# Reconstruction meets Recognition: Exploring Face Identity Embeddings

Mingrui Li

*Doctor of Philosophy*

University of York

Computer Science

April, 2024

# Abstract

Embedding a face image into a descriptor vector using a deep neural network is a standard technique in face recognition. These embeddings are designed to capture only identity (ID) information. Their effectiveness is assessed by evaluating recognition performance against datasets with diverse non-identity (non-ID) factors. This thesis examines the data within face embeddings and whether faces can be reconstructed from these embeddings using a Generative Adversarial Network (GAN) or a 3D Morphable Model (3DMM).

The first contribution of this work involves studying the ID and non-ID information encoded within face embeddings. Ideally, environmental data like background and lighting, along with variable facial aspects such as pose and accessories, should be ignored in recognition. However, we find that attributes, landmark positions, and image histograms can be retrieved from ID embeddings of networks like VGGFace2 and ArcFace. Building on this, we propose an adversarial training method that more effectively excludes non-ID information by deploying a novel network architecture that selectively penalizes non-ID attribute encoding, enhancing recognition performance.

The second contribution focuses on reconstructing 2D facial images directly from embeddings. Our study reveals that these reconstructed images contain not only identity but also certain non-ID characteristics such as pose, lighting, and background. These findings highlight the privacy risks inherent in facial recognition technologies.

The final contribution shifts focus to the potential reconstruction of 3D face geometry solely from recognition signals. By integrating a 3DMM with a Spatial Transformer Network, we demonstrate that the localiser network can learn 3D shape and pose parameters from identity signals in a warped UV space without additional geometric labels. Our findings confirm that face representations retain spatial information for 3D geometry. This potentially points to future face recognition architectures where most of the model capacity is used for alignment, leaving a relatively simple feature extraction and recognition problem on the aligned face image.

# Acknowledgement

I am deeply grateful for the guidance and support I received from my supervisors, Professor William A. P. Smith and Dr. Patrik Huber. Their expertise, insights, and especially their care during the challenging times of COVID-19 were invaluable in helping me navigate through the most difficult periods of my research journey. I would also like to extend my heartfelt thanks to my internal assessor, Professor Nick Pears, for his thoughtful and caring supervision of both my academic progress and personal well-being.

I am fortunate to have been surrounded by wonderful colleagues and friends who have made my PhD journey enjoyable and enriching. Special thanks to Dr Hao Sun, Dr Jie Zou, Yajie Gu, Yao Chen, Dr Zongyu Yin, Dr Jianjia Wang and Dr Dizhong Zhu for their companionship, encouragement, and shared wisdom.

Lastly, I would like to express my profound gratitude to my grandparents, whose love and support have been my constant source of strength and motivation. This accomplishment is as much theirs as it is mine.

# Declaration

I declare that this thesis is a presentation of original work and I am the principal author. This work has not previously been presented for an award at this or any other university. The main content has been published in the following paper, for which I made significant contributions in terms of design, implementation, experimentation, and writing. All sources are acknowledged in the References section.

1. Li, M., Smith, W. A., & Huber, P. (2023, April). ID2image: Leakage of non-ID information into face descriptors and inversion from descriptors to images. In Scandinavian Conference on Image Analysis (pp. 432-448). Cham: Springer Nature Switzerland.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

This thesis is concerned with face recognition and face reconstruction, two fundamental and well-studied problems in computer vision.

*Face recognition* has been studied for over five decades [56]. The problem is challenging due to: 1. the small differences that distinguish different identities, 2. the dramatic changes in appearance due to illumination, pose, background, clothing and camera properties, 3. the fact that the face itself changes with age and expression. After four decades of relatively little success on this task, a step change in performance was obtained through the rise of deep convolutional neural networks [64] and their application to face recognition [114]. The most successful paradigm for face recognition is to train a network that embeds a face image into a low-dimensional vector that depends only on the identity (ID) of the face. Faces can be compared by the similarity of their embeddings. Such approaches rely on having sufficiently diverse training data that they can learn to discard all non-identity (non-ID) factors in the embedding process. Hence, the goal is to learn to transform an image to a representation that is invariant to everything but ID.

*Face reconstruction* on the other hand aims to explain an image in terms of parameters that encode both identity and non-identity factors via a generative model. Sometimes ID and non-ID factors are disentangled. For example, a 3D Morphable Model (3DMM) explicitly models ID (in the form of 3D

geometry and texture), expression, lighting and pose in order to optimally reconstruct a given image. Classically, this was done via analysis-by-synthesis [14] though here too, CNN-based direct regression methods have dramatically improved performance [118, 39, 116]. In other cases, for example in a 2D Generative Adversarial Network (GAN) [57], both ID and non-ID factors are modelled together in a black box latent space.

In this thesis, we argue that face reconstruction and recognition are intimately related and that the relationship between them has been little explored. The theme of the thesis is the interplay between the two, particularly how advancements in one area can enhance the other. For instance, using reconstruction to demonstrate the challenges of removing non-ID factors in recognition, enhancing recognition by reducing these elements, and learning reconstruction techniques from recognition by disentangling 3D pose/shape from the feature extraction process.

## 1.1 Identity and Non-Identity Factor

Distinguishing between ID factors and non-ID factors is essential in face recognition and reconstruction. ID factors refer to intrinsic characteristics unique to individual identity, such as facial shape, skin texture, and specific facial features [86]. In contrast, non-ID factors refer to variations caused by external influences or environmental conditions, such as pose, lighting, expression, accessories (*e.g.,* glasses or hats), and even background or camera settings [17]. These non-ID factors can vary significantly, altering the appearance of a face without affecting the identity of the individual.

Certain factors may even contain elements of both ID and non-ID information. For example, facial landmarks encode intrinsic shape information

(an ID factor) but also change with pose (a non-ID factor). Effective face recognition systems, therefore, aim to be invariant to non-ID factors while maintaining high sensitivity to ID factors to ensure that only identity information is captured in the embedding [28]. However, fully achieving this balance remains challenging due to the entanglement of certain ID and non-ID attributes [36].

## 1.2 Exploration of Face Embedding

In recent years the hidden representations of CNNs, have dominated face recognition. These CNNs are structured to encode facial data into a compact, lower-dimensional space known as an embedding [17, 97, 86]. A measure of distance between face embedding is used to represent dissimilarity in identity. CNNs are structured hierarchically, where layers of abstraction progressively refine raw input data into identity-specific representations. The goal of training such networks is to minimise the within-class scatter while maximising the between-class scatter for all identities. The former goal requires that the face representation should depend only on the identity of the person in the image. Environmental conditions such as the lighting, background, and properties of the camera as well as changeable aspects of the face such as pose, expression and the presence of accessories should not affect the face representation (i.e. should not introduce within-class scatter). In other words, the embedding network should learn invariance to these factors.

To achieve this invariance, the engineering of training datasets, network architectures, and loss functions has been extensively explored in face recognition [17, 43, 28, 102]. Datasets are designed to expose a wide variety of non-ID factors, and loss functions are developed to ensure that the same

identity embeds to the same point across different conditions, facilitating invariance learning. However, despite these efforts, studies have shown that face recognition networks often encode soft biometrics such as age, race, and gender during training, posing significant privacy risks [36, 30, 15]. This raises substantial concerns regarding the unauthorized extraction of individual information. Additionally, recent research suggests that the removal of such soft biometric data can potentially improve the performance of facial recognition systems [81, 31, 40].

## 1.3 Leakage of Non-ID information

In this thesis, we investigate to what extent this has been achieved. In particular, we ask: how well do modern face embedding networks successfully remove non-ID factors when embedding a face image? Moreover, we propose innovative solutions aimed at mitigating these privacy risks. This investigation is segmented into two chapters, each addressing different facets of the problem and contributing to a comprehensive discussion on improving the security and efficacy of face recognition technology.

Our first study comprises two sections. Initially, we critically examine the nature of ID embeddings in state-of-the-art networks like VGGFace2 [17] and ArcFace [28], questioning whether these embeddings solely contain identity-related information. Our findings reveal that non-ID attributes such as landmark positions and image histograms can indeed be predicted from these ID embeddings, with accuracy comparable to predictions made from the original images. Additionally, we introduce a novel adversarial training approach within the facial recognition network aimed at penalizing the inclusion of non-ID information in ID embeddings. By doing so, we seek to purify the

embeddings of unnecessary non-ID data, thereby enhancing network performance. Our results demonstrate that eliminating unnecessary non-identity information can improve the performance of facial recognition networks to a certain extent.

In Chapter 4, we present an optimization strategy utilizing a generative model (specifically, StyleGAN2 for faces) to recover images from an ID embedding. We achieve photorealistic inversion from ID embedding to face image, where not only is the ID realistically reconstructed, but also the pose, lighting, and background/apparel to some extent. The successful visualization of non-ID factors from face embeddings provides crucial insights into privacy concerns and opens new possibilities for enhancing security measures in biometric systems.

## 1.4 Spatial information in face representation

Recent studies have primarily focused on improving the robustness of face recognition against various transformations. While CNNs inherently support a degree of spatial invariance through mechanisms like shared weights and localized receptive fields, they do not fully achieve spatial neutrality. The networks generally maintain sufficient spatial information to facilitate accurate recognition,effectively handling minor translations and shifts in position. However, the exploration of the extent of spatial invariance within face representations remains relatively underexplored. Additionally, the opaque nature of CNNs end-to-end training processes makes face representations akin to 'black boxes'—highly abstract and complex, thereby complicating further analysis.

