of generating synthetic face images was to use a 3D morphable model (3DMM). The original 3DMM of [Blanz & Vetter 1999] was learned from a relatively small set of approximately 200 3D shape and texture scans. More recently, several efforts have been made to build more representative 3D models. For example, the Large Scale Face Model (LSFM) [Booth *et al.* 2016] was constructed using 9663 facial scans, and the FLAME model (Faces Learned with an Articulated Model and Expressions) [Li *et al.* 2017b] was learned from 3800 scans and has separate male and female shape models. While the linear spaces of these models are known to capture most of the variation in the training datasets, generated faces still appear to be smooth with textures lacking in high frequency detail. This is thought to be a limitation of using a linear texture model.

In [Gecer *et al.* 2020] and [Gecer *et al.* 2019] the linear texture model of the LSFM is replaced by the nonlinear, CNN generator of a GAN trained to approximate the distribution of their dataset of high-quality texture scans. The quality of generated textures is outstanding. However, the dataset of scans is not available for general use. The difficulty of obtaining high-quality texture datasets motivates the development of methods such as our own, which aims to learn textures from natural (non-scanned) images. The method of [Tran & Liu 2018] has a similar aim and attempts to train an auto-encoder to reconstruct in-the-wild training images. Their disentangled auto-encoding pipeline involves generation of intermediate texture estimations for input images which are then rendered back into the reconstructed images. Since the method requires an input image to be encoded, new identities cannot easily be generated. The method proposed in this chapter is a GAN rather than an auto-encoder, and so can generate new, synthetic identities. The quality of our generated textures is also not limited by reconstruction losses, which tend to destroy high-frequency detail.

5.2.2 Large-pose 3D data-augmentation

As discussed in detail in Chapter 2, there are a number of works that have attempted data-augmentation for FR using techniques involving 3D models. Earlier methods extracted textures from images onto a 3D model's surface for manipulation of pose and sometimes illumination or expression [Masi *et al.* 2016, Crispell *et al.* 2017, Lv *et al.* 2017, Peng *et al.* 2017]. Due to self-occlusion in images and therefore holes in the textures, various in-filling techniques were employed. In [Zhao *et al.* 2017] this problem is tackled by refining the projected texture in image-space using a GAN. A similar idea is used in [Gecer *et al.* 2018] but it is synthetic 3DMM images that are refined by performing unsupervised translation to the real domain. These final refinement phases require identity-preserving losses, which is less than ideal for the purpose of data-augmentation for FR.

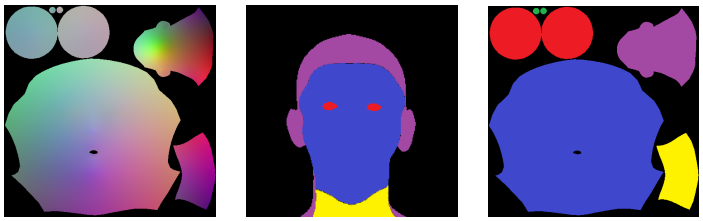
A preferable method is to produce a complete texture in the texture reference space to ensure that the identity remains consistent when projected to different poses. In [Deng *et al.* 2018], a texture-completion network is trained using a set of carefully prepared ground-truth textures. In [Kortylewski *et al.* 2018] and [Gecer *et al.* 2020] the problem of texture completion is avoided entirely by generating textures for synthetic identities. [Kortylewski *et al.* 2018] uses a linear texture model whereas [Gecer *et al.* 2020] trains a nonlinear model. Each of these methods makes use of datasets of scanned textures. The method proposed here also makes use of synthetic identities in order to avoid the problem of texture completion and reconstruction of existing identities. The method, however, does not require carefully prepared/captured ground-truth textures and, instead, learns textures directly from in-the-wild images.

5.3 The 3D GAN

Generative Adversarial Networks typically consist of a convolutional generator and discriminator that are trained alternately in a mini-max game: the discriminator is trained to distinguish generated images from those of a training set of real images, and the generator is trained to minimise the success of the discriminator. Although

generated images appear to represent real-world, 3D subjects, they are in fact nothing more than collections of 2D features learned by the 2D convolutional filters of the generator. For this reason, upon linearly traversing the latent space of a GAN's generator, one tends to see "lazy", 2D transformations between forms rather than transformations that are semantically meaningful in 3D space. For example, even if a direction in the latent space is identified that influences the pose of a face in a generated image, the 3D form of the face is unlikely to be maintained. Indeed, the generator may not even be capable of generating the same face at a different pose. In order to ensure that 3D form is maintained in synthesised images upon manipulation of pose, we enhance the generator by integrating a 3D morphable model (3DMM).

Typically a GAN's input is a random vector. The inputs to our 3D GAN are random texture and background vectors but also random 3DMM shape, expression and pose parameters. A differentiable renderer is then used to render random head-shapes into a generated "background image" with the facial texture being provided by the texture generator. No matter what the shape or pose of the random model instance, the rendered image must appear realistic to the discriminator. To achieve this, the texture generator learns to generate realistic textures with features that correctly correspond with the model shape. Figure 5.1 depicts the architecture of our 3D GAN. The lower half of the diagram depicts a standard conditional GAN in which some image is generated from random parameters and pose information, and is then fed to the discriminator. (In our implementation, pose information is repeated spatially and concatenated as additional channels of the image). The top half of the diagram depicts the integration of a 3DMM where a learned texture is rendered into this image via a differentiable renderer. With the main subject of the image being provided by the rendered texture, the background generator learns to generate only the background and features not modelled by the 3DMM, for example, the edges of glasses and hair. Since the texture generator is not conditioned on pose information, nor expression parameters, these aspects of the image can be manipulated without affecting the texture of the 3D model, as shown in Figures 5.3, 5.4, 5.7 and 5.8.



(a) (b) (c)

Figure 5.2: a) The FLAME 3DMM's texture map where RGB represents the corresponding 3D point on the mean model shape; b) a rendering of the texture shown in (c).

5.3.1 Implementation

Our full generator is a function of five sets of random input parameters and two sets of trained parameters:

$$\mathbf{x} = G([\mathbf{z}_T, \mathbf{z}_B, \beta, \psi, \phi]; [\theta_T, \theta_B]) \tag{5.1}$$

$$= (\mathbf{1} - \mathbf{K}) \circ G_B(\mathbf{z}_B, \mathbf{z}_T, \beta, \psi, \phi; \theta_B) + \mathbf{K} \circ \mathcal{M}(G_T(\mathbf{z}_T, \beta; \theta_T), \mathbf{y}) \tag{5.2}$$

where $\mathbf{x}$ is a generated image; G_B and G_T are the background and texture generators; $\mathbf{z}_T \in \mathcal{N}^{N_T}$ and $\mathbf{z}_B \in \mathcal{N}^{N_B}$ are vectors of random texture and background parameters of length N_T and N_B respectively, selected from standard normal distributions; $\beta \in \mathcal{N}^{N_s}$ and $\psi \in \mathcal{N}^{N_e}$ are vectors of shape and expression parameters that control the form of the 3DMM, again selected from standard normal distributions; ϕ is pose information, typically values of yaw and pitch selected at random from the labels of the training set of images; and θ_T and θ_B parametrise the texture and background generator networks. The background image and rendered texture are combined using a binary mask, $\mathbf{K}$, generated by the renderer. (Note that the masking by $\mathbf{K}$ is not shown in Figure 5.1.) $\mathbf{1}$ is a vector of ones of the same shape as the image and $\mathbf{a} \circ \mathbf{b}$ represents the element-wise product of vectors $\mathbf{a}$ and $\mathbf{b}$. $\mathcal{M}$ is an inverse texture-mapping function that maps interpolations from the generated texture map to appropriate locations in image space based on a rendering of texture coordinates in image-space, $\mathbf{y}$. Inverse texture mapping effectively allows the generated texture to be pasted onto the model surface rather than having only

single colours at each vertex and interpolating across facets. To make the most of texture-mapping, our texture generator operates at twice the resolution of the background generator. Rendering of $\mathbf{y}$ (and simultaneously, $\mathbf{K}$) is performed by the differentiable rendering function, R:

$$\mathbf{y}, \mathbf{K} = R(\mathbf{S}, \phi; \tau, \gamma) \tag{5.3}$$

where $\mathbf{S} \in \mathbb{R}^{N_v \times 3}$ is a vector of shape vertices for some random instance of the 3DMM; ϕ is pose information; $\tau \in \mathbb{Z}^{N_\tau \times 3}$ is the 3DMM's triangle list of N_τ vertex indices; $\gamma \in \mathbb{R}^{3N_\tau \times 2}$ is the vector of texture coordinates where each of the N_τ triangles has its own set of three 2D texture vertices. The rendering function, R, is implemented by DIRT (Differentiable Renderer for Tensorflow) [Henderson & Ferrari 2020] and we use the FLAME (Faces Learned with an Articulated Model and Expressions) [Li *et al.* 2017b] 3DMM. FLAME is an articulated model with joints controlling the head position relative to the neck, the gaze direction, and the jaw. During training of our 3D GAN we fix the joint parameters in their default positions such that the shape is given by the following, simplified equation

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{n=1}^{N_s} b_n \mathbf{s}_n + \sum_{n=1}^{N_e} c_n \mathbf{e}_n \tag{5.4}$$

where $\bar{\mathbf{S}}$ is the mean model shape; $\mathcal{S} = [\mathbf{s}_1, ..., \mathbf{s}_{N_s}]$ are the principal components of shape; $\varepsilon = [\mathbf{e}_1, ..., \mathbf{e}_{N_e}]$ are the principal components of expression; and $[b_1, ..., b_{N_s}]$ and $[c_1, ..., c_{N_e}]$ are the individual elements of the previously defined shape and expression vectors, β and ψ, that are also fed to the generator networks in equation (5.2). For the FLAME model, $N_s = 200$, $N_e = 200$, and $N_v = 5023$. We also set $N_T = N_B = 200$.

The architectures of G_T and G_B are based on that of the Progressive GAN [Karras *et al.* 2018]. However, to simplify implementation and speed up training, no progressive growing was used. We believe that use of a 3D model may act to stabilise training since it provides prior form that need not be learned from

scratch. The architecture was augmented with bilinear interpolation on upscaling (rather than nearest-neighbour upscaling), which helps to avoid checker-board artefacts, and with static Gaussian noise added to each feature map, as used in [Karras *et al.* 2019], which helps to prevent wave-like artefacts from forming. (See Figure 5.5 for examples.)

5.3.2 Training

Despite the more elaborate architecture of the generator, the 3D GAN can be trained like any other GAN. We choose to optimise a Wasserstein loss [Arjovsky *et al.* 2017] by alternately minimising equations (5.5) and (5.6). The values of all input vectors (with the exception of the conditional pose parameters) are selected from a standard Gaussian distribution. For simplicity of notation we agglomerate them into a single vector $\nu = [\mathbf{z}_T, \mathbf{z}_B, \beta, \psi]$.

$$\mathcal{L}_{\theta_D} = \mathbb{E}_{(\mathbf{x}_r,\phi)\sim p_{data}}[D(\mathbf{x}_r, \phi; \theta_D)] - \mathbb{E}_{\nu\sim\mathcal{N},\phi\sim p_{data}}[D(G(\nu, \phi; \theta_G), \phi; \theta_D)] + Reg. \tag{5.5}$$

$$\mathcal{L}_{\theta_G} = \mathbb{E}_{\nu\sim\mathcal{N},\phi\sim p_{data}}[D(G(\nu, \phi; \theta_G), \phi; \theta_D)] \tag{5.6}$$

where $(\mathbf{x}_r, \phi)$ is a real image and associated pose labels selected at random from the distribution of training data, p_{data}; $\theta_G = [\theta_T, \theta_B]$; and *Reg.* indicates the addition of a gradient penalty [Gulrajani *et al.* 2017] that acts to regularise the discriminator such that it approximately obeys the required k-Lipschitz condition [Arjovsky *et al.* 2017]. Note that, during training, the shape and expression parameters passed to the generator are random. There is never any direct reconstruction of training images via fitting of the 3D model. The only constraint on textures is that they must appear realistic (as judged by the discriminator) when projected at any angle and with any expression. Our motivation for training our generator as a GAN and avoiding reconstruction is to generate new identities and to avoid smoothed textures caused by reconstruction errors.

5.3.3 Limitations

The 3D GAN method has certain limitations, the most fundamental possibly being the fact that hair and glasses are not included in the 3D shape model. This can lead to projections of these features onto the surface of the model that do not necessarily look realistic when viewed from certain angles. The inclusion of such features in the shape model would be difficult at best. Instead, it may be better to detect and remove images containing unmodelled features from the training dataset and to seek another method for augmentation with glasses and occlusion by overhanging hair.

As currently formulated, the 3D GAN learns lighting effects and shadows as part of the texture. Although this helps generated images appear to be realistic, it is not ideal for our goal of improving FR since specific lighting conditions become part of the synthetic identities. Since we have the 3D shape for each generated image, a lighting model could be used to produce shading maps of randomised lighting conditions during training. Ideally, the random lighting conditions should follow the distribution of lighting in the training set. In this way the texture generator might avoid inclusion of the modelled lighting effects in the texture.

We also make the assumption that the distributions of shape and expression in the training dataset match the natural distributions of the 3DMM. This is not necessarily the case and improvements could be possible by first fitting the model to the dataset. N.B. we suggest this only for estimating the distributions, not for reconstructing images since fitting errors would be large in individual cases. We also assume that the distributions of feature points (used for alignment) and poses are known. For our in-the-wild experiments, these were detected automatically. We believe the mislabelling of poses to be one of the reasons for the drop in quality between our experiments using Multi-PIE and using either CelebA or FFHQ. (See the results in the following section.)

Finally, the texture map provided with the FLAME 3DMM (see Figure 5.2a) is spatially discontinuous. Since CNNs function by exploiting spatial coherence, these discontinuities in the texture-space lead to discontinuity artefacts in the rendered

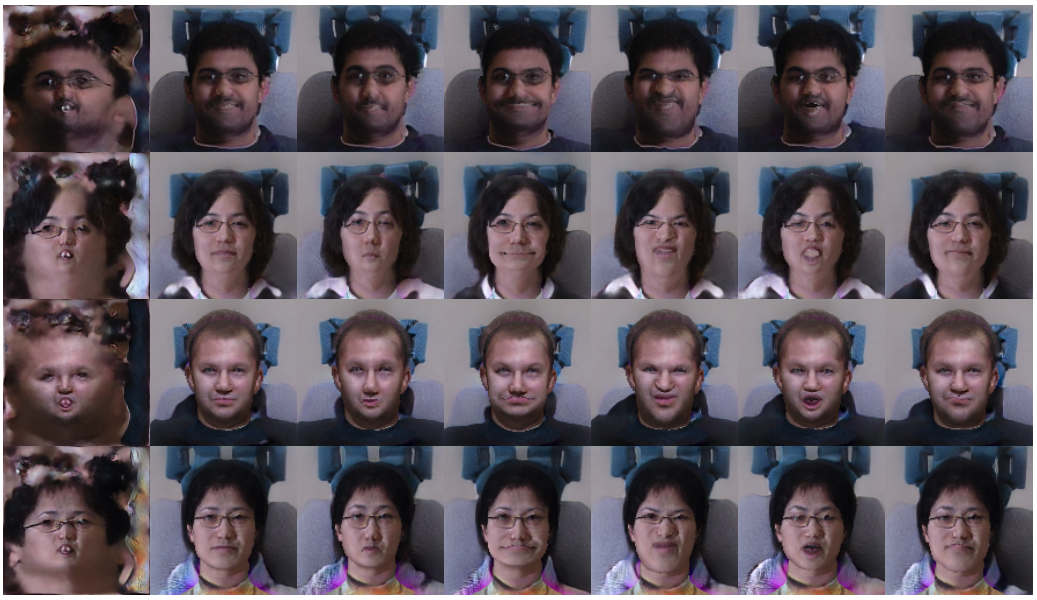


Figure 5.3: 3D GAN textures and renderings for various expressions trained using Multi-PIE.

images. This can be seen, for example, in Figure 5.3 where the facial texture meets the texture of the back of the head. These artefacts could be avoided by using an alternative, spatially continuous texture mapping.

5.4 Results

5.4.1 Controlled evaluation of the 3D GAN

During development of the 3D GAN, tests were conducted by training on the controlled, Multi-PIE dataset [Gross *et al.* 2010]. Doing so avoided potential problems that might have been caused by the incorrect detection of poses, which are required to condition the GAN. During these tests, the pitch angle was not varied and so we excluded Multi-PIE's CCTV-like camera angles (cameras 8 and 19). The first column of Figure 5.3 shows examples of random textures learned by the 3D GAN. To demonstrate the level of correspondence with the shape model, we render each texture for six different expressions. We see that features are well aligned and that expressions can be manipulated realistically. This is thanks to the requirement that

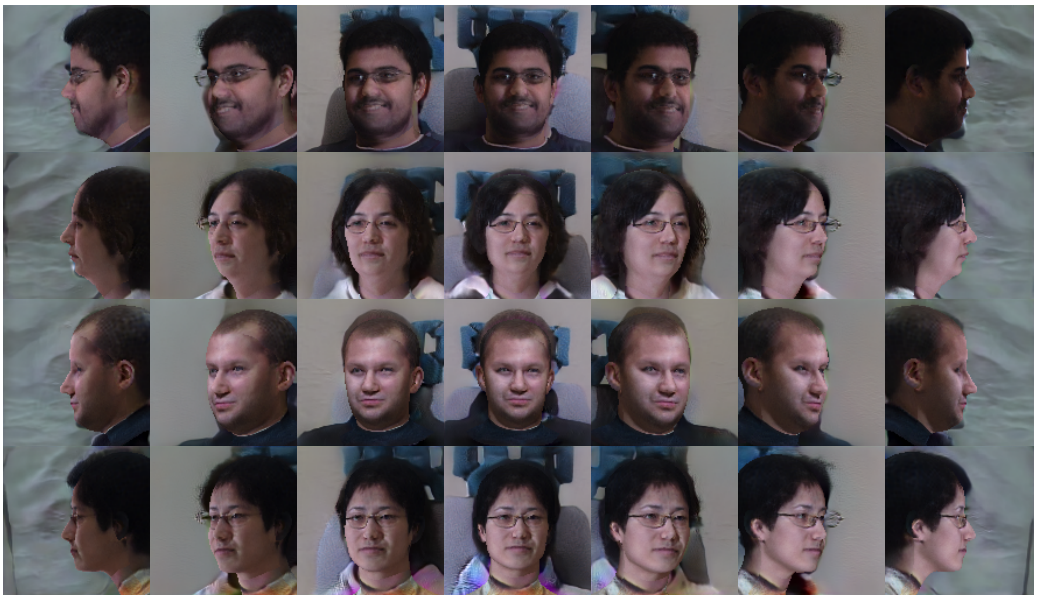
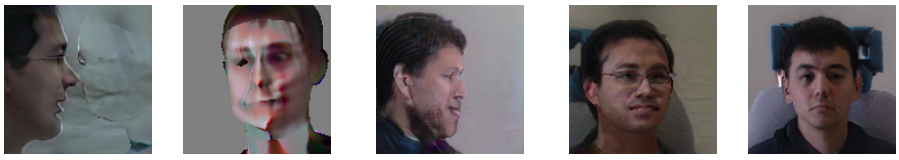


Figure 5.4: 3D GAN renderings at a range of yaw angles trained using Multi-PIE. (The model instances correspond to the *Neutral* column of Figure 5.3.

the texture look realistic for renderings of all poses and expressions. The texture is not dependent on the expression parameters and so the identity is implicitly maintained, at least to the limit of disentanglement present in the 3DMM. Figure 5.4 shows renderings of the same textures with a neutral expression at a selection of yaw angles in the range $[-90°, 90°]$. We see that the model heads are pleasingly integrated with the background with additional, unmodelled details such as hair and the edges of glasses being generated. In some cases, however, this is problematic. For example, in the final column, something resembling a protruding chin has been generated in the background for both of the male subjects. Note, however, that the background is only needed for training and that facial textures can be rendered onto arbitrary backgrounds.

Figure 5.5 shows a set of images that characterise the effects of disabling various aspects of our 3D GAN. Figure 5.5a shows that disabling the pose-conditioning can lead to degenerate solutions where the generators conspire to generate faces as part of the background and to camouflage the model. In the given example, pose-conditioning would have caused the discriminator to expect a leftward-facing subject and to therefore penalise such an image. Attempting to avoid this problem



a) No cond b) No bckd c) No bilinear d) No noise e) Final

Figure 5.5: Results characterising the effects of disabling various features of the final implementation of our 3D GAN.

by switching off the background generator causes a different problem. We can see this in Figure 5.5b where the texture generator now produces a mixture of face-like and background-like features in order to satisfy the discriminator. Figure 5.5c has the background and pose-conditioning enabled. It demonstrates, however, obvious checker-board artefacts in the texture. We found that this problem was caused by the nearest-neighbour up-sampling of feature-maps upon resolution doubling within the generator. Following the work of [Karras *et al.* 2019] we switched to bilinear up-sampling. Whilst this prevented the checkerboard artefacts, it led to wave-like artefacts being generated. These can be seen in Figure 5.5d. Finally, we added static, channel-wise Gaussian noise into the generator, similar to that used in [Karras *et al.* 2019]. See Figure 5.5e. The noise acts to provide high-frequency, stochastic features by default so that the generator need not attempt to derive these details from the random input vectors. Images generated by our full model are of comparable quality to those of [Tran & Liu 2018], which is perhaps the closest work to our own since it attempts to learn a non-linear texture model from in-the-wild images. Our method also has the benefit, however, of being able to 1) easily generate new identities, 2) generate full facial images, including the back of the head and the background, and 3) does not require the 3DMM to be fit to training images, thus avoiding reconstruction errors.

5.4.2 Data-augmentation in the wild

In the previous section we saw that it is possible to learn textures of good quality from a controlled dataset of images containing a wide range of pose. It is unlikely, however, that the synthetic 3D GAN data will be more informative than

Dataset	Num IDs	Num images
MS1M-V3	93.4k	5.2M
NetScrape (in-house)	26.8k	3.5M
CASIA Webface	10.6k	0.5M
CelebA	10.2k	0.2M
Flickr-Faces-HQ	N/A	0.07M

Table 5.1: Training dataset comparison.

the original, high-quality dataset. Although the 3D GAN is able to generate new identities and allows full control over the pose, the data also inevitably suffers from problems such as mode-collapse and from limited realism. In Chapter 2 we identified that many data-augmentation and data-normalisation methods in the literature make use of controlled datasets but do not perform fair comparisons by also including those data in baseline experiments. In this section we wish to demonstrate improvement to FR by making better use of noisy, in-the-wild datasets. We present experiments for various FR algorithms trained using one of three training datasets: either our in-house "NetScrape" datset, CASIA Webface [Yi *et al.* 2014], or MS1M-V3 [Deng *et al.* 2019b]. The datasets are augmented using synthetic data generated by the 3D GAN trained using either CelebA [Liu *et al.* 2015b] or Flickr-Faces-HQ (FFHQ) [Karras *et al.* 2019]. Since CelebA is a dataset of potential benefit to FR, it was also included in additional baseline experiments. Evaluation was performed for two challenging, large-pose datasets, Celebrities in Frontal-Profile in the Wild (CFP) [Sengupta *et al.* 2016] and Cross-Pose LFW (CPLFW) [Zheng & Deng 2018], as well as their frontal-frontal counterparts. Benefit from use of 3D GAN data arises from a combination of the balanced distribution of poses and expressions, the use of a 3D lighting model, the presence of additional synthetic identities, and the GAN's ability to clean noisy datasets.

5.4.2.1 Training datasets

Our baseline FR experiments are trained on either CASIA Webface, MS1M-V3 or our in-house dataset of 3.5 million images scraped from the internet, labelled as "NetScrape" in Figure 5.6 and Tables 5.2 and 5.3. (CelebA is also used for baseline

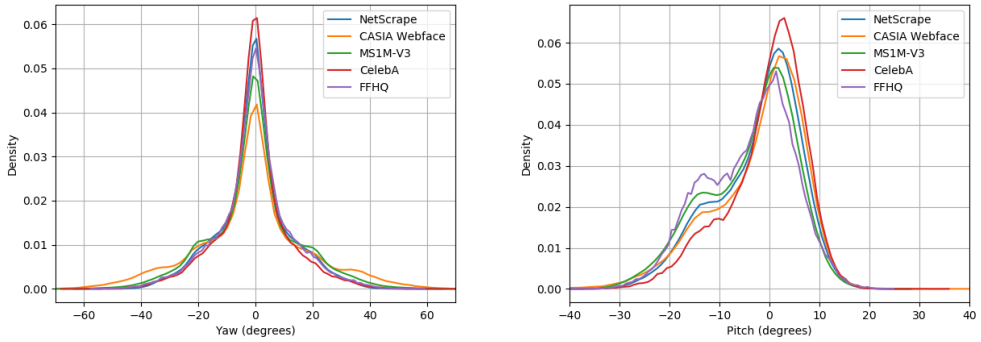


Figure 5.6: The relative pose distributions of the datasets used in the experiments described in Section 5.4.2.2 and Table 5.2.

training to provide a cleaner comparison where the dataset has been used to train the 3D GAN.) These datasets were then augmented using the 3D GAN trained on either CelebA or Flickr-Faces-HQ. Details of these datasets are presented in Table 5.1. We also show the distributions of detected yaw and pitch angles in Figure 5.6. CelebA was found to have the narrowest ranges of both yaw and pitch. Despite this, in conjunction with the 3D GAN, we were able to use the dataset to improve large-pose facial recognition. CASIA Webface displays a noticeably wider distribution of yaw angles than the other datasets. Again, despite this prior advantage, we were able to improve FR results above the CASIA baselines.

Synthetic datasets of 10k, 20k and 30k IDs were generated, each with 120 images per ID. Yaw and pitch angles were selected randomly from uniform distributions with ranges $[-90°, 90°]$ and $[-45°, 45°]$ respectively, whereas all other parameters (shape, expression, texture and background) were selected from a standard normal distribution, as during training. Synthetic images were augmented further using a spherical harmonic (SH) lighting model [Ramamoorthi & Hanrahan 2001]. We augmented using only white light and chose ambient and non-ambient lighting coefficients from random uniform distributions in the ranges $[0.6, 1.4]$ and $[-0.4, 0.4]$ respectively. In performing this lighting augmentation, we make the assumption that images in the synthetic training dataset are only ambiently lit. This is not the case, however, and learned textures contain problematic, embedded lighting effects. For example, a cast shadow may be coloured black in the texture. Applying the SH

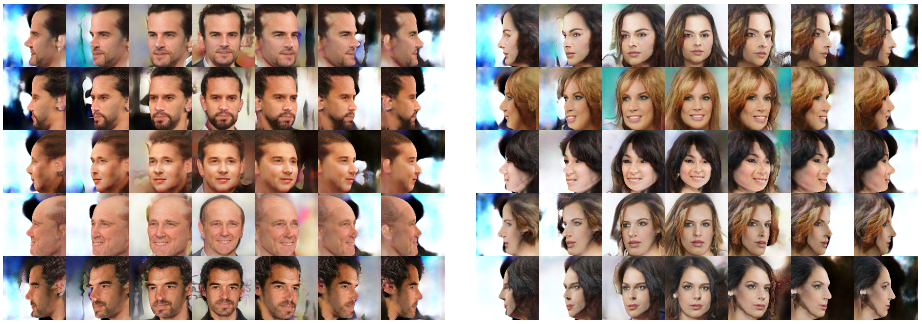


Figure 5.7: CelebA-like 3D GAN renderings at a range of yaw angles.

model may then brighten this region to give an unnatural grey colour rather than revealing a realistic facial texture. Nevertheless, performing this relatively crude lighting augmentation is shown to improve FR accuracy.

Examples of in-the-wild synthetic images can be seen in Figures 5.7 and 5.8. In Figure 5.7 we show a selection of images generated from CelebA with pitch, expression, background and lighting parameters set to 0. The images are generally of lower quality than those generated from Multi-PIE and display visible artefacts, particularly on the sides of the head. We suspect that this is due to a combination of the larger variation in textures and lighting conditions in CelebA, the lower number of images at large poses, and the absence of reliable pose labels. Despite these issues, our experiments show that the synthetic data is of adequate quality to successfully augment FR datasets. In Figure 5.8 we show a selection of images generated from FFHQ. These images have been cropped to 112×112 resolution as used in our data-augmentation experiments. All parameters were randomised, as described above.