Building on this foundational understanding, we propose a novel approach

in Chapter 5 that investigates the extent of spatial information retained in face embeddings. This method learns 3D face alignment and reconstruction using only the facial identity information captured through a recognition loss. It integrates a 3DMM within a Spatial Transformer Network (STN) framework. Initially, the model estimates the 3D shape and orientation of a face, which it then projects onto a UV texture map. This map serves as the input to our face recognition network, ensuring that the extracted facial representation is inherently spatially invariant. We show for the first time that 3D face alignment (i.e. estimation of pose and shape parameters) can be learnt using a recognition signal alone.

## 1.5 Outline

The structure of the remainder of this thesis is as follows:

1. Chapter 2 reviews all of the essential basic concepts for the theoretical groundings of all three works described by this thesis. Specifically, we review related work in the areas of generative face modelling, face recognition, face reconstruction, geometric alignment, face editing and privacy implications.

2. Chapter 3 presents our first significant finding: identity embeddings from state-of-the-art face embedding networks contain not only identity information but also non-ID attributes. Additionally, we incorporate adversarial training to eliminate non-ID information from the face embeddings. Our results suggest that removing non-identity information can improve facial recognition capabilities.

3. Chapter 4 details our second contribution: building on the findings from

the Chapter 3, we introduce an optimization approach using a generative model—specifically, StyleGAN2—to recover not only the identity but also non-ID attributes such as pose, expression, and to some extent, the background/lighting from these identity embeddings.

4. Chapter 5 describes our third contribution: we introduce an innovative approach to face recognition that involves reconstructing 3D facial geometry using only identity signals, without the need for additional geometric labels.

5. Chapter 6 concludes our contributions and discusses prospects for future work.

# 2 Literature Review

This chapter will cover all the related work for our three contributions. We start with fundamental ideas and methods in each section, then introduce state-of-the-art methods.

## 2.1 Generative Face Model

Generative face models represent a rapidly advancing field within facial recognition technologies. These models learn to identify and extrapolate the patterns and features present in a given distribution of facial data samples. Such capability allows the models to generate novel, realistic facial images that were not present in the original datasets. The study of generative models is critically important, as these can be optimized efficiently through scaling the training processes and datasets to produce higher quality and more convincing facial images. In this section, we introduce the foundational concepts and review related work that is crucial for understanding the research problems addressed in this thesis.

### 2.1.1 3D Morphable Model (3DMM)

The most commonly used 3D face model is the 3D Morphable Model (3DMM) proposed by Blanz and Vetter [14]. 3DMM is a statistical model with intrinsic

components of a face (*e.g.,* albedo, texture, and shape). It shows that it is possible to reconstruct facial features by solving a non-linear optimization problem that is constrained by linear statistical models of facial texture and shape.

In 3DMM, the face is in a dense point-to-point correspondence which allows the linear combinations of faces may produce more photo-realistic faces. The visual appearance of 3D faces and scenes is a function of their shape $s$ and surface texture $t$ properties referred to as albedo. The face model in 3DMM can be represented in terms of shape and texture as a pair of vectors:

$$\begin{aligned} s &= (x_1, y_1.z_1, ..., x_n, y_n, z_n)^T \\ t &= (r_1, g_1.b_1, ..., r_n, g_n, b_n)^T \end{aligned} \tag{2.1}$$

where $n$ is a number of vertices. The general approach of 3DMM to construct more statistical models is to remove redundancy by means of Principal Component Analysis (PCA). The key equation can be represented as:

$$\begin{aligned} s &= s_0 + S_\alpha \\ t &= t_0 + T_\beta \end{aligned} \tag{2.2}$$

where $s_0$ and $t_0$ are the mean shape and the mean texture, respectively. $\alpha$ and $\beta$ are the shape coefficients and texture coefficients, respectively. By changing $\alpha$ and $\beta$ we can generate, or morph new faces, as seen in Fig. 2.1.

3DMM could handle the ill-posed reconstruction problem and produce acceptable results. However, the reconstruction result lacks fine-scale skin detail, *e.g.,* freckles, wrinkles. Moreover, it is hard to guarantee that the produced result is physical plausibility. Therefore, many recent works try to

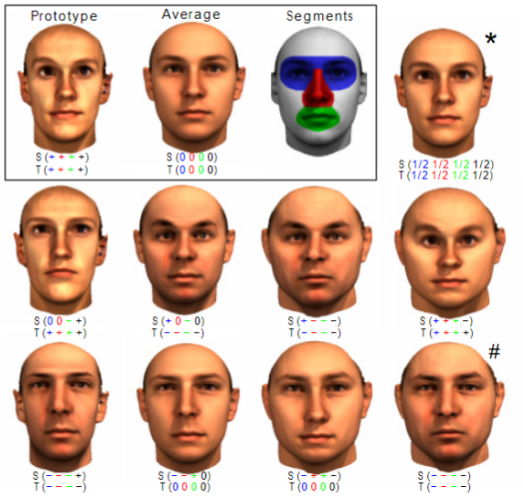



Figure 2.1. The expressiveness of the 3DMM. The deviation of a prototype from the average is added (+) or subtracted (−) from the average. Adding and subtracting deviations independently for shape $S$ and texture $T$ on each of the four segments produces several distinct faces. Taken from [14].

go beyond this coarse reconstruction to personalize the model further and recover the missing dimensions. In the following, we will give a detailed overview of these approaches.

### 2.1.2 Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GANs) are a relatively new concept in Machine Learning, introduced by Goodfellow *et al.* [41]. Their goal is to synthesize artificial samples, such as images, that are indistinguishable from authentic images. The resolution and quality of images produced by generative methods, especially GANs, are improving rapidly. While GANs images became more realistic over time, one of their main challenges is controlling their output, *i.e.,* changing specific features such as pose, face shape, and hairstyle in an image of a face.

GANs consist of two main components: a generator $G$ and a discrimi-

nator $D$. These two components play a minimax two-player game with the following value function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (2.3)$$

Here, $x$ represents real data samples, and $z$ is a noise vector sampled from a probability distribution $p_z(z)$. The generator $G$ attempts to generate data that are indistinguishable from real data by learning to map the noise vector $z$ to the data space. On the other hand, the discriminator $D$ aims to distinguish between samples provided by the generator and real data samples. The discriminator maximizes its ability to identify real and generated samples, while the generator tries to minimize the probability that the discriminator makes the correct classifications (see Fig. 2.2).

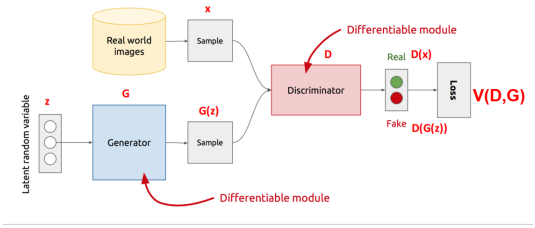



Figure 2.2. Generative Adversarial Network framework.

Generative Adversarial Networks (GANs) have emerged as a powerful tool in computer vision and machine learning, particularly for synthesizing photo-realistic images. Among the notable advancements, NVIDIA's introduction of the Style-Based Generator Architecture for GANs, known as StyleGAN, and its subsequent iteration, StyleGAN2, by Karras *et al.* in 2019 and 2020 [57, 58], respectively, have set new benchmarks in data-driven unconditional generative image modeling. These models are capable of generating high-resolution, lifelike face images, showcasing state-of-the-art results. A

distinctive feature of StyleGAN lies in its generator architecture, which innovatively uses an input latent code $z \in Z$, not just at the network's inception but transforms it through a mapping network into an intermediate latent code $w \in W$. This code then undergoes affine transformations to produce styles that meticulously control the synthesis network's layers, enhancing the model's ability to create novel images with remarkable detail and variation. Additionally, the incorporation of stochastic variation through random noise maps further enriches the synthesized images realism and diversity. Fig. 2.3 shows a set of novel images generated from the StyleGAN.

Figure 2.3. A set of images produced by StyleGAN generator. StyleGAN uses baseline progressive GAN architecture which means the size of generated image increases gradually from a very low resolution ($4 \times 4$) to high resolution ($1024 \times 1024$). Taken from [57]

StyleGAN introduces a progressive approach to image generation, starting from a very low resolution and increasing to high resolutions (*e.g.,* $1024 \times 1024$ pixels). By separately modifying the input at each resolution level, the generator finely tunes visual features ranging from coarse aspects like pose and face shape to finer details such as hair colour, without affecting other

levels.

One of the critical innovations in StyleGAN is its Mapping Network, which transforms the initial latent vector $z \in \mathcal{Z}$ into the intermediate vector $w \in \mathcal{W}$. This mechanism allows StyleGAN to produce images that more closely mimic the training data set, as it facilitates more precise control over the visual features by the input vector. The transformation output $w$ is then passed through a learned affine transformation $A$, utilized via adaptive instance normalization (AdaIN), to generate *styles* that control the stylistic elements of the generated image. The Fig. 2.4. summarizes the StyleGAN generator structure.