5.4.2.2 Data-augmentation experiments

In all of our experiments we use the ResNet architecture of [Deng *et al.* 2019a] trained for 15 epochs. The only changes made were to the number of layers and to the loss function, as noted in Tables 5.2 and 5.3. Table 5.2 presents results for a series of experiments in which we augmented the NetScrape dataset with 3D GAN data generated from CelebA. Experiment 1 gives our baseline, trained only on the "NetScrape" dataset. Experiment 2 shows that the effect of adding in CelebA is to

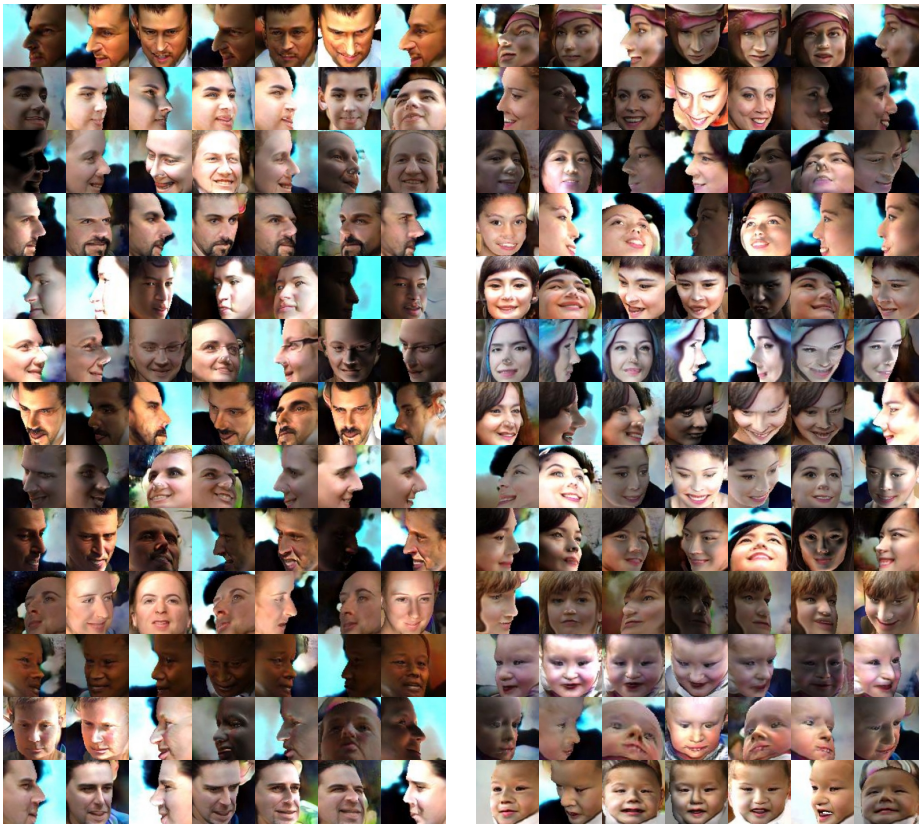


Figure 5.8: Random examples for a selection of IDs generated by the 3D GAN trained on FFHQ. The images have been cropped to 112×112 pixels for use in the experiments recorded in Table 5.3.

increase accuracy on CFP-FP and CPLFW by 0.47% and 0.25% respectively. The effect of adding the synthetic data, however, is to increase accuracy by up to 1.7% for CFP-FP, with an accuracy of 95.29% being achieved in Experiment 6, and by up to 1.69% for CPLFW, with an accuracy of 86.25% being achieved in Experiment 5; i.e. the 3D GAN was able to exploit the images of CelebA somewhere between three to six times more effectively. Experiments 3 and 4 show that disabling the spherical harmonic lighting, and limiting the variance of the pose to that detected in CelebA itself, each decrease accuracy on both CFP-FP and CPLFW, with limitation of the pose having the largest effect. Both experiments, however, still perform better than the baseline. Finally, in Experiments 5, 6 and 7, we augment the dataset with 10k, 20k and 30k synthetic identities. For each experiment the measured accuracies are above those of the baseline experiments, although performance drops for either 20k

Exp	Network	Loss	Training sets	Num IDs	Num images	CFP-FP	CPLFW
1	ResNet-34	ArcFace	NetScrape (in-house)	26.8k	3.5M	93.59%	84.56%
2	ResNet-34	ArcFace	NetScrape + CelebA	26.8k + 10.2k	3.5M + 0.2M	94.06%	84.81%
3	ResNet-34	ArcFace	NetScrape + 3D Synth (no SH)	26.8k + 10k	3.5M + 1.2M	94.46%	85.55%
4	ResNet-34	ArcFace	NetScrape + 3D Synth (narrow pose)	26.8k + 10k	3.5M + 1.2M	93.76%	84.93%
5	ResNet-34	ArcFace	NetScrape + 3D Synth	26.8k + 10k	3.5M + 1.2M	94.89%	**86.25%**
6	ResNet-34	ArcFace	NetScrape + 3D Synth	26.8k + 20k	3.5M + 2.4M	**95.29%**	85.96%
7	ResNet-34	ArcFace	NetScrape + 3D Synth	26.8k + 30k	3.5M + 3.6M	94.63%	85.91%

Table 5.2: A comparison of the effect dataset-augmentation on verification accuracies for the 7000 positive and negative frontal-profile pairs of the CFP dataset [Sengupta *et al.* 2016], and the 6000 positive and negative image pairs of CPLFW [Zheng & Deng 2018].

or 30k identities depending on the evaluation dataset. The reason for this decrease in performance could be due to synthetic identities being too densely sampled, i.e. with too many look-alikes being generated. Alternatively, it could be due to overfitting of the biometric network to 3D GAN data since, in Experiments 6 and 7, significant proportions of the training dataset were synthetic (40.7% and 50.7% as opposed to only 25.5% in Experiment 3).

Table 5.3 presents the results of experiments for comparison with the 3D model-based data-augmentation methods of [Deng *et al.* 2018] and [Gecer *et al.* 2020], and also with [Deng *et al.* 2019a] which had the state of the art accuracy for CPLFW. Results taken from the literature are highlighted in grey. The cleanest comparison is with the method of [Gecer *et al.* 2020] in which synthetic data generated by their TB-GAN was used to augment CASIA Webface giving an improvement of 1.56% from 95.56% to 97.12% verification accuracy on the Frontal-Profile protocol of CFP. Augmentation using 20k synthetic identities generated from FFHQ using our 3D GAN gave an improvement of 1.24% from the slightly lower baseline of accuracy of 95.50% up to 96.74%. Note that, in this experiment, the 3D GAN extracts useful information from the noisy FFHQ dataset, which is not accompanied by identity information, whereas the TB-GAN of [Gecer *et al.* 2020] is trained using a dataset of high-quality texture scans. Improvements in accuracy were also seen for CPLFW with addition of 10k and 20k synthetic identities leading to improvements of 0.84% and 1.16% respectively. Evaluation on the frontal protocol of CFP and on LFW gave only small improvements.

Method	FR network	Loss	Training sets	Method type	CFP-FF	CFP-FP	LFW	CPLFW
Human	Brain	-	-	-	96.24%	94.57%	97.27%	81.21%
Baseline	ResNet-27	Softmax	CASIA	-	98.59%	87.74%	99.02%	-
[Deng *et al.* 2018]	ResNet-27	Softmax	CASIA, (Multi-PIE, UMDFaces)	Aug (existing IDs)	98.83%	93.09%	99.22%	-
[Deng *et al.* 2018]	ResNet-27	Softmax	CASIA (Multi-PIE, UMDFaces)	Norm to 15°	-	94.05%	-	-
Baseline	ResNet-28	Softmax	CASIA	-	94.74%	84.76%	95.47%	68.01%
3D GAN (FFHQ)	ResNet-28	Softmax	CASIA	Aug (10k synth IDs)	95.44%	85.70%	95.97%	68.52%
Baseline	ResNet-50	ArcFace	CASIA	-	-	95.56%	-	-
[Gecer *et al.* 2020]	ResNet-50	ArcFace	CASIA	Aug (10k synth IDs)	-	97.12%	-	-
Baseline	ResNet-50	ArcFace	CASIA	-	99.37%	95.50%	99.30%	85.69%
3D GAN (FFHQ)	ResNet-50	ArcFace	CASIA	Aug (10k synth IDs)	99.49%	96.40%	99.35%	86.53%
3D GAN (FFHQ)	ResNet-50	ArcFace	CASIA	Aug (20k synth IDs)	99.40%	96.74%	99.42%	86.85%
[Deng *et al.* 2019a]	ResNet-100	ArcFace	MS1M-V2	-	-	-	99.82%	92.08%
Baseline	ResNet-100	ArcFace	MS1M-V3	-	**99.90%**	98.47%	**99.87%**	93.36%
3D GAN (FFHQ)	ResNet-100	ArcFace	MS1M-V3	Aug (20k synth IDs)	**99.90%**	**98.51%**	99.85%	**93.53%**

Table 5.3: A comparison of data-augmentation using synthetic identities generated by the 3D GAN with various similar methods from the literature (highlighted in grey). Evaluation is performed for the frontal-frontal (FF) and frontal-profile (FP) protocols of the CFP dataset as well as for LFW (view 2) and CPLFW. Datasets parenthesised in the "Training sets" column are FR datasets used to train the data-generation networks but not the FR network.

In a second set of experiments we scaled down the ResNet to have 28 layers and trained using a standard softmax loss in order to compare more closely with the work of [Deng *et al.* 2018]. These experiments, again, showed consistent improvements above baseline accuracies for all evaluation datasets. The frontal-profile accuracy for CFP did not come close to the accuracies of 93.09% and 94.05% achieved by [Deng *et al.* 2018]. This is perhaps not surprising, however, given the initially higher baseline accuracy and the additional high-quality data used during training of their texture-completion network. (Use of this additional training data is indicated in the "Training sets" column of the table.)

Finally, experiments were performed for a ResNet-100 architecture trained on the MS1M-V3 dataset. Augmentation using 20k synthetic identities generated from FFHQ using our 3D GAN gives a state-of-the-art accuracy of 93.53% on CPLFW.

5.5 Conclusions

We proposed a novel 3D GAN formulation for learning a nonlinear texture model from in-the-wild images and thereby generating synthetic images of new identities with fully disentangled pose. Unlike other similar methods, the 3D GAN does not require a training set of specially captured texture scans. We demonstrated that images synthesised by our 3D GAN can be used successfully to improve the accuracy of large-pose facial recognition. Finally, since the 3D GAN can generate images of new identities, it provides an avenue for extraction of useful information from noisy datasets such as FFHQ.

Chapter 6

Robustness of facial recognition to morphing attacks

6.1 Introduction

The potential threat of morphing attacks to systems secured by facial recognition (FR) was first identified in the 2014 paper "The Magic Passport" [Ferrara *et al.* 2014]. The paper demonstrated the relative ease with which images can be manipulated to simultaneously resemble multiple identities using commercially available tools, and the vulnerability of FR systems to those images. The extent to which face-morphing as a method of attack has been adopted by criminals is not known since, by definition, successful attacks remain undetected. Nevertheless, a pre-emptive arms race was spawned in the literature, with evermore sophisticated morphing methods being proposed in conjunction with tools for their detection [Damer *et al.* 2018, Debiasi *et al.* 2018, Ferrara *et al.* 2018, Scherhag *et al.* 2018, Seibold *et al.* 2018]. Various datasets of morphed examples have been made publicly available [Mahfoudi *et al.* 2019, Raghavendra *et al.* 2017] and an ongoing morphing detection benchmark has been included as part of NIST's Face Recognition Vendor Test (FRVT) [Ngan *et al.* 2020].

There are three broad approaches that might be taken to prevent morphing attacks:

1. **Trusted capture.** In the prelude to a morphing attack, the accomplice exploits his freedom to provide an image to the issuing authority. Removing this freedom by enforcing live image-capture at the time of application would make attacks significantly more challenging to perpetrate.

2. **Morph-detection.** Although currently known morphing methods produce images of high quality, none of them is perfect. Morphed images may contain certain features that betray their dubious provenance. Deploying automated detection of these features, either prior to creation of the identity document or at the time of use, could potentially prevent attacks.

3. **Robustness of recognition.** A morphed image contains an identity that is neither that of the accomplice nor of the imposter. A facial recognition system that is effective enough to recognise the identity as such would not be vulnerable to the attack.

In this chapter we consider the third approach and evaluate the effect that improvements to FR systems have on the success rate of morphing attacks. We evaluate the robustness of two FR algorithms to two morphing methods that make use of style-based generative networks; specifically, we make use of the generator of StyleGAN [Karras *et al.* 2019] pre-trained on the Flickr-Faces-HQ (FFHQ) dataset. At the time of writing, style-based GAN morphing methods had not been evaluated in the literature in the context of face-morphing attacks. The similar work of [Venkatesh *et al.* 2020] has since been published which evaluates a method similar to the "midpoint method" presented here. Whereas [Venkatesh *et al.* 2020] focusses on assessment of the extent to which FR systems are vulnerable to GAN-based face-morphing attacks in comparison to landmark-based attacks, and also on detection of such attacks, here we focus on the changing response of FR systems to morphed images as fidelity improves. We observe that improvements to FR systems do not necessarily translate to improved robustness to morphing attacks and that morphed

images should be taken into account when setting acceptance thresholds. We also introduce and evaluate a second style-based morphing method: the "dual biometric method". Finally, we show that FR networks trained using synthetic, 3D GAN images demonstrate improved robustness to morphing attacks.

The rest of the chapter is organised as follows: in Section 6.2 we discuss work in the literature proposing to complement FR systems with algorithms for detection of morphed images, and also work leading to the development of the style-based morphing methods proposed here; in Section 6.3 we describe the style-based morphing methods being evaluated, providing results in Section 6.4; In Section 6.4.3 we analyse the effect of training with synthetic 3D GAN images on the success of simulated morphing attacks; and in Section 6.5 we draw conclusions.