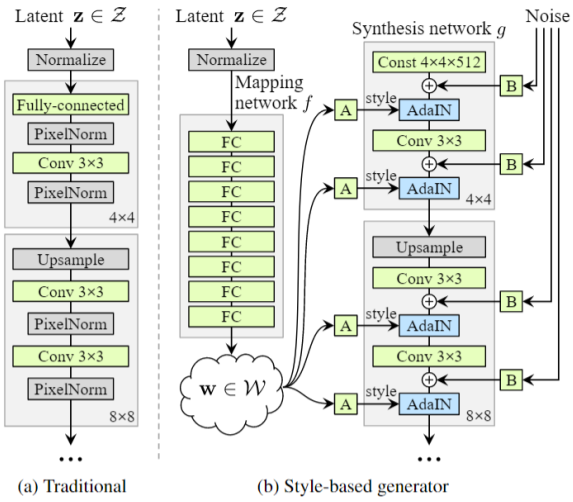



Figure 2.4. The framework of StyleGAN generator. The input to the *AdaIN* is generated by applying $A$ to $w$. Taken from [57].

**GAN inversion** Recent work has shown that GANs can encode a rich set of semantics in their latent space. Beyond image generation, recently, attempts have been made to invert the GAN generating process from the image back to latent space for the purpose of image manipulation or analysis, which is widely known as GAN inversion. Predominantly, existing methods

employ one of three approaches: they either develop an additional encoder or regressor external to the GAN structure [9, 94], directly optimize the latent code to align with a specific target image [126], or utilize a hybrid strategy that begins with an initial optimization seeded by the outputs of a regressor [141].

For StyleGAN specifically, recently, several works have shown that it is possible to retrieve the latent code $w$ of a target image [101, 117]. The works show that inverting to the latent space $W$ is easier than to $Z$. However, accurately reconstructing a target image is still an ongoing challenge. In another recent work, Abdal *et al.* [2, 3] proposed a framework that projects an image into an extended latent space $W+$, where $W+$ contains separates latent vectors for the specific scales of StyleGAN. This approach effectively allows for the reconstruction of features at multiple levels of the target image. Additionally, Yin *et al.* [135] introduced a deep-inversion method that employs a pretrained neural network to generate synthesized class-conditional input images, facilitating data-free knowledge transfer.

### 2.1.3 Hybrids

Due to the complexity and high dimensionality of human face images, modelling faces accurately using solely the 3DMM or GANs presents significant challenges. The 3DMM approach offers the advantage of generating realistic morphological face structures that adhere to the distribution of actual human faces. In contrast, GANs excel in creating detailed texture models within a high-resolution UV space. Recent efforts aim to integrate these two techniques to develop a hybrid model that yields more convincing and realistic facial representations.

Shammai *et al.* [100] presents a facial synthesis approach that begins with

transforming facial scans into 2D images to create new textures via a GAN. These textures are then correlated with 3D facial geometries using 3DMM, providing a cohesive method for generating realistic facial textures and their corresponding geometries. Gecer *et al.* [38] propose GANFIT, which utilizes a GAN to enhance the texture representation in 3DMM, enabling high-fidelity 3D face reconstruction from single images. This method has excellent results in photorealistic 3D face reconstructions and achieves facial texture reconstruction with high-frequency details for the first time.

This approach complements other influential models like FaceVerse [124], which employs a coarse-to-fine strategy merging PCA-based shape and texture models with conditional StyleGAN networks for detailed facial feature synthesis from hybrid datasets. The Next3D framework by Galanakis *et al.* [105], utilizing neural texture rasterization in a 3D GAN framework, synthesizes multi-view consistent, high-fidelity facial images from unstructured single-view 2D imagery, integrating 3DMM for fine-grained control over facial attributes.

The Exp-GAN [67] and cGOF [106] models exemplify the effectiveness of hybrid approaches in providing robust and detailed control over various facial attributes, including identity, expression, pose, and illumination. Recently, Rai *et al.* [91] introduced AlbedoGAN, which leverages StyleGAN2 to produce high-quality albedo textures and precise 3D facial shapes. This method combines 2D face generation with semantic face manipulation to allow direct control of facial expressions in 3D through latent space exploitation, enabling text-based editing of 3D faces. These models excel at disentangling these attributes, significantly enhancing the realism and diversity of the generated facial images.

## 2.2 Face recognition

Face recognition has emerged as a key area of research within artificial intelligence and computer vision, boasting considerable advancements over the past four decades. Initially, the field faced significant hurdles due to limited computing resources and rudimentary algorithms, which impeded the achievement of high accuracy [122, 73, 25, 1]. The emergence of deep convolutional neural networks (DCNNs) marked a turning point, substantially mitigating these early [64, 102, 114, 28, 108]. Today, DCNNs facilitate the creation of highly accurate and reliable face recognition systems. This section will begin with an overview of traditional methodologies in face recognition, transitioning to an in-depth discussion of contemporary DCNNs. We will explore pivotal contributions to the development of feature extraction techniques and face recognition, including influential network architectures, loss functions, and datasets that have defined modern face recognition systems.

### 2.2.1 Pre-deep learning methods

Face recognition techniques have advanced remarkably since the early 1990s, transitioning from basic algorithms to sophisticated image analysis methods that map facial features to a lower dimensional subspace. Initially, holistic or appearance-based approaches treated the entire face region using both linear and non-linear methods.

Principal Component Analysis (PCA), known as eigenfaces [122], linear discriminative analysis, known as fisherfaces [11] and independent component analysis (ICA) [24] represent the most commonly used linear techniques in facial recognition systems. The eigenface technique, a landmark in face recognition, describes each image as a vector of weights derived by project-

ing the image onto a set of principal components, enabling the system to identify the nearest face class by comparing distances within this face space. However, these linear holistic algorithms often underperform when the input data lacks a linear structure. This is particularly relevant for human facial features, which vary widely due to factors such as facial expressions, self-occlusion, lighting conditions, and accessories like glasses or hats. These variations introduce non-linearities and complex spatial relationships that linear methods may fail to capture effectively. Consequently, Support Vector Machines (SVMs) emerged as non-linear holistic approaches that differentiate faces more effectively by finding an optimal hyperplane [25]. By the 2000s, local feature-based face recognition methods gained popularity. These methods utilize hand-crafted features to describe the face, such as Gabor filters [73]and Local Binary Patterns (LBP) [4], along with their advanced extensions [136, 21], demonstrating robust performance due to their invariant properties. Subsequently, the face recognition community began integrating learning-based local descriptors that learn discriminant image filters using shallow techniques [1, 68]. Unlike earlier methods that relied heavily on predefined features, these learning-based descriptors adaptively adjust their parameters to maximize discriminative power, making them more resilient to variations in facial expressions, lighting, and other environmental changes. This progression marked a pivotal shift towards more adaptive methodologies, setting the stage for the next wave of innovation with deep learning approaches.

### 2.2.2 Deep learning based methods

The evolution of face recognition has been significantly accelerated by the introduction of DCNNs, especially after the landmark success of AlexNet in

the 2012 ImageNet competition [64]. The utilization of DCNNs for face representation through embeddings has emerged as the predominant method in contemporary face recognition systems. Deep learning methods use a cascade of multiple layers of processing units for feature extraction and transformation. These methods learn multiple levels of representations corresponding to varying degrees of abstraction, creating a hierarchical structure of concepts. This structure exhibits strong invariance to changes in face pose, lighting, and expression, enabling CNNs to extract more effective features for distinguishing between different faces.

A number of successful systems, such as DeepFace [114], FaceNet [97], VGGFace [17, 86] have demonstrated impressive performance in face identification and verification. Notably, DeepFace [114] marked a significant milestone by nearly reaching human-level performance under unconstrained conditions [49]. Subsequent innovations, including the use of the ResNet architecture and its variations [137, 92, 76, 45], have continued to refine the accuracy and efficiency of face recognition systems. Liu *et al.* [74] introduced a novel approach known as Deep Hypersphere Embedding (SphereFace). In particular, they proposed an angular Softmax loss (A-Softmax), which allows deep CNN to learn discriminative facial features with the angular margin by imposing constraints on a hypersphere manifold. IN 2019, Deng *et al.* [28] presented an additive angular margin loss (ArcFace) achieved 99.83%verification performance on LFW with ResNet-100 architecture, and MS-Celeb-1M training dataset [43].

**Network Architecture** In this section, we focus on three widely recognized CNN architectures—AlexNet, VGGNet, and ResNet—that have significantly advanced face recognition technologies. These networks employ

convolutional layers, pooling layers, and non-linear activation functions to progressively transform raw image data into highly abstract representations. AlexNet [64] marked a pivotal shift in deep learning by introducing deeper architectures with five convolutional layers and three fully connected layers. It leveraged the ReLU activation function for non-linearity and adopted dropout to combat overfitting, demonstrating the feasibility of training large-scale networks on GPUs and leading to significant improvements in recognition tasks. Building on this, VGGNet [102] refined deep networks by consistently employing small $3 \times 3$ convolutional filters across its layers, enabling a gradual increase in depth and feature map dimensions. This uniform design validated the effectiveness of deeper architectures in improving large-scale recognition accuracy. ResNet [46] addressed the challenges of training very deep networks through its introduction of residual blocks with skip connections, which effectively mitigated vanishing gradient problems and enabled the learning of complex patterns in extremely deep models. These architectures collectively laid the foundation for modern face recognition systems by enabling robust and scalable hierarchical feature extraction.

**Learning Metrics** The choice of loss functions plays a pivotal role in optimizing the performance of face recognition systems. These functions can be broadly categorized into classification-based and distance-based approaches. Classification-based losses, such as softmax loss and its variants, focus on maximizing inter-class dispersion. In contrast, distance-based approaches like contrastive and triplet losses aim to optimize the feature space by enhancing discrimination between different classes, effectively penalizing dissimilarities.

**Classification-based loss function** The softmax loss [110] is commonly

used as the supervision signal in object recognition, simplifying the output layer into a multi-class classifier where the cross-entropy between the predicted and actual distributions is minimized, which can be written as follows:

$$L = -\sum_{i} y_i \log(p_i) \tag{2.4}$$

where $y_i$ are the true labels , and $p_i$ denotes the predicted probabilities. The softmax loss function, while effective at achieving inter-class dispersion, does not inherently promote intra-class compactness. To address this limitation and enhance feature discriminability, several modifications based on softmax loss have been proposed [74, 123, 127, 29, 89]. An extension of softmax loss named center-loss [127] attempted to achieve the missing intra-class compactness by taking into account the Euclidean distance between the feature vector and the centre of the class.

Further enhancing the traditional softmax loss, angular or cosine-margin-based losses introduce an angular margin to improve the discriminability of the learned features. This method adjusts the decision boundary in the angular space, ensuring that the learned embeddings not only separate classes effectively but also do so with a margin that enhances generalization to new examples. Several angular margin-based losses, such as SphereFace [74], AM-softmax [113], CosFace [123] progressively improve the performance on various benchmarks to the newer level. In 2019, Deng *et al.* [28] introduce an additive angular margin loss (ArcFace) achieving a considerable improvement on LFW with 99.83% accuracy. their mathematical equation is as follows:

$$L = -\log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j \neq y_i} e^{s\cos\theta_j}} \tag{2.5}$$

Where $\theta_{y_i}$ is the angle for the correct class$m$ is the margin added, and $s$ is

a scaling parameter.

**Distance-based approaches** Distance-based approaches represent an alternative optimization strategy distinct from softmax loss and its variants. In metric learning, the network is trained using sample images and penalized based on whether the samples belong to the same class or different classes. Popular methods include contrastive loss and triplet loss.

The contrastive loss [108, 107, 111, 112, 134] initially proposed by Hadsell *et al.* [44] requires pairs of face images, pulling together positive pairs and pushing apart negative pairs. It has been utilized extensively in face recognition systems, notably the DeepID series [107, 111, 108], and other [112, 134]. The contrastive loss can be mathematically defined as follows:

$$L = (1 - y)\frac{1}{2}(D)^2 + y\frac{1}{2}\max(0, m - D)^2 \tag{2.6}$$

Where, $D$ represents the Euclidean distance between the two samples, $y$ is a binary label (0 for similar pairs and 1 for dissimilar pairs), and $m$ is a predefined margin.

Along with FaceNet [97] proposed by Google, Triplet loss [97, 32, 85, 104] was introduced into FR. IN The triplet loss function involved an anchor $a$, a positive example $p$ of the same class and a negative example $n$ of a different class. The objective is to train the CNN such that the distance between the matching pair is minimized, and the distance between the non-matching pair is maximized. The triplet loss function is mathematically defined by:

$$L = \max(D(a, p) - D(a, n) + m, 0) \tag{2.7}$$

Where $D(a, p)$ and $D(a, n)$ represent the distances from the anchor $a$ to the positive example $p$ and the negative example $n$, respectively, with $m$ again

being the margin.

### 2.2.3 Datasets

In the realm of face recognition, datasets serve a dual role as both training material and benchmarks for system validation. The quality of training data significantly influences the performance of deep neural networks, while the quality of validation data impacts the reliability of benchmark results.

Initially, face datasets such as CelebFaces [77] and CASIA-Webface [134] consisted mostly of high-quality images featuring celebrities. These datasets included images captured in less controlled environments, presenting more realistic challenges that systems might face in real-world applications. Modern datasets, such as MS-Celeb-1M [43] and VGGFace2 [17] have expanded vastly, providing vast amounts of data critical for training DNNs with robust face recognition capabilities. For instance, VGG-Face2 [17] includes images from 2,622 identities, each represented by approximately 1,000 samples, covering a diverse spectrum of ethnicities, lighting conditions, poses, and expressions. MS-Celeb-1M [43], currently the largest public face dataset, contains 10 million images of 10,000 celebrities.

For benchmark datasets, LFW [49] is the most typical benchmark for unconstrained face recognition. It comprises 13,233 facial images of 5,749 people under varying conditions of pose, lighting, focus, and resolution, targeting the pair-matching problem/face verification. Following LFW, Youtube face (YTF) [128] serves as another important testing datasets. The dataset comprises 3425 videos of 1595 individuals. YTF evaluates pair matching under two protocols: restricted and unrestricted. The IJB series [62, 79] is known for its stringent testing protocols and includes both images and videos, providing a robust platform for evaluating the performance of face recog-

nition systems across a variety of pose angles and illumination conditions. MegaFace [60] further evaluates face recognition and verification performance against up to 1 million distractors (approximately 672K identities), crucial for applications that require identification from very large populations. A detailed comparison of the datasets discussed in this section is provided in Table 2.1, highlighting their size, the number of identities and key features.

| Dataset | Size | Identities | Key Features |
|---|---|---|---|
| CelebFaces [77] | 202,599 | 10,177 | Unconstrained, celebrity images |
| CASIA-Webface [134] | 494,414 | 10,575 | Realistic challenges, diverse environments |
| VGGFace2 [17] | 3.3M+ | 2,622 | Diverse poses, lighting, and expressions |
| MS-Celeb-1M [43] | 10M+ | 10K | Largest public dataset, celebrity images |
| LFW [49] | 13,233 | 5,749 | Pair-matching, various conditions |
| YTF [128] | 3,425 | 1,595 | Video-based pair matching |
| IJB-C [79] | 31K+ | 3,531 | Benchmarking, varied poses |
| MegaFace [60] | 1M+ | 672K | Large-scale distractors |

Table 2.1. This table compares popular face datasets in terms of size, number of identities, and key features relevant to training and evaluation.

## 2.3 Face reconstruction

This section reviews existing works on 3D face reconstruction from monocular images. Reconstructing a face from a single 2D image presents an ill-posed problem, necessitating the estimation of various parameters including intrinsic and extrinsic camera parameters, lighting conditions, shape, and texture. Despite these challenges, face reconstruction has significant applications in fields such as surveillance, medicine, security, and entertainment.

The classical approach to face reconstruction employs an analysis-by-synthesis architecture. This methodology explains an observed data vector, such as an image, in terms of the hidden features that generated it. In the context of face reconstruction, this involves recovering the facial shape

by comparing a target image with an image produced through computer graphics-based rendering. The specific reconstruction approach discussed below adheres to this analysis-by-synthesis framework.

### 2.3.1 3DMM fitting

A 3DMM-based face reconstruction typically involves two primary stages: model building and model fitting. In the model building stage, a 3D statistical face model is generated using training datasets. During the model fitting stage, the 3D face model is projected onto the face in a given image to facilitate reconstruction.

Zhmoginov *et al.* [140] present a gradient descent approach to invert FaceNet embeddings for image recovery. This process utilizes an autoencoder structure, where the autoencoder is designed to approximate the identity mapping by coupling an encoding stage with a decoding stage to learn a compact intermediate representation, known as the code vector. These architectures have been widely used to extract facial features from images, offering the significant advantage of being generally unsupervised.

Their result is shown in Fig. 2.5, while not perfect, demonstrate that neural network embedding losses, when combined with simple regularization functions, have the potential to reconstruct faces that look realistically human.

In recent years, people have done much work on how to employ DCNNs for face reconstruction. The first such methods are trained differentiable render (i.e. image-to-parameter mapping) to regress 3DMM parameters.

Richardson *et al.* [95] proposed a methodology utilizing an iterative Convolutional Neural Network (CNN) trained on synthetic datasets for estimating parameters of a 3DMM. This initial geometric prediction is further refined


Figure 2.5. Sample result of face reconstruction: (a) generic "male" guiding image; (b) generic "female." guiding image; (c) image used for calculating the target embedding e; (d) reconstruction of e with the guiding image (a); (e) reconstruction of e with the guiding image (b). Taken from [140].

using a real-time shape-from-shading technique. To enhance detail extraction, they later developed an end-to-end CNN framework that incorporates a coarse-to-fine approach, significantly advancing the field of facial geometry reconstruction [96].

Training deep neural networks usually requires a great quantity of data, but face images with 3D ground truth shapes are hardly available. Tewari *et al.* [118] which can be trained on unlabeled photographs to predict shape, expression, texture, pose, and lighting simultaneously. MoFA integrates deep learning-based and model-based capture within an end-to-end trainable architecture, facilitated by a differential renderer that enables unsupervised training on in-the-wild face images.