6.2 Related Work

6.2.1 Securing systems against morphing attacks

The largest part of the face-morphing attack literature consists of the development of methods for the detection of morphs, for example, by using deep learning techniques [Damer *et al.* 2018], analysis of sensor noise in images [Debiasi *et al.* 2018], detection of landmark shifts [Scherhag *et al.* 2018], verification of the consistency of lighting conditions [Seibold *et al.* 2018], or by de-morphing images to reveal the original subject [Ferrara *et al.* 2018]. It is generally accepted, however, that detection methods are ineffective and suffer from high error rates that worsen when morphed images are printed and scanned [Makrushin & Wolf 2018, Scherhag *et al.* 2017b]. In [Makrushin & Wolf 2018] it is recommended that identity document-issuing authorities enforce the submission of high-resolution digital images. However, they also point out that attackers could still manipulate digital noise signatures to obfuscate traces of image editing. In a recent survey of morphing attacks and detection methods [Scherhag *et al.* 2019] it was concluded that morphing attack detection methods do not generalise well to datasets incorporating real-world capture conditions. Indeed, in the most recent FRVT morph detection report [Ngan *et al.* 2020], the best value of APCER@BPCER=0.01 (Attack Pre-

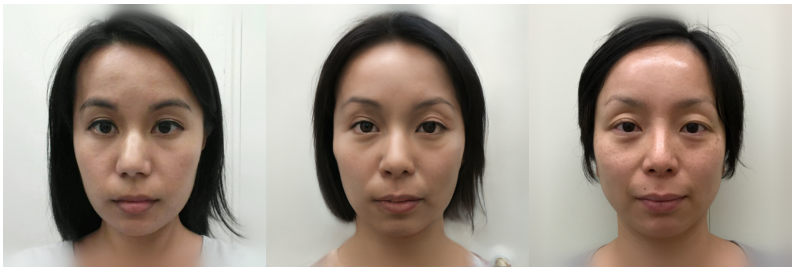


Figure 6.1: StyleGAN midpoint morph of NIST subjects A and B. Images of subject A (left) and B (right) were taken from [Ngan *et al.* 2020]. The central image is the morph.

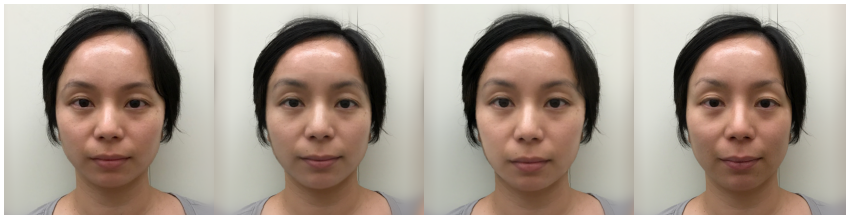


Figure 6.2: Examples of the output of various alternative, automated morphing methods taken from [Ngan *et al.* 2020]. These correspond to Figure 2 (g), (i), (j) and (l) of the NIST report.

sentation Classification Error Rate at a *Bona fide* Presentation Classification Error Rate of 0.01) for detection of morphs of the types shown in Figure 6.2 (i.e. the "Local Morph Colorized Match", "Splicing", "'Combined" and "DST" methods) was 88% for the "Splicing" method.

An assessment of the vulnerability of FR to the *average* of images of two identities [Raghavendra *et al.* 2017] showed it to be a more effective method than morphing. They also showed, however, that the averaged images were much easier to detect. It is unlikely, therefore, that an attacker would choose this type of method. In this work we propose two StyleGAN-based *morphing* methods and, in light of the evident difficulty of detecting morphs, we instead focus on demonstrating the effect on morphing attacks of improvements to the robustness of FR algorithms.

6.2.2 The development of style-based face-morphing

Generative Adversarial Networks (GANs) [Goodfellow *et al.* 2014] learn to map latent vectors of random values to points on a manifold in data-space, usually image-

space, representing realistic data-samples that can fool a concurrently trained discriminator into classifying them as real samples. Typically, generator architectures take a similar form to other deep neural networks, starting with the input - in this case the random vector - and applying a series of convolutions. In [Karras *et al.* 2019], however, the vector of random values is projected to each convolutional layer of the network and used to directly influence the scale of variation in each feature map of each convolutional layer. This is achieved via conditional instance normalisation [Dumoulin *et al.* 2017b], which was originally introduced as a method to manipulate the styles of images via the use of image-to-image translation networks [Isola *et al.* 2017]. Since the generators of GANs do not translate images but *grow* them, each convolutional layer naturally learns to control image features at different scales. For example, pose is controlled by early, large-scale features at low resolutions, whereas the presence of wrinkles is controlled later on, at higher resolutions [Karras *et al.* 2019]. This natural, scale-wise disentanglement (in conjunction with the "style mixing" used in [Karras *et al.* 2019]) causes the projections of the latent vector - the so called "w^+" vectors - to be largely independent of one another. It is this independence that makes style-based GANs particularly suitable for image reconstruction and then morphing.

In order to use GANs to perform morphing attacks, one needs to be able to *invert* the generator, i.e. to find the latent vector that best describes some input image. There are various one-shot ways to do this, for example, one could train an encoder to regress the latents from synthetic images or, alternatively, train an encoder via Adversarially Learned Inference (ALI) [Dumoulin *et al.* 2017a, Donahue *et al.* 2017] as was done in [Damer *et al.* 2018]. However, it is more effective, albeit slower, to find the latents using some iterative gradient descent method. Typically, it is difficult to fit GANs to non-synthetic images; so much so that the failure to reproduce images precisely has been used as a evidence that memorisation of images is not taking place in GANs [Webster *et al.* 2019]. However, in [Abdal *et al.* 2019] it was noticed that precise reconstructions could be achieved by treating the projected latents of StyleGAN independently during fitting, thereby taking advantage of the aforementioned scale-wise disentanglement. (This increase

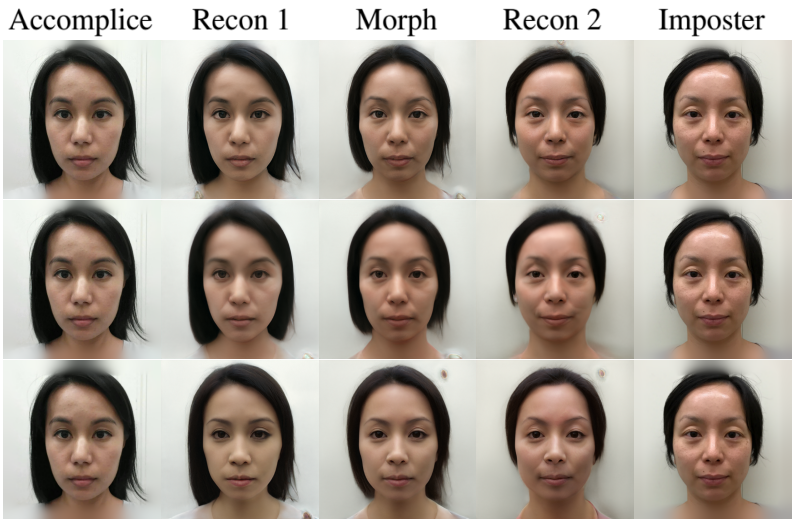


Figure 6.3: Ablation tests of the midpoint morphing method. Top - Results of the full method as described in Section 6.3.1; middle - perceptual loss and regularisation of the latent vector removed (reconstruction of pixel intensities only); bottom - using non-independent latent vectors at each convolutional layer. (The full loss was used, as in the top row.)

in precision can be seen by comparing the reconstructed images in the top and bottom rows of Figure 6.3.) It is then straightforward to generate realistic face-morphs by linearly interpolating between two sets of recovered w^+ vectors. In this work, we also manipulate images in the w^+ latent space. Further details of our methods are given in the following section.

6.3 Face-morphing with StyleGAN

We will evaluate robustness of FR algorithms to two different methods of face-morphing based on StyleGAN: the "midpoint method", which is similar to that demonstrated in [Abdal *et al.* 2019] and [Venkatesh *et al.* 2020], and the "dual biometric method", which has been developed for this study. In both methods we optimise the loss functions using Adam [Kingma & Ba 2015]. To speed up convergence and improve reconstruction quality, initialisations of the latent vectors are provided by a one-shot encoder trained on pairs of random vectors and associated synthetic images.

6.3.1 The midpoint method

Face-morphing using the midpoint method consists of two steps: recovering the two latent vectors that best describe two input images, and generating a synthetic image from the midpoint interpolation of those two vectors. To recover $\mathbf{w}^+$ for an input image $\mathbf{x}$, the following loss function is minimised:

$$\mathcal{L}_{w^+} = \mathcal{P}(G(\mathbf{w}^+), \mathbf{x}) + \frac{\lambda_r}{N_x}||G(\mathbf{w}^+) - \mathbf{x}||_2^2 + \lambda_w||\mathbf{w}^+ - \overline{\mathbf{w}}||_1 \tag{6.1}$$

where G is the generator (with StyleGAN's mapping network removed), N_x is the number of image pixels, $\overline{\mathbf{w}}$ is the average of the $\mathbf{w}^+$ seen during training of G, and $\mathcal{P}$ is a perceptual loss given by

$$\mathcal{P}(G(\mathbf{w}^+)), \mathbf{x}) = \frac{\lambda_v}{N_v}||VGG_9(G(\mathbf{w}^+) - VGG_9(\mathbf{x})||_2^2 + \lambda_m(1 - MSSSIM(G(\mathbf{w}^+), \mathbf{x})) \tag{6.2}$$

where VGG_9 is the output of the ninth layer of the VGG classification network [Simonyan & Zisserman 2015] used to extract discriminative features, N_v is the number of VGG features, and $MSSSIM$ is the Tensorflow implementation of the MS-SSIM metric described in [Wang *et al.* 2003]. The generator, G, is the official version of the StyleGAN generator trained on the Flickr-Faces-HQ dataset. The code implementing the inversion of StyleGAN's generator was taken from [Baylies 2019] and the coefficients weighting each term of equations (6.1) and (6.2) are left at their default values of $\lambda_r = 1.5$, $\lambda_w = 0.5$, $\lambda_v = 0.4$ and $\lambda_m = 200$. Once two vectors, $\mathbf{w}_1^+$ and $\mathbf{w}_2^+$ have been recovered, the final morphed image is given by

$$\mathbf{x}_{morph} = G(\frac{\mathbf{w}_1^+ + \mathbf{w}_2^+}{2}) \tag{6.3}$$

In the middle row of Figure 6.3 we demonstrate the effect of reverting the method to that used in [Abdal *et al.* 2019] by removing the perceptual loss term and the regularisation of $\mathbf{w}^+$ from equation (6.1) during latent recovery. The reconstructed images as well the midpoint morph become more blurred, lacking in high-frequency detail. This result motivates our use of the full, perceptual loss function in our

experiments. Results showing the level of robustness of FR to morphing attacks using this method are given in Section 6.4.1.

6.3.2 The dual biometric method

Although the latent space of StyleGAN is disentangled with respect to some scale-dependent features, identity features are not necessarily disentangled. This means that the equality

$$B(\mathbf{x}_{morph}) = \frac{1}{2}B(G(\mathbf{w}_1^+)) + \frac{1}{2}B(G(\mathbf{w}_2^+)) \tag{6.4}$$

where B is a biometric network producing an identity feature vector, does not necessarily hold; i.e. the identity of the midpoint morph does not necessarily lie between the identities of the two reconstructed images in a biometric feature space. A more reliable method of ensuring that the morphed identity remains close to each of the original identities could be to explicitly minimise those distances in the feature space. This motivates our dual biometric method in which the following cost function is minimised:

$$\mathcal{L}_{w^+} = ||B(G(\mathbf{w}^+)) - B(\mathbf{x_1})||_2^2 + ||B(G(\mathbf{w}^+)) - B(\mathbf{x_2})||_2^2 + \lambda_w||\mathbf{w}^+ - \overline{\mathbf{w}}||_1 \tag{6.5}$$

For B we used a Keras implementation of the VGGFace2 "SENet" network [Cao *et al.* 2018b] taken from [Malli 2020]. We have also included the same L1 regularisation of the latent vector as was used in the midpoint method. Since the biometric loss terms are robust to (i.e. ignore) all image features except for the identity, the L1 regularisation is important for maintaining a realistic looking image. λ_w was tuned by hand based on the appearance of a handful of morphed images and set to a value of 3. Results showing the level of robustness of FR to morphing attacks using this second method are given in Section 6.4.2.

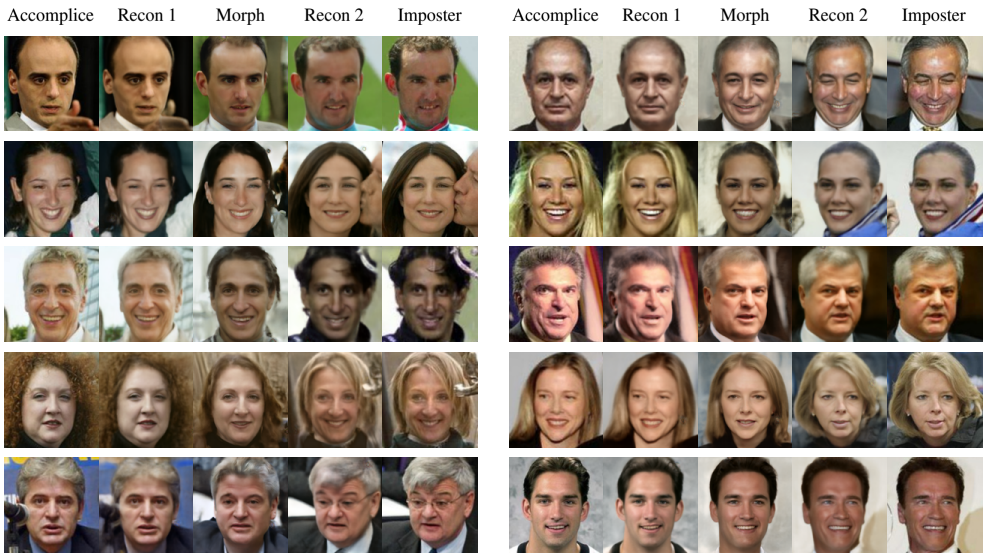


Figure 6.4: Examples of image-reconstructions and morphs produced using the midpoint method. The set of morphs in the left half of the figure represent successful attacks against Algo. 2017 but not Algo. 2019 with an acceptance threshold at FRR=0.25%. Attacks using the set of morphs to the right were successful against both Algo. 2017 and Algo. 2019.

6.4 Experiments

We evaluate robustness of FR to morphing attacks using the Labeled Faces in the Wild (LFW) dataset [Huang *et al.* 2007]. To simulate realistic morphing attacks we first select the highest quality image for each of the 5478 identities. We then assign fifty random "friends" to each identity and select the strongest identity match to be the accomplice, as judged by a biometric matching algorithm. Morphed images were produced for each of these image pairs using both the midpoint method and our dual biometric method. The original, *bona fide* images were then matched against the morph mated with those two images. Note that we do not compare morphs with independent images of the mated subjects. Comparisons are made with the *bona fide* images used to create the morph meaning that matching scores are likely to be at a maximum thus giving exaggerated, conservative estimates of FR system vulnerability. We present results for two matching algorithms, the first based on DeepVisage [Hasnat *et al.* 2017] that we refer to as "Algo. 2017" and the second based on ArcFace [Deng *et al.* 2019a] that we refer to as "Algo. 2019".

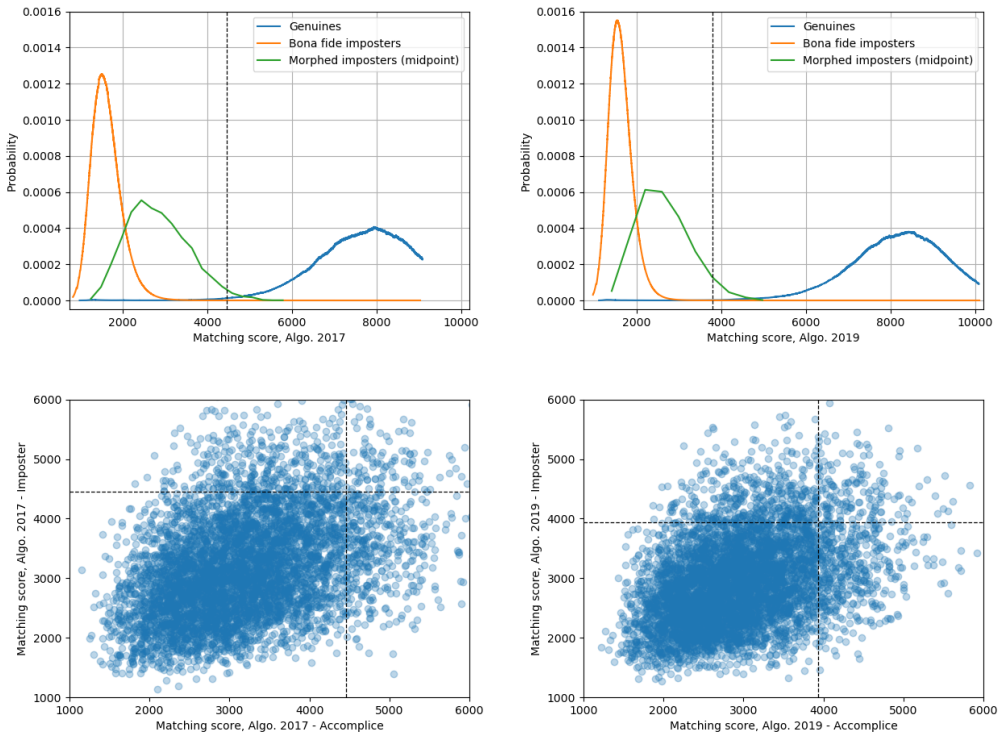


Figure 6.5: Distributions of matching scores for Algo. 2017 (left) and Algo. 2019 (right). Morphed imposters were produced using the **midpoint method**. Dashed lines represent thresholds of FAR=1×10^{-5} for *bona fide* imposters.

6.4.1 Results - The midpoint method

Figure 6.4 gives examples of face-morphs generated from pairs from LWF using the midpoint method. In Figure 6.5 we plot distributions of matching scores produced for this type of morph by the 2017 and 2019 algorithms. The green, "Morphed imposters" curves show the distributions of the Minimum Mated Morph Similarity Scores (MMMSS), i.e. the minimum of either the accomplice-morph or morph-imposter matching score. The minimum score is interesting since is the strength of the weakest similarity that determines whether the attack as a whole succeeds. The blue, "Genuines" curves show the distribution of mated matching scores for sets of *bona fide* images from LFW sharing the same identity, and the orange, "Imposters" curves show non-mated matching scores. In each figure we have drawn a threshold at the score corresponding to a False Acceptance Rate (FAR) of 1×10^{-5} based on the distribution of imposters (not of the morphs). Values of Mated Morph

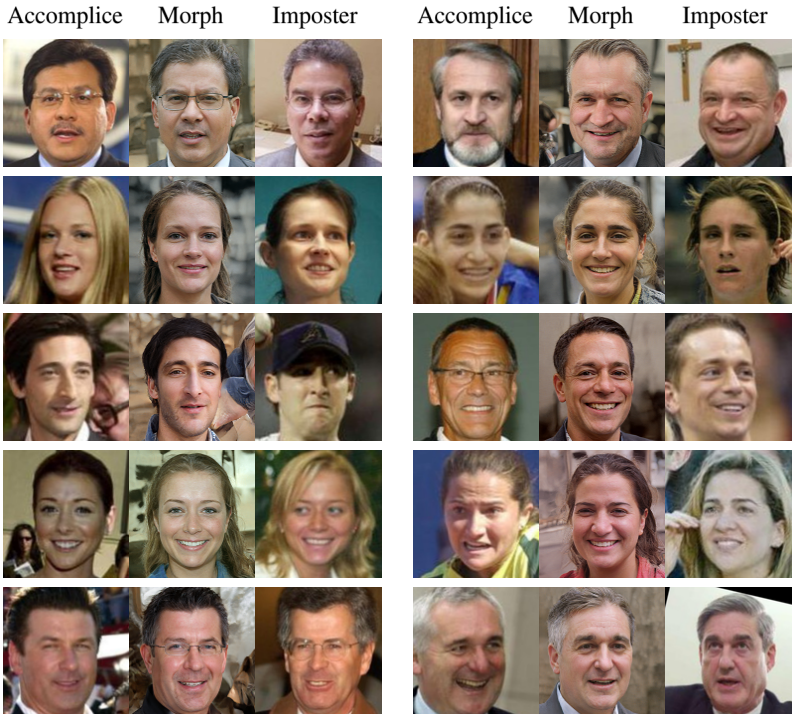


Figure 6.6: Examples of morphs produced using the dual biometric method. The set of morphs in the left half of the figure represent successful attacks against Algo. 2017 but not Algo. 2019 with an acceptance threshold at FRR=0.25%. Attacks using the set of morphs to the right were successful against both Algo. 2017 and Algo. 2019.

Presentation Match Rate (MMPMR) [Scherhag *et al.* 2017a] at this threshold are presented in Table 6.1 and correspond to the proportions of points lying in the top-right quadrant of the scatter plots in Figures 6.5 and 6.7.

In typical circumstances, both algorithms are able to well separate the distributions of mated and non-mated matching scores. However, inclusion of the MMMSS blurs this separation and at FAR=1×10^{-5} the success rate of simulated morphing attacks, the MMPMR, is 1.99% for Algo. 2017 and 2.96% for Algo. 2019, i.e. the MMPMR is three orders of magnitude larger than the FAR. What is more, despite the Genuine and Imposter curves clearly being better separated by Algo. 2019, the value of MMPMR@FAR=1×10^{-5} increases. This is not because the MMMSS become less distinguishable from *bona fide* mated matching scores; in fact, from Figure 6.5 we also see an improvement in separation of the Genuine and Morphed imposter curves. The issue arises from the fact that the improvement in separation of the

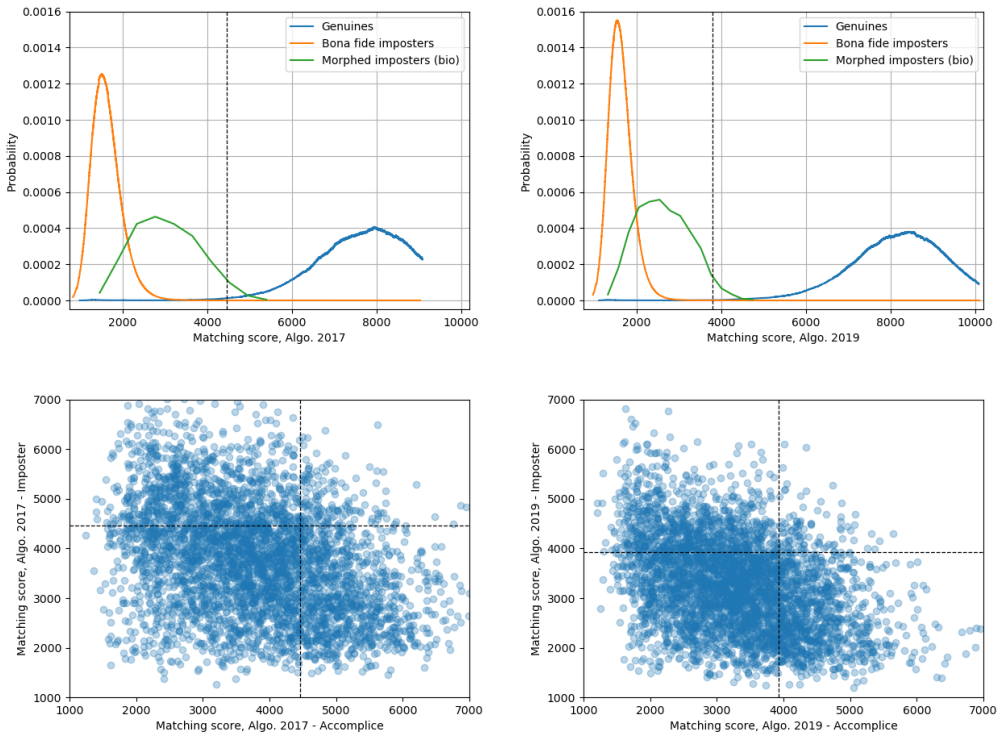


Figure 6.7: Distributions of matching scores for Algo. 2017 (left) and Algo. 2019 (right). Morphed imposters were produced using the **dual biometric method**. Dashed lines represent thresholds of FAR=1×10^{-5} for *bona fide* imposters.

bona fide mated and non-mated matching scores is larger than for the MMMSS. This acts to shift the threshold of FAR=10^{-5} to a lower score that "overtakes" the improvements in MMMSS. This means that we cannot assume that improvements to FR systems will lead directly to increased robustness to morphing attacks. Instead, the response of the FR system to datasets of morphed images should be considered when setting operational acceptance thresholds.

Table 6.2 shows values of MMPMR, and also FAR, at a threshold corresponding to FRR=0.73% for the *bona fide* mated pairs of LFW. We see that this corresponds to the original threshold of FAR=1×10^{-5} for Algo. 2017 but that for Algo. 2019 it corresponds to FAR=1.81×10^{-6}. At this much more stringent threshold, MMPMR drops to 0.07% for Algo. 2019 (and drops to 0% for the morphs produced by the dual biometric method). This means that, by compromising on improvements to FRR, essentially *all* face-morphing attacks of the type presented here can be prevented.

Morphing method	Algo. 2017	Algo. 2019
Bone fide imposters (FAR)	1×10^{-5}	1×10^{-5}
Genuines (FRR)	0.73%	0.25%
Midpoint Morphs (MMPMR)	1.99%	2.96%
Biometric Morphs (MMPMR)	3.88%	2.34%

Table 6.1: MMPMRs and FRR at a False Acceptance Rate of 1×10^{-5} for two different face-recognition algorithms.

Morphing method	Algo. 2017	Algo. 2019
Bone fide imposters (FAR)	1×10^{-5}	1.81×10^{-6}
Genuines (FRR)	0.73%	0.73%
Midpoint Morphs (MMPMR)	1.99%	0.07%
Biometric Morphs (MMPMR)	3.88%	0.00%

Table 6.2: MMPMRs and FAR at a False Rejection Rate of 0.73% for two different face-recognition algorithms.

[Scherhag *et al.* 2017a] suggests reporting Relative Morph Match Rate (RMMR) as a measure of FR system vulnerability where RMMR = MMPMR + FRR. This measure varies with threshold, however, and implicitly weights robustness to morphs and low FRR as being equal in priority, which is not necessarily the case. We find it preferable to observe the compromise between MMPMR and FRR by plotting the relevant ROC curve, as has been done in Figure 6.8. Here we see that the ROC curves for Algo. 2019 are significantly steeper than for Algo. 2017 indicating that accepting only a small increase in FRR can cause the success rates of morphing attacks to plummet relative to those measured against Algo. 2017.

6.4.2 Results - The dual biometric method

Figure 6.9 gives a direct comparison of face-morphs generated using the dual biometric method ("BioMorph") with those generated by the midpoint method ("MidMorph"). We see that the biometrically morphed identities are plausible and that they are distinct from the midpoint morphs. We also notice that the image-quality of the bio-morphs is higher than that of the midpoint morphs. This is because the images of LFW are of a lower quality than the images of FFHQ used to train

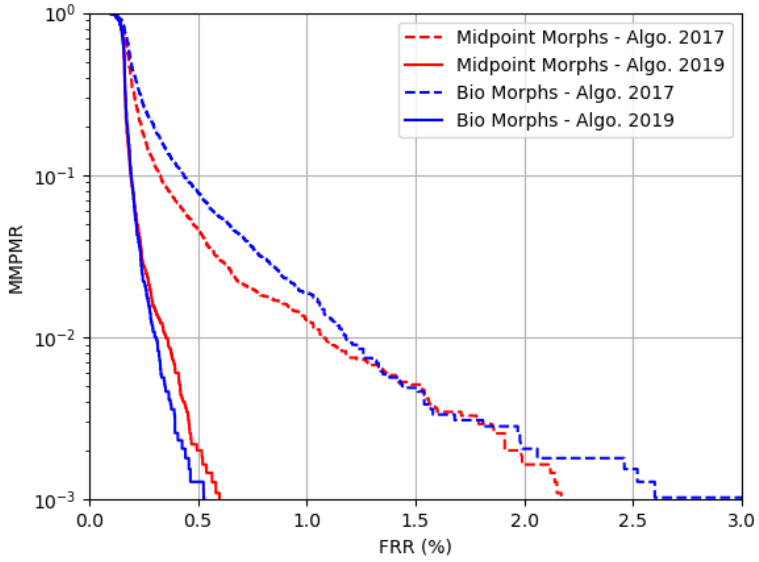


Figure 6.8: ROC curves showing the trade-off between MMPMR and FRR.