The framework of MoFA is illustrated in Fig. 2.6, Here, a differential decoder processes the code vector produced by a convolutional encoder network that captures faces, effectively reconstructing the face. This semantic vector stores 3DMM parameters, perspective camera parameters, and spherical harmonics in a unified manner:

$$x = (\alpha, \theta, \beta, T, t, \gamma) \tag{2.8}$$

Where shape $\alpha \in \mathbb{R}^{80}$, facial expression $\gamma \in \mathbb{R}^{80}$,skin reflectance $\beta \in \mathbb{R}^{80}$,

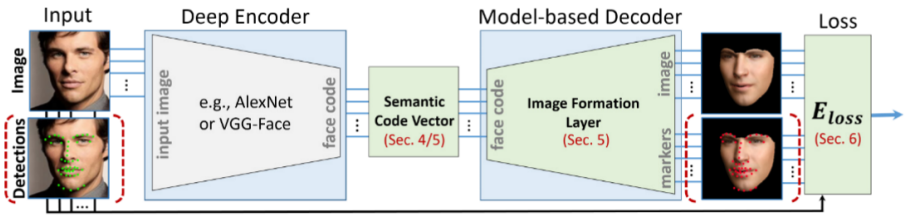



Figure 2.6. The framework of MOFA [118]. A deep model-based face autoencoder enables unsupervised end-to-end learning of semantic parameters. Taken from [118].

camera rotation $T \in SO^{80}$ and translation $t \in \mathbb{R}^3$, and the scene illumination $\gamma \in \mathbb{R}^{27}$.

MoFA employs a robust dense photometric loss function that enables end-to-end training of the encoder. It is possible because the decoder is differentiable, which allows MOFA to compute the gradients of the rendered image with respect to the parameters of the 3DMM model. The loss combines three terms:

$$E_{loss(X)} = w_{land}E_{land}(X) + w_{photo}E_{photo}(X) + w_{reg}E_{reg}(X) \tag{2.9}$$

Here they enforce sparse landmark alignment $E_{land}$, dense photometric alignment $E_{photo}$ and statistical plausibility $E_{reg}$ of the modelled faces.

Genova *et al.* [39] also propose an encoder-decoder architecture for 3D face reconstruction. Their method utilizes a CNN followed by fully connected layers, which effectively map extracted features to 3D Morphable Model (3DMM) parameters. This architecture enables the learning of realistic facial geometries through iterative refinement from a diverse dataset comprising both synthetic and real-world images. A distinctive feature of their approach is the implementation of a batch distribution loss, which is specifically designed to align the distribution of the output with that of the morphable model. This method shows improved resistance to confounding

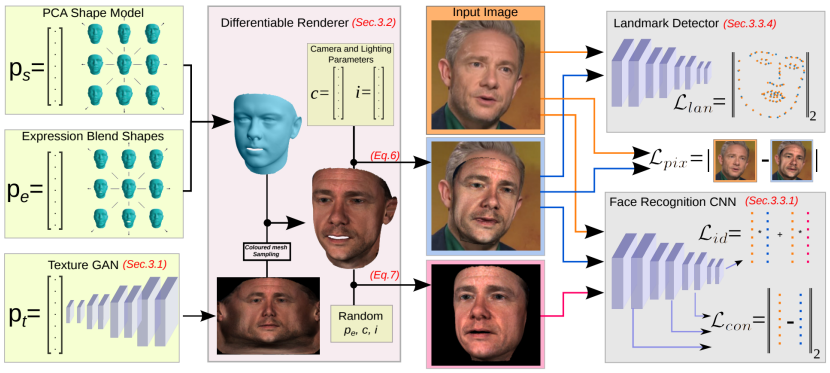



Figure 2.7. Overview of the architecture of GANFIT [38]. GANFIT optimize the parameters with the supervision of pretrained deep identity features through our end-to-end differentiable framework. Taken from [38].

variables such as identity, expression, skin tone, and lighting, compared to MoFA.

Despite these technological advancements, the methods we mentioned above, continue to face challenges in capturing high-frequency details in textures. Moreover, because the reconstruction process permits deviations from the 3DMM space, it remains vulnerable to outliers, such as glasses or earrings, which can be inaccurately represented in shape and texture.

### 2.3.2 Face reconstruction with GANs

Face reconstruction with GANs can utilize the characteristics of GANs to generate high-quality facial texture. In the past few years, some work has explored the use of GANs for face reconstruction [100, 38]. The fundamental concept behind these methods involves employing GANs in conjunction with a differentiable renderer to develop a potent generator of facial textures. This differentiable renderer operates within a model-based encoder-decoder architecture, where the decoder processes the 3DMM parameters predicted by the encoder to reconstruct a 3D face onto a 2D image plane. Addition-

ally, the differentiable render allows the latent vector of GANs to be easily optimized by backpropagation via gradient descents.

In [38], the authors employ an end-to-end differentiable renderer combined with GANs to train a sophisticated generator of facial texture in UV space. This setup integrates 3DMM to determine the optimal latent parameters that not only reconstruct the test image but also adapt it to a new scene. The overall architecture of this system, GANFIT, is illustrated in Fig. 2.7.

Within the GANFIT structure, the reconstruction mesh is formed by a 3D morphable shape model and textured by the generator network's output UV map. A differentiable renderer is used to project the 3D reconstruction onto a 2D image plane using a deferred shading model, with specified camera and illumination parameters. Furthermore, to enhance the robustness of identity-related parameters, the authors render a secondary image with random expression, pose, and illumination [38]. This end-to-end differentiable architecture enables the propagation of loss all the way back to the latent parameters through gradient descent optimization, thereby allowing deep networks to function either as a generator or as part of the cost function. The results of GANFIT (shown in Fig. 2.8) show that GANFIT has higher photorealistic texture reconstructions than has higher photorealistic texture reconstructions than other methods that regress their texture from the MOFA model.

Building upon GANFIT method, Lattas *et al.* [66] introduced AvatarMe, a groundbreaking method designed to reconstruct photorealistic 3D faces from single "in-the-wild" images, incorporating an unprecedented level of detail. Initially, AvatarMe employs a 3D Morphable Model (3DMM) to reconstruct the basic shape and texture of a 3D face from a single image at low resolution. Subsequently, a completed UV texture is synthesized to enhance

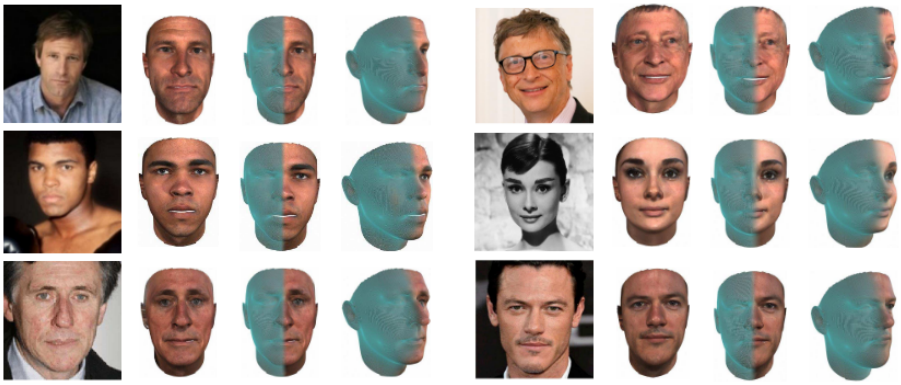


Figure 2.8. Qualitative results of GANFIT for the images from various datasets [38]. It is worth noting that GANFIT is not only good at capturing high-frequency details of the identities but robust to occlusion (*e.g.,* glasses), low resolution and black-white in the photos and generalizes well with ethnicity, gender and age. Taken from [38].

the facial details. The GAN framework aids in the inference of the final material properties such as specular albedo and normals. These elements are essential for simulating realistic light interactions with the facial surface, thereby significantly enhancing the overall photorealism of the reconstructed faces. This process vividly demonstrates the capability of GANs to generate high-fidelity, detailed textures from limited input data, making it integral to the AvatarMe system's performance in rendering photorealistic 3D faces from in-the-wild images.

### 2.3.3 Reconstruction from features

Both the 3DMM-based and GAN-based reconstruction methods are based on analysis-by-synthesis loops of forming a face image and minimizing the difference between the input image and the model appearance. However, recent studies suggest that analysis-by-synthesis fitting algorithms can be computationally burdensome and prone to getting trapped in local minima [143].

Where the reconstruction from features method is opposed to the reconstruction from images. An alternative to this approach is the reconstruction from features method, which contrasts sharply with image-based reconstruction. Instead of using images as input, this method utilizes features extracted from images, such as face ID descriptors, for the fitting task. This approach reduces dependency on the initial quality of reconstruction and increases the efficiency of the fitting process.

Despite its potential, the reconstruction of a face solely from face descriptors remains a relatively unexplored area in research. Only a few studies have investigated the inversion of a face descriptor back into a face image [140, 118, 23, 39, 93].

Huber *et al.* [51] and Zhu *et al.* [143] use local features with regression-based methods to fit a 3D Morphable Model (3DMM) to 2D face images. Zhu *et al.* [143] particularly focus on reconstructing shape from the histogram of oriented gradients (HOG) features around face landmarks. More recent work employs face recognition networks to generate face ID descriptors, which serve as input features for their regression networks, verifying that the output closely resembles the input photograph [90, 116].

Concerning non-ID information contained in face descriptors, early work by Kumar *et al.* [65] found that using an 'inverse crop'—removing the face from an image—still allowed for surprisingly high face recognition rates on LFW. Notably, these inverse crops included hair and parts of ears/chin, contributing to the recognition rates exceeding 99%, even a decade ago. However, it remains largely unexplored whether any background or other non-ID information is present in face descriptors, particularly in modern state-of-the-art networks like VGGFace2. To the best of our knowledge, no work so far has investigated if any non-ID properties can be recovered from iden-

tity descriptors. In Chapter 5, we will introduce our attempt to reconstruct non-ID attributes from face descriptors.

### 2.3.4 Geometric alignment

Face alignment is the process of moving and deforming a face model to an image, so as to extract the semantic meanings of facial pixels. It is an essential preprocessing step for many face analysis tasks, recognition, animation, tracking, attribute classification and image restoration. In the field of computer vision, face alignment remains a critical challenge that has garnered significant interest. Initially, numerous 2D facial alignment techniques were developed to locate fiducial 2D facial landmarks, treating 2D Face Alignment as a regression problem where the landmark locations are directly regressed from the face images [75, 129]. Additionally; CNN-based methods are also largely used on 2D landmark location [87, 16, 72]. Sun *et al.* [109] firstly uses CNN to regress landmark locations with the raw face image. Liang *et al.* [72] improve the flexibility by estimating the landmark response map. Zhang *et al.* [137] further combine face alignment with attribute analysis through multi-task CNN to boost the performance of both tasks. Jourabloo and Liu [55] further propose a CNN architecture that enables the end-to-end training ability of their network cascade to improve its alignment. However, the limitation of 2D landmarks is that they only regress visible points, which fails to adequately describe face shape under large pose variations. For faces with large poses or occlusions, the incorporation of strong 3DMM face shape priors has proven advantageous, which begins with fitting a 3DMM to a 2D image [80, 42, 142]. Recently, regression based 3DMM fitting, which estimates the model parameters by regressing the features at projected 3D landmarks, has looked to improve the efficiency [54, 118, 121]. Nonetheless,

given that 3DMM parameters vary in importance during the fitting process, directly minimizing parameter error may not always yield optimal alignment results, as smaller parameter errors do not necessarily correlate with reduced alignment discrepancies [18].

Jaderberg *et al.* [53] proposed the Spatial Transformer Network (STN), a neural network module that explicitly handles pose and nonrigid deformations in input data, enabling it to handle inputs with significant translation and pose variations more effectively. Bas *et al.* [8] extended this work to align a 3DMM using STN (3DMM-STN). The author uses the localiser to predict 3DMM shape parameters and pose. According to the predicted parameter, the grid generator projects the 3D geometry to 2D. At the same time, an occlusion mask is computed from the estimated 3D geometry. Subsequently, a bilinear sampler resamples the input image onto a regular output grid, which is then masked by the previously computed occlusion mask. This methodology demonstrates robust performance in handling images with significant pose variations. Our work presented in Chapter 5 is based on their architecture, we leverage the STN-3DMM approach to enhance face recognition, focusing on pose-invariance through advanced mapping and occlusion handling, distinctly aiming to achieve efficiency with less data compared to existing methods.

### 2.3.5 Semantic Face Editing

Semantic face editing aims at manipulating the facial attributes of a given image. Different from simply changing the greyscale and other low-complex information of the facial image, manipulating attributes of a face (*e.g.,* changing the pose, expression, ageing or even gender) is a more complex and challenging modification to perform. In this case, in order to obtain realistic

results, a skilled human with image edition software would often be required. In recent years, many learning-based methods are deployed to solve this problem and have achieved certain results.

**Face editing with 3DMMs** The robust statistical priors inherent in 3D Morphable Models (3DMMs) prove highly beneficial for face editing tasks [12, 13]. 3DMMs allows for the parameterization of face manipulations, providing a controlled framework for altering facial attributes. Typically, face editing involves an initial 3D shape reconstruction by fitting the 3DMM to an image within an analysis-by-synthesis loop. Once this model is fitted, prior knowledge about the shape and texture of faces becomes available, enabling substantial and plausible edits to the full 3D shape and texture of a face, even from a single image.

Early research in this area focused on modifying specific attributes such as pose [12, 55, 88], expression [13, 119, 20], and aging [61, 139]. However, these approaches often cater to specific editing tasks and rely on prior knowledge that may not be universally applicable to new editing challenges. This limitation underscores the need for more adaptive and generalized methods that can handle a broader range of editing tasks without extensive retraining or recalibration.

**Face editing with GANs** GANs have shown significant potential in face editing due to their ability to generate plausible and realistic data. However, despite recent advancements in high-fidelity image synthesis using GANs, there remains a limited understanding of how facial semantics are encoded within their latent space. This lack of understanding restricts our ability to manipulate facial attributes effectively using GAN inversion methods.

One popular approach, known as *Invert and Edit*, involves a two-step

process during which an image is first inverted into the latent space of a GAN, and then the resulting latent code is edited in a semantically meaningful way [130]. This method allows for intuitive, semantic editing of the image, as demonstrated by Shen *et al.* [101] in their work on InterFaceGAN. The author finds that for any binary semantic (*e.g.,* male v.s. female, young vs old) there exists a hyperplane in latent space as the separation boundary such that all samples from the same side are with the same attribute. Each hyperplane has a specific unit average vector $n$, the distance between a latent space $z$ to this hyperplane as:

$$d(n, z) = n_T z \tag{2.10}$$

A semantic scoring function $f : X \rightarrow S$ that is a mapping from image space $S$ to semantic space $X$. When a sample semantic lies near the boundary and is moved toward and across the hyperplane, both the "distance" and the semantic score can change accordingly. Therefore, the correspondence between facial semantics and hyperplane can be modelled as:

$$f(g(z)) = \lambda d(n, z) \tag{2.11}$$

The manipulation of a single attribute of a synthesized image can straightforwardly adjust the original latent $z$ with $z_{edit} = z + \alpha$. However, when editing involves multiple semantic attributes, unintended interactions between these attributes may occur due to potential couplings within the semantic features. To address this complexity, the conditional manipulation technique employs orthogonal projection of vectors, which effectively decouples these attributes, thereby facilitating precise control. This work underscores the potential of GANs in semantic face editing, yet it also highlights the neces-

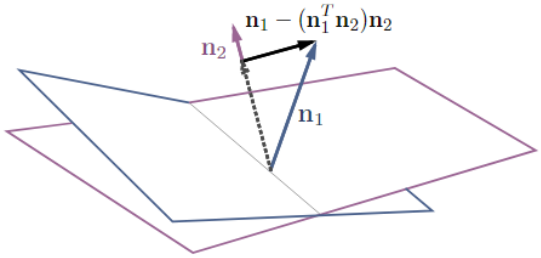



Figure 2.9. Illustration of the conditional manipulation in subspace. Two hyperplanes with normal vectors $n_1$ and $n_2$. Moving samples along the projected direction $n_1 - n_1^T n_2$ can change the "semantic from the blue plane" without affecting the "semantic of the purple plane". Taken from [101].

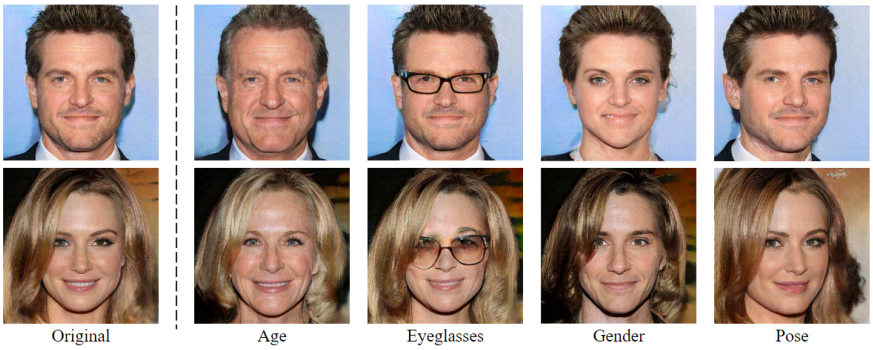



Figure 2.10. The results of manipulating a specific attribute by InterfaceGAN. Taken from [101].

sity for a deeper understanding of latent encoding mechanisms to fully utilize this technology. An illustration of this conditional manipulation is shown in Fig. 2.9, with some results depicted in Fig. 2.10.

One of the primary challenges in GAN-based face synthesis models is the difficulty in controlling the images they generate, primarily because a random distribution typically serves as the input for generators. To address this issue, modified GAN architectures such as Conditional GANs (C-GAN) [82] have been developed. These models set conditions on both the generative and discriminative networks to facilitate conditional image synthesis. However, the mapping in C-GAN does not restrict the output strictly to the

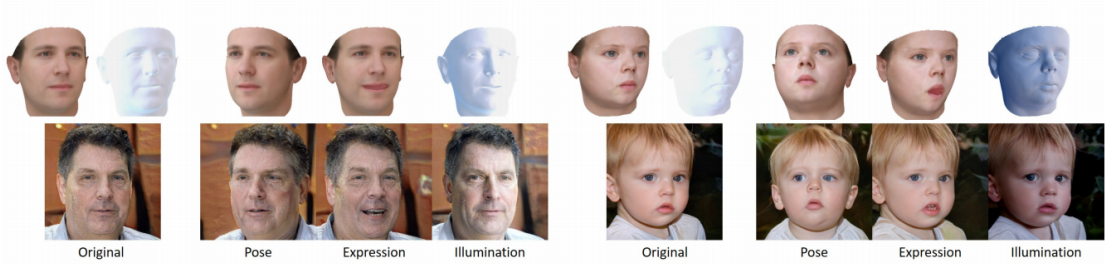



Figure 2.11. The results for various controls over StyleGAN images: pose, expression, and illumination edits. Taken from [117].

target manifold, allowing outputs that may deviate significantly from the desired results. Additionally, generating identity-preserving faces remains an unsolved challenge within these frameworks.