StyleGAN. Image artefacts from LFW therefore seep into the midpoint morphs via the image-reconstructions. During generation of the bio-morphs, the original images are never reconstructed. The only constraint is that similar identity-related features be generated.

Figure 6.7 shows the distributions of matching scores for the biometrically morphed images. From the scatter plots, we see that the matching scores are less balanced between accomplices and imposters, i.e. a large proportion of bio morphs were found to match one identity much more strongly than the other. This contrasts with the midpoint method where a larger proportion of morphs were found to give weak matching scores for both original images. From Table 6.1 we see that despite the imbalanced matching scores, MMPMR@FAR=1×10^{-5} is larger for biometric morphs than for midpoint morphs as measured by Algo. 2017. This situation reverses, however, for Algo. 2019 which succeeds in reducing the number of successful simulated attacks without modification to the threshold of FAR=1×10^{-5}. Table 6.2 shows that by sacrificing improvements to FRR and modifying the acceptance threshold of Algo. 2019 to FAR=1.81×10^{-6}, all simulated biometric morphing attacks can be prevented. Figure 6.6 (right) shows examples of bio-morphs that re-

Figure 6.9: Comparison of morphs generated using the midpoint and dual biometric methods.

main to be problematic for the 2019 algorithm with a threshold at FAR=1×10^{-5}. Those shown in Figure 6.6 (left) were previously problematic for Algo. 2017 but are correctly rejected by Algo. 2019.

By evaluating the dual biometric method on an in-the-wild dataset, we have inadvertently disguised the fact that desirable (for the attacker), non-identity image characteristics are lost. For example, Figure 6.10 (top) demonstrates the result of applying the biometric morphing method to the passport-style photographs of Figure 6.1. The generated morph resembles an in-the-wild image from FFHQ and would likely not be accepted for use on an identity document. To avoid this problem, a pixel-wise reconstruction loss can be applied to background regions. Figure 6.10 (bottom) shows an example of a biometric morph produced in this way. An alternative approach for an attacker to overcome this issue could be to train a GAN using solely passport-style images.

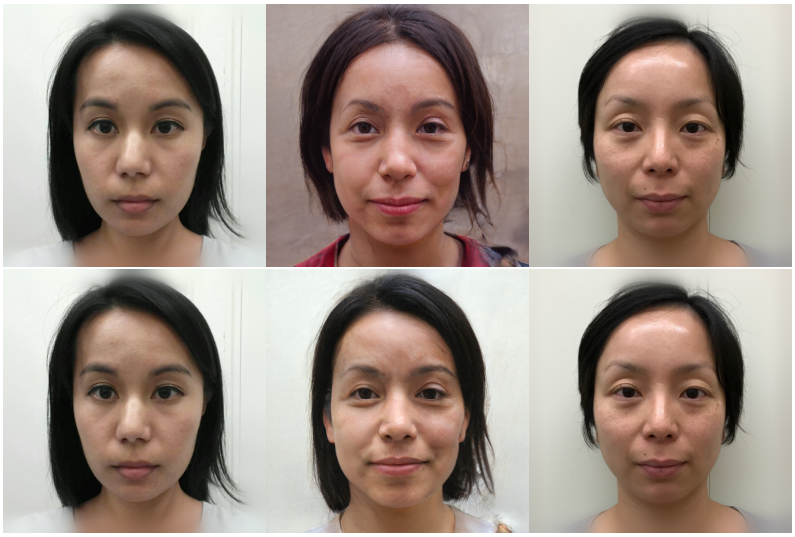


Figure 6.10: Demonstration of the dual biometric method applied to the passport-style images of Figure 6.1. Top: as described in equation (6.5); Bottom: with added reconstruction loss on the background regions.

6.4.3 The effect of training with synthetic identities on morphing attacks

In the previous chapter we saw that synthetic data generated by a GAN integrating a 3D morphable model can be used to improve the accuracy of large-pose facial recognition. (See Table 5.3.) The effect on performance when evaluating on datasets containing little pose variation, however, was less clear. Small improvements in accuracy were seen for LFW and so one would hope that this translates to improved robustness to our LFW-based, simulated morphing attacks. As we saw for the midpoint morphs, however, changes in response of the FR system to *bona fide* imposters are not necessarily equal to those to morphed imposters.

To evaluate the effect of training with synthetic data, we produced matching scores for LFW and for our two sets of morphed images using our baseline ResNet-50 network from the previous chapter, trained on CASIA Webface, and also the version augmented with 20k synthetic identities. In Figure 6.11 we plot the corresponding ROC curves showing the compromise between the attack success rate (MMPMR) and FRR. Overall, simulated attacks conducted using the midpoint morphing method were more successful against both networks with higher

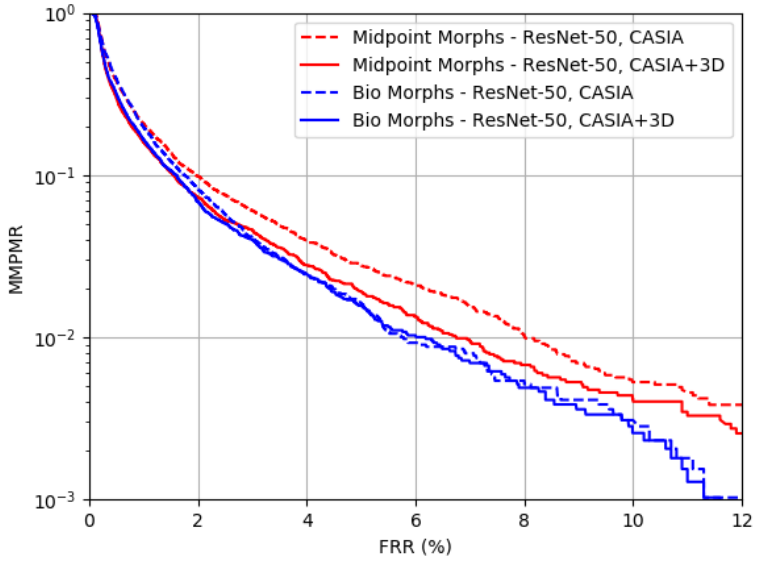


Figure 6.11: ROC curves showing the trade-off between MMPMR and FRR for biometric networks trained with and without synthetic 3D GAN data.

MMPMR at most values of FRR. Training with synthetic data was found to reduce the MMPMR of midpoint morphs at all FRR. For biometric morphs, the change in performance is noisier. However, clear improvements can be seen at lower values of FRR.

6.5 Conclusions

In this chapter we evaluated the robustness of two facial recognition algorithms to face-morphing attacks using a recent, StyleGAN-based morphing method. We also proposed and evaluated a second, related morphing method in which biometric distances of morphed faces from the contributing identities are minimised explicitly. Both morphing methods were found to be of potential threat with their relative success rates depending on the FR algorithm under attack. Assuming that we have been able to simulate realistic attack scenarios, it is likely that fewer than 3% of StyleGAN-based morphing attacks would succeed against a state-of-the-art facial recognition algorithm with a matching threshold set at FAR=1×10^{-5}. We also observed that improvements to FR algorithms do not necessarily translate directly

to increased robustness to face-morphing attacks and recommend that matching scores for datasets of morphed images be considered when setting operational acceptance thresholds. Finally, we showed that augmenting FR training datasets with synthetic 3D GAN data leads to increased robustness to both kinds of StyleGAN-based morphing method at low FRR.

Chapter 7

Conclusions and Future Work

Augmentation of facial recognition datasets with synthetic identities promises various advantages. For example, necessitation of the learning of a more discriminative feature-space, and the possibility of extracting useful information from noisy / unlabelled datasets. The work of this thesis investigated the use of GANs to generate such synthetic datasets and demonstrated these particular advantages.

We began by assessing the ability of GANs to generate images of subjects not found in the training dataset. It is widely believed that GANs are indeed capable of doing so. However, this belief is mainly based on observing qualitative differences between a handful of synthetic images and their nearest neighbours from the training dataset. Such analyses give little idea of whether the generators of GANs over-fit to the training dataset or not. Here, we provided analyses of full sets of biometric matching scores between synthetic and real datasets showing that they display similar dynamics to non-mated matching scores within training datasets of non-synthetic images. This allowed us to conclude that any over-fitting of generators is minimal and that subjects from training datasets are not being preferentially generated relative to any other possible identity in the biometric feature space. This conclusion validates the use of GANs for dataset anonymisation and for data-augmentation with synthetic identities.

Chapter 4 investigated the ability of GANs to disentangle identity from other image characteristics. We proposed a new method of gaining fine-control of labelled image characteristics but found that an additional biometric constraint was necessary to ensure consistent identity. We also evaluated the "SD-GAN" method of disentangling identity and proposed a modified, triplet-style loss incorporating an imposter term. Evaluation of these methods demonstrated our SD-GAN with

triplet loss to be the most successful method of disentangling identity. However, False Rejection Rates remain an order of magnitude larger than for real data indicating that identity-drift within synthetic sets of mated images remains to be a problem, particularly when augmenting pose to large angles.

In Chapter 5 we proposed a method of cleanly disentangling pose from identity in GANs by introducing a 3D shape model into the architecture of the generator. Unlike other generative methods employing 3D models and adversarial losses, our formulation does not involve reconstruction of training images and is therefore adapted for generation of new, synthetic identities. Augmenting FR datasets with sets of these synthetic identities was shown to improve biometric verification accuracy and we demonstrated state-of-the-art performance on the Cross-Pose LFW dataset.

Finally, analysis of the susceptibility of two recent FR algorithms to simulated face-morphing attacks showed the two proposed StyleGAN-based morphing methods to be threats, although fewer than 3% of attacks were successful for the plausible acceptance threshold of FAR=10^{-5}. A potentially important observation was made in that improvements to the fidelity of FR systems do not necessarily translate directly to improved robustness to morphing attacks. The response of FR algorithms to datasets of morphed images should therefore be taken into account when setting operational acceptance thresholds, and compromises to FRR made where necessary in order to ensure that morphed images can be reliably excluded. A similar analysis for an FR algorithm augmented with synthetic, 3D GAN data showed some improvement in resilience to attacks.

7.1 Future Work

Robustness to pose remains to be one of the greatest challenges for facial recognition. We believe data-augmentation using synthetic identities to be the simplest and most elegant solution, and strongly advocate further research into improved 3D methods. The current state-of-the-art in 3D generation is represented by [Gecer *et al.* 2020]. As discussed previously, however, the method depends on availability of a dataset

of high-quality 3D scans and corresponding textures. In contrast, the work of this thesis confirmed that data-augmentation using synthetic identities is feasible whilst making use of only in-the-wild images. Bridging the gap in image quality between the 3D GAN proposed here and that of [Gecer *et al.* 2020] whilst learning only from in-the-wild images is an important research direction.

In Chapter 5, various limitations of our 3D GAN method were identified. It was speculated that it might be possible to learn more realistic textures if the 3D-modelled conditions more closely reflected those found in the training datasets. This could be achieved, for example, by first fitting the 3DMM to the images of the training dataset to get sets of shape, expression and pose parameters, and then selecting the corresponding random parameters from those distributions during training of the 3D GAN. Similarly, shading of generated textures during training using a plausible 3D lighting model and distribution of lighting parameters may help to avoid lighting effects being learned as part of the texture. The linear shape model is also a limitation in the current formulation of the 3D GAN. Some nonlinear deviation from the model shape could be introduced in a similar way to [Tran & Liu 2018], although this is likely to lead to decreased stability during training.

Bibliography

[Abdal *et al.* 2019] Rameen Abdal, Yipeng Qin and Peter Wonka. *Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?* In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 4431–4440. IEEE, 2019. 109, 110, 111

[Ahonen *et al.* 2006] Timo Ahonen, Abdenour Hadid and Matti Pietikäinen. *Face Description with Local Binary Patterns: Application to Face Recognition.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pages 2037–2041, 2006. 1

[Arjovsky *et al.* 2017] Martín Arjovsky, Soumith Chintala and Léon Bottou. *Wasserstein Generative Adversarial Networks.* In Doina Precup and Yee Whye Teh, editeurs, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017. xi, 16, 17, 18, 19, 92

[Bansal *et al.* 2017] Ankan Bansal, Anirudh Nanduri, Carlos Domingo Castillo, Rajeev Ranjan and Rama Chellappa. *UMDFaces: An annotated face dataset for training deep networks.* In 2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017, pages 464–473. IEEE, 2017. 31

[Bao *et al.* 2018] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li and Gang Hua. *Towards Open-Set Identity Preserving Face Synthesis.* In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6713–6722. IEEE Computer Society, 2018. 57

[Baylies 2019] Peter Baylies. *stylegan-encoder*. https://github.com/pbaylies/stylegan-encoder/tree/266d1eb2da09894adcb49879fbd0674e12cab739, 2019. 111

[Bengio *et al.* 2014] Yoshua Bengio, Eric Laufer, Guillaume Alain and Jason Yosinski. *Deep Generative Stochastic Networks Trainable by Backprop*. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 226–234. JMLR.org, 2014. xi, 11, 12

[Blanz & Vetter 1999] Volker Blanz and Thomas Vetter. *A Morphable Model for the Synthesis of 3D Faces*. In Warren N. Waggenspack, editeur, Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, Los Angeles, CA, USA, August 8-13, 1999, pages 187–194. ACM, 1999. 3, 26, 44, 65, 71, 87

[Booth *et al.* 2016] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah and David Dunaway. *A 3D Morphable Model Learnt from 10, 000 Faces*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 5543–5552. IEEE Computer Society, 2016. 27, 87

[Bozorgtabar *et al.* 2019] Behzad Bozorgtabar, Mohammad Saeed Rad, Hazim Kemal Ekenel and Jean-Philippe Thiran. *Using Photorealistic Face Synthesis and Domain Adaptation to Improve Facial Expression Analysis*. In 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019, pages 1–8. IEEE, 2019. 57

[Brock *et al.* 2017] Andrew Brock, Theodore Lim, James M. Ritchie and Nick Weston. *Neural Photo Editing with Introspective Adversarial Networks*. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. 63

[Cao *et al.* 2010] Zhimin Cao, Qi Yin, Xiaoou Tang and Jian Sun. *Face recognition with learning-based descriptor*. In The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pages 2707–2714. IEEE Computer Society, 2010. 1

[Cao *et al.* 2014] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong and Kun Zhou. *FaceWarehouse: A 3D Facial Expression Database for Visual Computing*. IEEE Trans. Vis. Comput. Graph., vol. 20, no. 3, pages 413–425, 2014. 26