Mokhayeri *et al.* [83] introduced a novel approach to this problem with a cross-domain face synthesis method that utilizes a controllable GAN framework. This method employs a 3D face model as a simulator to generate face images under various poses. These simulated images, along with noise, are then input into a C-GAN, which incorporates an additional adversarial game involving a third player to preserve specific attributes. The C-GAN is designed to generate highly consistent, realistic, and identity-preserving synthetic face images under specific pose conditions. However, this model requires the training of a specific refiner for each attribute and does not control other aspects of facial appearance such as illumination and expression during the synthesis process.

Tero *et al.* [2] proposed an embedding algorithm to project images to the latent space of StyleGAN. The pseudo-cod of their embedding algorithm is illustrated in Fig. 2.12. While this algorithm facilitates semantic image editing operations, it lacks explicit rig-like 3D control over the generative model, a feature that remains highly sought after in the field.

**Algorithm 1:** Latent Space Embedding for GANs

**Input:** An image $I \in \mathbb{R}^{n \times m \times 3}$ to embed; a pre-trained generator $G(\cdot)$.

**Output:** The embedded latent code $w^*$ and the embedded image $G(w^*)$ optimzed via $F'$.

Initialize latent code $w^* = w$;

**while** *not converged* **do**

  $L \leftarrow L_{percept}(G(w^*), I) + \frac{\lambda}{N}\|G(w^*) - I\|_2^2$ ;

  $w^* \leftarrow w^* - \eta F'(\nabla_{w^*} L)$;

**end**

Figure 2.12. The pipeline of the embedding algorithm. The loss function that is weighted combination of the VGG-16 perceptual loss and pixel-wise MSE loss, where $F'$ is the feature output of VGG-16 layers *conv1_1*, *conv1_2*, *conv3_2* and *conv4_2*. Taken from [2].

To explicit control over a set of semantic face parameters that are interpretable in 3D, Tewari *et al.* [117] proposed a method that obtain both the controllable parametric nature of face models and the high-photo realism of generative face models (Sample result shows in Fig. 2.11). Their method is to provide a face rig-like control over a pretrained and fixed StyleGAN via a 3DMM. A new rigging network, RigNet is trained between the 3DMM's semantic parameters and StyleGAN's input. This method is trained in a self-supervised manner and does not require any additional images or manual annotations. Fig. 2.13 shows an overview of StyleRig architecture.

However, StyleRig is not able to exploit the full expressivity of the parametric face model. The author attributes these problems to the bias in the images StyleGAN has been trained on. For instance, StyleGAN is trained on the FFHQ dataset [7] that without in-plane rotations, hence styleRig will also ignore the in-plane rotation of the face mesh.

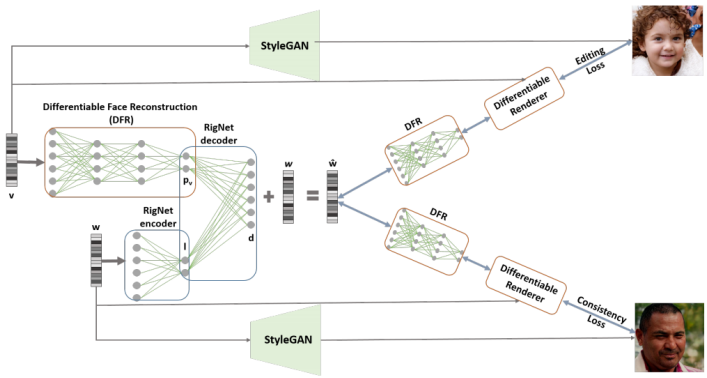



Figure 2.13. StyleRig Structure: The rig-like is based on a RigNet that is trained between the 3DMM's semantic parameters and StyleGAN's input. The differentiable face reconstruction (DFR) and StyleGAN networks are trained, and their weights are fixed, The consistency and edit losses in the image domain using a differentiable renderer. Taken from [117].

## 2.4 Privacy leakage & Adversarial learning

Past studies have shown that face recognition networks encode soft biometric attributes (*e.g.,* race and gender) while training [36, 37, 31]. This indicates that face descriptors are at risk of privacy leaks. Tinsley *et al.* [120] present that there is also privacy leakage in GAN. In addition, respectable studies [81, 31, 40] have shown that privacy leaks will affect the performance of facial recognition systems.

To address bias, Alvi *et al.* [6] proposed a joint learning and unlearning method to reduce bias in neural network embeddings, employing confusion loss to unlearn bias through the calculation of cross-entropy between the classifier output and a uniform distribution. In the realm of GANs, approaches [131, 15, 19, 71] such as differential privacy have been employed to protect privacy by adding noise, thereby masking the maximum change in data-related functions. This method has been further refined by Xu *et al.* [133] with the introduction of GANobfuscator, which mitigates informa-

tion leakage by injecting noise into gradients during the learning process. Similarly, Chen *et al.* [22] have implemented differential privacy on the discriminator using autoencoders to protect privacy while preserving data utility.

Parallel to concerns about bias and privacy, there is a growing interest in understanding and mitigating the risks associated with ID descriptor information and its inversion. Mahendran and Vedaldi's seminal work in 2015 [78] explored the potential for reconstructing images from their encoded representations, showing that deep networks retain rich visual information that can support image reconstruction. This research is particularly relevant for face identity descriptors derived from the final layer of a deep network, representing the most invariant and abstract image representation.

Building on this foundation, several works employ an identity loss in 3D face model fitting, for example, GANFIT [38]. Cole *et al.* [23] also perform an ID-only inversion for face frontalisation. They assume that the face encoder successfully removes all non-ID information and trains to reconstruct only frontal image landmarks and textures. Since they do not use a GAN, their results are not photorealistic. Some recent works [93, 33] have investigated a so-called *black-box attack*, which is whether given only an ID descriptor, and no access to an attacked model, one could reconstruct an image of the face, and they have presented encouraging (technically) as well as concerning (from a privacy perspective) results. These investigations have yielded both technically impressive and privacy-concerning results, underscoring the need for continued research into methods that can protect individual privacy without compromising the utility of facial recognition technologies.

## 2.5 Conclusion

The comprehensive review of literature presented in this chapter highlights significant advancements and challenges in generative face modeling, face recognition, face reconstruction, and privacy considerations. Recent progress in face reconstruction has been primarily driven by the integration of deep generative models, enabling high-fidelity and realistic facial reconstructions. Techniques that leverage unsupervised learning, hybrid approaches combining 3D geometry with neural rendering have gained significant traction. In domain of face recognition, current trends focus on optimizing embeddings to improve both discriminability and fairness. Methods such as contrastive learning and angular margin-based losses have improved the robustness of recognition systems under diverse conditions.

Despite these advancements, critical gaps remain, particularly concerning the leakage of non-ID information in state-of-the-art face recognition systems. Although identity embeddings are designed to capture only identity-specific features, research has shown that these embeddings inadvertently encode non-ID attributes such as soft biometrics (*e.g.,* age, gender, and race), as well as pose, lighting, and background. This unintended encoding raises significant privacy concerns and ethical challenges in the deployment of facial recognition technologies [36, 37, 31].

Building on this foundation, the subsequent chapters of this thesis aim to address these gaps by investigating the extent and implications of non-ID information leakage. Specifically, this work introduces adversarial training techniques to minimize its impact. Furthermore, by leveraging the spatial information retained in identity embeddings, this research demonstrates that spatial features encoded in recognition signals can be repurposed for tasks like face alignment and 3D reconstruction. This dual nature of non-ID leak-

age is critical: while it poses risks to privacy and fairness, it also presents opportunities for advancing related applications.

These contributions aim to enhance the technical rigor, security, and ethical deployment of face recognition technologies, paving the way for systems that are not only more accurate but also more accountable.

# 3 Leakage of non-ID information into face descriptors and its mitigation

## 3.1 Introduction

State-of-the-art face recognition relies on the use of deep neural networks (usually CNNs) to embed a face image to an identity vector [17, 97, 86]. A measure of distance in this embedding space is used to represent dissimilarity in identity. The goal of training such networks is to minimise the within-class scatter while maximising the between-class scatter for all identities. The former goal necessitates that the embedding should depend only on the identity of the person in the image. Environmental conditions such as the lighting, background and properties of the camera as well as changeable aspects of the face such as pose, expression and the presence of accessories should not affect this embedding (i.e. should not introduce within-class scatter). In other words, the embedding network should learn invariance to these factors.