[Cao *et al.* 2018a] Jie Cao, Yibo Hu, Bing Yu, Ran He and Zhenan Sun. *Load Balanced GANs for Multi-view Face Image Synthesis*. CoRR, vol. abs/1802.07447, 2018. 24, 32, 33

[Cao *et al.* 2018b] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi and Andrew Zisserman. *VGGFace2: A Dataset for Recognising Faces across Pose and Age*. In 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018, pages 67–74. IEEE Computer Society, 2018. 112

[Chen *et al.* 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever and Pieter Abbeel. *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets*. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 2172–2180, 2016. 57

[Chen *et al.* 2018] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao and Yung-Yu Chuang. *Deep Photo Enhancer: Unpaired Learning for Image Enhancement From Photographs With GANs*. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6306–6314. IEEE Computer Society, 2018. 21, 38, 63

[Choi *et al.* 2018] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim and Jaegul Choo. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation.* In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 8789–8797. IEEE Computer Society, 2018. 56, 57

[Ciresan *et al.* 2011] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella and Jürgen Schmidhuber. *Flexible, High Performance Convolutional Neural Networks for Image Classification.* In Toby Walsh, editeur, IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011, pages 1237–1242. IJCAI/AAAI, 2011. 1

[Crispell *et al.* 2017] Daniel E. Crispell, Octavian Biris, Nate Crosswhite, Jeffrey Byrne and Joseph L. Mundy. *Dataset Augmentation for Pose and Lighting Invariant Face Recognition.* CoRR, vol. abs/1704.04326, 2017. 24, 25, 26, 27, 85, 88

[Dai *et al.* 2017] Zihang Dai, Zhilin Yang, Fan Yang, William W. Cohen and Ruslan Salakhutdinov. *Good Semi-supervised Learning That Requires a Bad GAN.* In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6510–6520, 2017. 46

[Damer *et al.* 2018] Naser Damer, Alexandra Mosegui Saladie, Andreas Braun and Arjan Kuijper. *MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network.* In 9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018, pages 1–10. IEEE, 2018. 105, 107, 109

[Debiasi *et al.* 2018] Luca Debiasi, Ulrich Scherhag, Christian Rathgeb, Andreas Uhl and Christoph Busch. *PRNU-based detection of morphed face images.* In 2018 International Workshop on Biometrics and Forensics, IWBF 2018, Sassari, Italy, June 7-8, 2018, pages 1–7. IEEE, 2018. 105, 107

[Deng *et al.* 2018] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou and Stefanos Zafeiriou. *UV-GAN: Adversarial Facial UV Map Completion for Pose-Invariant Face Recognition.* In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 7093–7102. IEEE Computer Society, 2018. vii, viii, 24, 29, 31, 32, 37, 39, 85, 88, 101, 102

[Deng *et al.* 2019a] Jiankang Deng, Jia Guo, Niannan Xue and Stefanos Zafeiriou. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 3, 47, 99, 101, 102, 113

[Deng *et al.* 2019b] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia and Stefanos Zafeiriou. *RetinaFace: Single-stage Dense Face Localisation in the Wild.* CoRR, vol. abs/1905.00641, 2019. 97

[Donahue *et al.* 2017] Jeff Donahue, Philipp Krähenbühl and Trevor Darrell. *Adversarial Feature Learning.* In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. 109

[Donahue *et al.* 2018] Chris Donahue, Zachary C. Lipton, Akshay Balsubramani and Julian J. McAuley. *Semantically Decomposing the Latent Spaces of Generative Adversarial Networks.* In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. 55, 72, 73, 77, 78

[Dumoulin *et al.* 2017a] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro and Aaron C. Courville. *Adversarially Learned Inference*. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. 109

[Dumoulin *et al.* 2017b] Vincent Dumoulin, Jonathon Shlens and Manjunath Kudlur. *A Learned Representation For Artistic Style*. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. 63, 109

[Ferrara *et al.* 2014] Matteo Ferrara, Annalisa Franco and Davide Maltoni. *The magic passport*. In IEEE International Joint Conference on Biometrics, Clearwater, IJCB 2014, FL, USA, September 29 - October 2, 2014, pages 1–7. IEEE, 2014. 5, 105

[Ferrara *et al.* 2018] Matteo Ferrara, Annalisa Franco and Davide Maltoni. *Face demorphing in the presence of facial appearance variations*. In 26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018, pages 2365–2369. IEEE, 2018. 105, 107

[Gecer *et al.* 2018] Baris Gecer, Binod Bhattarai, Josef Kittler and Tae-Kyun Kim. *Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model*. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss, editeurs, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI, volume 11215 of *Lecture Notes in Computer Science*, pages 230–248. Springer, 2018. 24, 29, 30, 35, 45, 88

[Gecer *et al.* 2019] Baris Gecer, Stylianos Ploumpis, Irene Kotsia and Stefanos Zafeiriou. *GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction*. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 1155–1164. Computer Vision Foundation / IEEE, 2019. 87

[Gecer *et al.* 2020] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou and Stefanos Zafeiriou. *Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks.* In Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan-Michael Frahm, editeurs, Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX, volume 12374 of *Lecture Notes in Computer Science*, pages 415–433. Springer, 2020. 24, 37, 39, 45, 86, 87, 88, 101, 102, 124, 125

[Geirhos *et al.* 2019] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.* In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. 26

[Goodfellow *et al.* 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville and Yoshua Bengio. *Generative Adversarial Nets.* In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence and Kilian Q. Weinberger, editeurs, Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2672–2680, 2014. 3, 4, 9, 13, 108

[Gross *et al.* 2008] Ralph Gross, Latanya Sweeney, Fernando De la Torre and Simon Baker. *Semi-supervised learning of multi-factor models for face de-identification.* In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE Computer Society, 2008. 44

[Gross *et al.* 2010] Ralph Gross, Iain A. Matthews, Jeffrey F. Cohn, Takeo Kanade and Simon Baker. *Multi-PIE.* Image Vis. Comput., vol. 28, no. 5, pages 807–813, 2010. 94

[Gulrajani *et al.* 2017] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin and Aaron C. Courville. *Improved Training of Wasserstein GANs.* In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5767–5777, 2017. 19, 20, 21, 38, 63, 92

[Guo *et al.* 2018] Jianzhu Guo, Xiangyu Zhu, Zhen Lei and Stan Z. Li. *Face Synthesis for Eyeglass-Robust Face Recognition.* In Jie Zhou, Yunhong Wang, Zhenan Sun, Zhenhong Jia, Jianjiang Feng, Shiguang Shan, Kurban Ubul and Zhenhua Guo, editeurs, Biometric Recognition - 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings, volume 10996 of *Lecture Notes in Computer Science*, pages 275–284. Springer, 2018. 24, 36

[Härkönen *et al.* 2020] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen and Sylvain Paris. *GANSpace: Discovering Interpretable GAN Controls.* In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan and Hsuan-Tien Lin, editeurs, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. 77

[Hasnat *et al.* 2017] Md. Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric and Liming Chen. *DeepVisage: Making Face Recognition Simple Yet With Powerful Generalization Skills.* In 2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017, pages 1682–1691. IEEE Computer Society, 2017. 3, 61, 65, 113

[Hassner *et al.* 2015] Tal Hassner, Shai Harel, Eran Paz and Roee Enbar. *Effective face frontalization in unconstrained images.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June

7-12, 2015, pages 4295–4304. IEEE Computer Society, 2015. xii, 24, 25, 26, 28, 85

[Henderson & Ferrari 2020] Paul Henderson and Vittorio Ferrari. *Learning Single-Image 3D Reconstruction by Generative Modelling of Shape, Pose and Shading.* Int. J. Comput. Vis., vol. 128, no. 4, pages 835–854, 2020. 91

[Heusel *et al.* 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler and Sepp Hochreiter. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.* In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6626–6637, 2017. 42

[Hinton & Sejnowski 1983] Geoffrey E. Hinton and Terrence J. Sejnowski. *Optimal Perceptual Inference.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1983. 9

[Hu *et al.* 2018] Yibo Hu, Xiang Wu, Bing Yu, Ran He and Zhenan Sun. *Pose-Guided Photorealistic Face Rotation.* In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 8398–8406. IEEE Computer Society, 2018. 24, 32, 33

[Huang *et al.* 2007] Gary B. Huang, Manu Ramesh, Tamara Berg and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Rapport technique 07-49, University of Massachusetts, Amherst, October 2007. 113

[Huang *et al.* 2017] 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 IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 2458–2467. IEEE Computer Society, 2017. 24, 32, 33

[Hukkelas *et al.* 2019] Hakon Hukkelas, Rudolf Mester and Frank Lindseth. *DeepPrivacy: A Generative Adversarial Network for Face Anonymization.* In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang and Panpan Xu, editeurs, Advances in Visual Computing - 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7-9, 2019, Proceedings, Part I, volume 11844 of *Lecture Notes in Computer Science*, pages 565–578. Springer, 2019. 44

[Inc. 2016] Singular Inversions Inc. *FaceGen.* `http://www.facegen.com`, 2016. 27

[Isola *et al.* 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks.* In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 5967–5976. IEEE Computer Society, 2017. 109

[Karras *et al.* 2018] Tero Karras, Timo Aila, Samuli Laine and Jaakko Lehtinen. *Progressive Growing of GANs for Improved Quality, Stability, and Variation.* In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. 4, 41, 47, 63, 66, 75, 91

[Karras *et al.* 2019] Tero Karras, Samuli Laine and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. 41, 47, 92, 96, 97, 106, 109

[Karras *et al.* 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen and Timo Aila. *Analyzing and Improving the Image Quality of StyleGAN.* In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 8107–8116. IEEE, 2020. 41

[Kemelmacher-Shlizerman *et al.* 2016] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller and Evan Brossard. *The MegaFace Benchmark: 1 Million Faces for Recognition at Scale*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 4873–4882. IEEE Computer Society, 2016. 36

[Kingma & Ba 2015] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. In Yoshua Bengio and Yann LeCun, editeurs, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 110

[Kortylewski *et al.* 2018] Adam Kortylewski, Andreas Schneider, Thomas Gerig, Bernhard Egger, Andreas Morel-Forster and Thomas Vetter. *Training Deep Face Recognition Systems with Synthetic Data*. CoRR, vol. abs/1802.05891, 2018. vii, 24, 35, 36, 45, 86, 88

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou and Kilian Q. Weinberger, editeurs, Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, pages 1106–1114, 2012. 1

[Kynkäänniemi *et al.* 2019] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen and Timo Aila. *Improved Precision and Recall Metric for Assessing Generative Models*. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 3929–3938, 2019. 42

[Lai & Lai 2018] Ying-Hsiu Lai and Shang-Hong Lai. *Emotion-Preserving Representation Learning via Generative Adversarial Network for Multi-View Facial Expression Recognition.* In 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018, pages 263–270. IEEE Computer Society, 2018. 56

[Li *et al.* 2017a] Jerry Li, Aleksander Madry, John Peebles and Ludwig Schmidt. *Towards Understanding the Dynamics of Generative Adversarial Networks.* CoRR, vol. abs/1706.09884, 2017. 21

[Li *et al.* 2017b] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li and Javier Romero. *Learning a model of facial shape and expression from 4D scans.* ACM Trans. Graph., vol. 36, no. 6, pages 194:1–194:17, 2017. 86, 87, 91

[Lindt *et al.* 2019] Alexandra Lindt, Pablo V. A. Barros, Henrique Siqueira and Stefan Wermter. *Facial Expression Editing with Continuous Emotion Labels.* In 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019, pages 1–8. IEEE, 2019. 57

[Liu & Wechsler 2002] Chengjun Liu and Harry Wechsler. *Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition.* IEEE Trans. Image Process., vol. 11, no. 4, pages 467–476, 2002. 1

[Liu *et al.* 2015a] Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang. *Deep Learning Face Attributes in the Wild.* In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 3730–3738. IEEE Computer Society, 2015. 65

[Liu *et al.* 2015b] Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang. *Deep Learning Face Attributes in the Wild.* In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 3730–3738. IEEE Computer Society, 2015. 97

[Lucic *et al.* 2018] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly and Olivier Bousquet. *Are GANs Created Equal? A Large-Scale Study.* In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 698–707, 2018. 43

[Lv *et al.* 2017] Jiang-Jing Lv, Xiaohu Shao, Jia-Shui Huang, Xiang-Dong Zhou and Xi Zhou. *Data augmentation for face recognition.* Neurocomputing, vol. 230, pages 184–196, 2017. xii, 24, 25, 26, 27, 29, 30, 36, 88

[Mahfoudi *et al.* 2019] Gaël Mahfoudi, Badr Tajini, Florent Retraint, Frédéric Morain-Nicolier, Jean-Luc Dugelay and Marc Pic. *DEFACTO: Image and Face Manipulation Dataset.* In 27th European Signal Processing Conference, EUSIPCO 2019, A Coruña, Spain, September 2-6, 2019, pages 1–5. IEEE, 2019. 105

[Makrushin & Wolf 2018] Andrey Makrushin and Andreas Wolf. *An Overview of Recent Advances in Assessing and Mitigating the Face Morphing Attack.* In 26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018, pages 1017–1021. IEEE, 2018. 5, 107

[Malli 2020] Refik Can Malli. *keras-vggface.* https://github.com/rcmalli/keras-vggface/tree/9ac97da84f18e392f5009d83cafcc5359204a408, 2020. 112

[Marmoset 2019] Marmoset. *Marmoset Toolbag.* https://marmoset.co/toolbag, 2019. 37

[Marriott *et al.* 2020a] R. Marriott, S. Romdhani and L. Chen. *Taking Control of Intra-Class Variation in Conditional GANs Under Weak Supervision.* In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG), pages 283–290, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. xxiii, 55, 78

[Marriott *et al.* 2020b] Richard T. Marriott, Safa Madiouni, Sami Romdhani, Stéphane Gentric and Liming Chen. *An Assessment of GANs for Identity-related Applications*. In 2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020, pages 1–10. IEEE, 2020. xxiii, 55

[Marriott *et al.* 2020c] Richard T. Marriott, Sami Romdhani and Liming Chen. *A 3D GAN for Improved Large-pose Facial Recognition*. CoRR, vol. abs/2012.10545, 2020. xxiii

[Marriott *et al.* 2020d] Richard T. Marriott, Sami Romdhani, Stéphane Gentric and Liming Chen. *Robustness of Facial Recognition to GAN-based Face-morphing Attacks*. CoRR, vol. abs/2012.10548, 2020. xxiii

[Masi *et al.* 2016] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut and Gérard G. Medioni. *Do We Really Need to Collect Millions of Faces for Effective Face Recognition?* In Bastian Leibe, Jiri Matas, Nicu Sebe and Max Welling, editeurs, Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V, volume 9909 of *Lecture Notes in Computer Science*, pages 579–596. Springer, 2016. xi, 24, 25, 26, 85, 88

[Meden *et al.* 2017] Blaz Meden, Refik Can Malli, Sebastjan Fabijan, Hazim Kemal Ekenel, Vitomir Struc and Peter Peer. *Face deidentification with generative deep neural networks*. IET Signal Process., vol. 11, no. 9, pages 1046–1054, 2017. 44

[Mescheder *et al.* 2018] Lars M. Mescheder, Andreas Geiger and Sebastian Nowozin. *Which Training Methods for GANs do actually Converge?* In Jennifer G. Dy and Andreas Krause, editeurs, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 3478–3487. PMLR, 2018. 20, 21

[Mirza & Osindero 2014] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets.* CoRR, vol. abs/1411.1784, 2014. 22

[Miyato & Koyama 2018] Takeru Miyato and Masanori Koyama. *cGANs with Projection Discriminator.* In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. 63

[Miyato *et al.* 2018] Takeru Miyato, Toshiki Kataoka, Masanori Koyama and Yuichi Yoshida. *Spectral Normalization for Generative Adversarial Networks.* In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. 64

[Ngan *et al.* 2020] Mei Ngan, Patrick Grother, Kayee Hanaoka and Jason Kuo. *Face Recognition Vendor Test (FRVT) Part 4: MORPH Performance of Automated Face Morph Detection.* National Institute of Technology (NIST), Tech. Rep. NISTIR, vol. 8292, 2020. xvi, 105, 107, 108

[Odena *et al.* 2017] Augustus Odena, Christopher Olah and Jonathon Shlens. *Conditional Image Synthesis with Auxiliary Classifier GANs.* In Doina Precup and Yee Whye Teh, editeurs, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR, 2017. 64

[Odena 2016] Augustus Odena. *Semi-Supervised Learning with Generative Adversarial Networks.* CoRR, vol. abs/1606.01583, 2016. 45

[Peng *et al.* 2017] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas and Manmohan Chandraker. *Reconstruction-Based Disentanglement for Pose-Invariant Face Recognition.* In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 1632–1641. IEEE Computer Society, 2017. viii, 24, 25, 26, 29, 37, 88

[Perarnau *et al.* 2016] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu and Jose M. Álvarez. *Invertible Conditional GANs for image editing*. In NIPS Workshop on Adversarial Training, 2016. 63

[Pumarola *et al.* 2018] Albert Pumarola, Antonio Agudo, Aleix M. Martínez, Alberto Sanfeliu and Francesc Moreno-Noguer. *GANimation: Anatomically-Aware Facial Animation from a Single Image*. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss, editeurs, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X, volume 11214 of *Lecture Notes in Computer Science*, pages 835–851. Springer, 2018. 56, 57

[Raghavendra *et al.* 2017] Ramachandra Raghavendra, Kiran B. Raja, Sushma Venkatesh and Christoph Busch. *Face morphing versus face averaging: Vulnerability and detection*. In 2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017, pages 555–563. IEEE, 2017. 105, 108

[Ramamoorthi & Hanrahan 2001] Ravi Ramamoorthi and Pat Hanrahan. *An efficient representation for irradiance environment maps*. In Lynn Pocock, editeur, Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, Los Angeles, California, USA, August 12-17, 2001, pages 497–500. ACM, 2001. 28, 71, 98

[Roth *et al.* 2017] Kevin Roth, Aurélien Lucchi, Sebastian Nowozin and Thomas Hofmann. *Stabilizing Training of Generative Adversarial Networks through Regularization*. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 2018–2028, 2017. 20, 21

[Russakovsky *et al.* 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya

Khosla, Michael S. Bernstein, Alexander C. Berg and Fei-Fei Li. *ImageNet Large Scale Visual Recognition Challenge*. Int. J. Comput. Vis., vol. 115, no. 3, pages 211–252, 2015. 1

[Sajid *et al.* 2018] M. Sajid, Nouman Safdar Ali, S. H. Dar, Naeem Iqbal Ratyal, Asif Raza Butt, Bushra Zafar, Tamoor Shafique, M. J. A. Baig, Imran Riaz and S. Baig. *Data Augmentation-Assisted Makeup-Invariant Face Recognition*. Mathematical Problems in Engineering, vol. 2018, pages 1–10, 2018. vii, 24

[Sajjadi *et al.* 2018] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet and Sylvain Gelly. *Assessing Generative Models via Precision and Recall*. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 5234–5243, 2018. 42

[Salakhutdinov & Hinton 2009] Ruslan Salakhutdinov and Geoffrey E. Hinton. *Deep Boltzmann Machines*. In David A. Van Dyk and Max Welling, editeurs, Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009, volume 5 of *JMLR Proceedings*, pages 448–455. JMLR.org, 2009. 9

[Salimans *et al.* 2016] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford and Xi Chen. *Improved Techniques for Training GANs*. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 2226–2234, 2016. 35, 42, 45

[Samarzija & Ribaric 2014] Branko Samarzija and Slobodan Ribaric. *An approach to the de-identification of faces in different poses.* In 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014, Opatija, Croatia, May 26-30, 2014, pages 1246–1251. IEEE, 2014. 44

[Scherhag *et al.* 2017a] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Marta Gomez-Barrero, Raymond N. J. Veldhuis, Luuk J. Spreeuwers, Maikel Schils, Davide Maltoni, Patrick Grother, Sébastien Marcel, Ralph Breithaupt, Ramachandra Raghavendra and Christoph Busch. *Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting.* In Arslan Brömme, Christoph Busch, Antitza Dantcheva, Christian Rathgeb and Andreas Uhl, editeurs, International Conference of the Biometrics Special Interest Group, BIOSIG 2017, Darmstadt, Germany, September 20-22, 2017, volume P-270 of *LNI*, pages 149–159. GI / IEEE, 2017. 115, 117

[Scherhag *et al.* 2017b] Ulrich Scherhag, Ramachandra Raghavendra, Kiran B. Raja, Marta Gomez-Barrero, Christian Rathgeb and Christoph Busch. *On the vulnerability of face recognition systems towards morphed face attacks.* In 5th International Workshop on Biometrics and Forensics, IWBF 2017, Coventry, United Kingdom, April 4-5, 2017, pages 1–6. IEEE, 2017. 5, 107

[Scherhag *et al.* 2018] Ulrich Scherhag, Dhanesh Budhrani, Marta Gomez-Barrero and Christoph Busch. *Detecting Morphed Face Images Using Facial Landmarks.* In Alamin Mansouri, Abderrahim Elmoataz, Fathallah Nouboud and Driss Mammass, editeurs, Image and Signal Processing - 8th International Conference, ICISP 2018, Cherbourg, France, July 2-4, 2018, Proceedings, volume 10884 of *Lecture Notes in Computer Science*, pages 444–452. Springer, 2018. 105, 107

[Scherhag *et al.* 2019] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt and Christoph Busch. *Face Recognition Systems Under Morphing*

Attacks: A Survey. IEEE Access, vol. 7, pages 23012–23026, 2019. 107

[Schroff *et al.* 2015] Florian Schroff, Dmitry Kalenichenko and James Philbin. *FaceNet: A unified embedding for face recognition and clustering.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 815–823. IEEE Computer Society, 2015. vii, 3, 36, 74

[Seibold *et al.* 2018] Clemens Seibold, Anna Hilsmann and Peter Eisert. *Reflection Analysis for Face Morphing Attack Detection.* In 26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018, pages 1022–1026. IEEE, 2018. 105, 107

[Sengupta *et al.* 2016] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo Castillo, Vishal M. Patel, Rama Chellappa and David W. Jacobs. *Frontal to profile face verification in the wild.* In 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016, pages 1–9. IEEE Computer Society, 2016. viii, 31, 97, 101

[Sáez Trigueros *et al.* 2021] Daniel Sáez Trigueros, Li Meng and Margaret Hartnett. *Generating photo-realistic training data to improve face recognition accuracy.* Neural Networks, vol. 134, pages 86 – 94, 2021. vii, 24, 32, 33, 34, 36, 77

[Shen *et al.* 2018] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang and Xiaoou Tang. *FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis.* In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 821–830. IEEE Computer Society, 2018. 24, 36, 85

[Shen *et al.* 2020] Yujun Shen, Jinjin Gu, Xiaoou Tang and Bolei Zhou. *Interpreting the Latent Space of GANs for Semantic Face Editing.* In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 9240–9249. IEEE, 2020. xiii, xiv, 55, 58, 65, 68, 77, 78, 79, 80

[Simonyan & Zisserman 2015] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. In Yoshua Bengio and Yann LeCun, editeurs, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 111

[Song *et al.* 2018] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun and Tieniu Tan. *Geometry Guided Adversarial Facial Expression Synthesis*. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart and Tao Mei, editeurs, 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, pages 627–635. ACM, 2018. 24, 32, 33

[Sun *et al.* 2018a] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele and Mario Fritz. *Natural and Effective Obfuscation by Head Inpainting*. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 5050–5059. IEEE Computer Society, 2018. 44

[Sun *et al.* 2018b] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt and Bernt Schiele. *A Hybrid Model for Identity Obfuscation by Face Replacement*. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss, editeurs, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I, volume 11205 of *Lecture Notes in Computer Science*, pages 570–586. Springer, 2018. 44

[Taigman *et al.* 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato and Lior Wolf. *DeepFace: Closing the Gap to Human-Level Performance in Face Verification*. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pages 1701–1708. IEEE Computer Society, 2014. 1

[Tran & Liu 2018] Luan Tran and Xiaoming Liu. *Nonlinear 3D Face Morphable Model*. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 7346–7355. IEEE Computer Society, 2018. 87, 96, 125

[Tran *et al.* 2019] Luan Tran, Xi Yin and Xiaoming Liu. *Representation Learning by Rotating Your Faces*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 12, pages 3007–3021, 2019. vii, viii, 24, 32, 33, 34, 37, 57, 77

[Turk & Pentland 1991] Matthew Turk and Alex Pentland. *Eigenfaces for Recognition*. Journal of Cognitive Neuroscience, vol. 3, no. 1, pages 71–86, 1991. PMID: 23964806. 2

[Venkatesh *et al.* 2020] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran B. Raja, Naser Damer and Christoph Busch. *Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection*. In 8th International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020, pages 1–6. IEEE, 2020. 7, 106, 110

[Vincent *et al.* 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio and Pierre-Antoine Manzagol. *Extracting and composing robust features with denoising autoencoders*. In William W. Cohen, Andrew McCallum and Sam T. Roweis, editeurs, Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, volume 307 of *ACM International Conference Proceeding Series*, pages 1096–1103. ACM, 2008. 11

[Wang *et al.* 2003] Z. Wang, E. P. Simoncelli and A. C. Bovik. *Multiscale structural similarity for image quality assessment*. In The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003, volume 2, pages 1398–1402 Vol.2, 2003. 111

[Wang *et al.* 2018] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian and Chen Change Loy. *The Devil of Face Recognition Is in the*

Noise. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss, editeurs, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX, volume 11213 of *Lecture Notes in Computer Science*, pages 780–795. Springer, 2018. 2

[Webster *et al.* 2019] Ryan Webster, Julien Rabin, Loïc Simon and Frédéric Jurie. *Detecting Overfitting of Deep Generative Networks via Latent Recovery*. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 11273–11282. Computer Vision Foundation / IEEE, 2019. 43, 109

[Weinberger & Saul 2009] Kilian Q. Weinberger and Lawrence K. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification*. J. Mach. Learn. Res., vol. 10, pages 207–244, 2009. 74

[Wu *et al.* 2018] Xiang Wu, Ran He, Zhenan Sun and Tieniu Tan. *A Light CNN for Deep Face Representation With Noisy Labels*. IEEE Trans. Inf. Forensics Secur., vol. 13, no. 11, pages 2884–2896, 2018. 33

[Wu *et al.* 2019] Yifan Wu, Fan Yang, Yong Xu and Haibin Ling. *Privacy-Protective-GAN for Privacy Preserving Face De-Identification*. J. Comput. Sci. Technol., vol. 34, no. 1, pages 47–60, 2019. 44

[Yi *et al.* 2014] Dong Yi, Zhen Lei, Shengcai Liao and Stan Z. Li. *Learning Face Representation from Scratch*. CoRR, vol. abs/1411.7923, 2014. 97

[Yin *et al.* 2017] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu and Manmohan Chandraker. *Towards Large-Pose Face Frontalization in the Wild*. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 4010–4019. IEEE Computer Society, 2017. 24, 32, 33, 34, 36

[Zhang *et al.* 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman and Oliver Wang. *The Unreasonable Effectiveness of Deep Features as a*

Perceptual Metric. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 586–595. IEEE Computer Society, 2018. 81

[Zhao *et al.* 2017] 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 Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, editeurs, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 66–76, 2017. 24, 29, 30, 31, 85, 88

[Zheng & Deng 2018] T. Zheng and W. Deng. *Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments*. Rapport technique 18-01, Beijing University of Posts and Telecommunications, February 2018. viii, 97, 101

[Zheng *et al.* 2017] Zhedong Zheng, Liang Zheng and Yi Yang. *Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro*. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 3774–3782. IEEE Computer Society, 2017. 45, 46

[Zhu *et al.* 2015] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi and Stan Z. Li. *High-fidelity Pose and Expression Normalization for face recognition in the wild*. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 787–796. IEEE Computer Society, 2015. xii, 24, 25, 26, 27, 28, 29, 85

[Zhu *et al.* 2016] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi and Stan Z. Li. *Face Alignment Across Large Poses: A 3D Solution*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las

Vegas, NV, USA, June 27-30, 2016, pages 146–155. IEEE Computer Society, 2016. 37

[Zhu *et al.* 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A. Efros. *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.* In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 2242–2251. IEEE Computer Society, 2017. 31, 57