In this chapter, we ask whether ID embeddings truly contain only ID-related information. The engineering of training datasets, network architectures and loss functions has been widely studied in the face recognition literature in order to satisfy the goal of invariance to non-ID factors in the input image. Datasets are created specifically to introduce lots of variation in these non-ID factors. Then, by designing a loss function that encourages the

same ID to embed to the same point, invariance to these factors is hopefully learnt. In this thesis, we investigate to what extent this has been achieved. In particular, we ask: how well do modern face embedding networks successfully remove non-ID factors when embedding a face image?

In an early work, Kumar *et al.* [65] came to the perhaps surprising result that using a so-called *inverse crop*, where the face is cut out of an image, leads to surprisingly high face recognition rates on LFW. It should be noted though that these inverse crops do contain hair and part of ears/chin, and that LFW is a fairly simple dataset (recognition rates achieved are over 99%, even a decade ago) - so it is largely unexplored, if any background or other non-ID information is present in face descriptors, especially in today's state-of-the-art networks.

Having established that non-ID information does leak into the ID descriptors of modern face embedding networks, we secondly investigate whether this leakage can be mitigated and whether this improves face recognition performance. Any leakage of non-ID information degrades the value of the ID descriptors for recognition by introducing distractor information that is unhelpful for recognition. We therefore expect that the removal of this information might improve recognition performance. Consider a concrete example of a person who regularly wears hats. Leaking the non-ID information "wearing hat" into the descriptor is useful for recognising this individual but will degrade performance when trying to recognise this person not wearing a hat.

Recent studies have underscored significant privacy and bias concerns associated with facial recognition technologies. Works such as [36, 48, 115] have demonstrated that face recognition networks encode protected attributes like race, gender, and age while being trained for identity classification. The encoding of such sensitive attributes raises serious privacy and bias issues. One

approach to mitigating these issues involves training face recognition systems with datasets that are balanced in terms of sensitive attributes. However, compiling large datasets balanced in these attributes is challenging, costly, and time-consuming. Another prevalent strategy is the application of differential privacy [34], which reduces information leakage by injecting noise into the gradients during the learning process [133, 22, 131, 15, 19, 71]. Nevertheless, introducing additional noise in descriptors during network training typically results in a degradation of overall performance [50]. Moreover, in contrast to the impacts of soft biometric privacy on facial recognition, the influence of non-ID information—such as background elements and occlusions—on biases within facial identification systems remains insufficiently explored.

In this work, we focus on on the leakage of non-ID information into ID descriptors and explore how mitigating this leakage can enhance face recognition performance. We introduce a technique to detect the presence of non-ID information in face descriptors and propose an adversarial training procedure designed to minimize the leakage of protected attributes. This approach aims to improve the performance and reliability of face recognition systems by ensuring that the ID descriptors are less influenced by irrelevant or sensitive attributes.

## 3.2 Non-ID attribute prediction from ID

We begin by exploring to what extent we are able to estimate non-ID "attributes" from an ID descriptor provided by a pretrained face encoder CNN. We use "attribute" here in very general terms, including image-based attributes such as landmark positions and colour histograms and non-ID face attributes such as the presence or absence of a smile, glasses or hat. For each

attribute, we train an MLP that maps from an ID descriptor to the target attribute. All of our MLPs are trained on the CelebA dataset [77], which contains 202,599 celebrity face images annotated with 40 binary attributes. CelebA is selected for its rich annotations and the wide diversity it offers in terms of facial attributes, poses, and lighting conditions, making it particularly suitable for evaluating attribute prediction tasks. These detailed annotations also support robust training and evaluation of models, ensuring high-quality learning and attribute-specific performance. In this study, we use the `smiling`, `glasses`, and `wearing hat` annotations provided by CelebA as representative non-ID face attributes for training purposes. We aligned and cropped the original CelebA to a VGGFace2 compatible version and scaled all images to resolution 224.

We train in a supervised fashion using either labelled real data or synthetically generated data. In all cases, we embed images $\mathcal{I}$ to an ID descriptor, $\mathbf{d} = f_{\text{ID}}(\mathcal{I})$, where $f_{\text{ID}} : [0,1]^{H\times W\times 3} \rightarrow \mathbb{R}^N$ is the network that embeds to an $N$ dimensional latent space. In practice, we experiment with both the VGGface2 face encoder [17], where $N = 2048$, and the ArcFace encoder [28], where $N = 512$. Fig. 3.1 illustrates our proposed Non-ID attribute regression framework.

### 3.2.1 Discrete Binary Attributes

We begin by estimating discrete binary attributes. These have been manually labelled as part of the CelebA [77] dataset. We train an MLP, $f : \mathbb{R}^N \rightarrow [0,1]$, that predicts the probability of the binary class from the ID descriptor. Our MLP consists of 3 fully connected layers with 256 hidden neurons and a sigmoid output layer. We train the classification MLP using Binary Cross Entropy loss, the Adam optimiser with learning rate $10^{-3}$, for 20 epochs. We

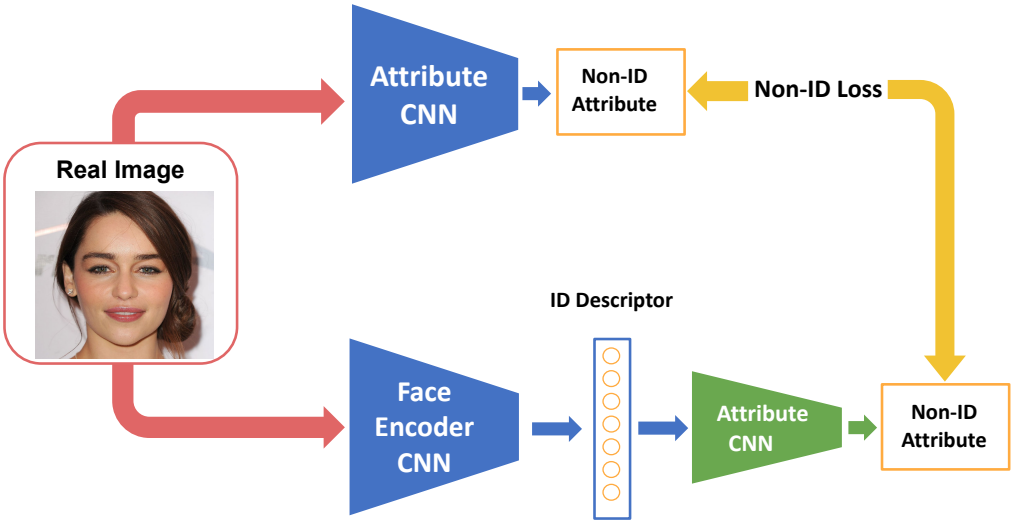



Figure 3.1. Non-ID attribute regression via an ID bottleneck. Only the green component is trained: an MLP that maps an ID vector to the appropriate attribute (such as expression, landmarks, image histogram etc). The labels either come from a pretrained (and fixed) attribute estimation network that takes an image as input, or they are provided as manually assigned labels or they are computed directly from the input image (in the case of the image histogram attribute).

train separate networks for the `smiling`, `glasses` and `wearing hat` binary, non-ID attributes.

### 3.2.2 Histogram regression

Histograms of RGB intensities provide a global summary of an image that encapsulates not only ID but also environment-related features such as background, camera settings and lighting. We therefore investigate whether image histograms can be recovered from ID descriptors.

For each image, we compute a ground truth histogram of intensities for each colour channel and use this as the label for training. Here, the three colour values are binned into fixed width bins, with one histogram per colour channel.

We found that a very small MLP provides best performance for the task

of image histogram regression from ID vector. We use 2 fully connected layers with ReLU activation and 8 neurons per layer. We apply softmax to the output layer such that the output represents a normalised histogram. We use $N = 10$ bins. We train this network using mean squared error loss, Adam with learning rate of $10^{-6}$, and batch size 32, for 20 epochs.

### 3.2.3 Landmark regression

Finally, we attempt to directly regress the coordinates of 68 face landmarks from the ID vector. We apply the dlib landmark detector [59] to all images in the CelebA dataset [77]. We use these as pseudo ground truth labels for our regressor. We use a three-layer MLP to predict landmarks from ID vector with 256 neurons per hidden layer and 136 outputs for the 2D coordinates of each landmark. For comparison, we also train a more conventional image to landmark regressor using a CNN. We use a simple architecture comprising 5 convolutional layers followed by 2 fully connected layers. The activation function is ReLUs. Both image to landmark and ID vector to landmark networks are trained using mean squared error loss and the SGD optimiser with learning rate $10^{-3}$, batch size 16, and for 150 epochs. Tinsley *et al.* [120] is the first work to present identity leakage in StyleGAN. They find that identity information in face images can leak from the training datasets into synthetic result.

### 3.2.4 Results

We now evaluate our prediction of non-ID attributes from ID vectors, as provided by the VGGFace2 network. We show quantitative results for all attributes in Table 3.1. The evaluation images we used are 15k test images from CelebA with the remaining 188k images for training. To validate our

| | Attribute | | | Landmarks | Histogram |
|---|---|---|---|---|---|
| | Smiling | Wearing_Hat | Eyeglasses | mean | EMD |
| From ID (VGGFace2 [17]) | 91.0% | 99.0% | 99.7% | 9.3% | 2.68 |
| From ID (ArcFace [28]) | 81.7% | 96.7% | 96.7% | 7.3% | 2.45 |
| From image [77] | 92% | 99% | 99% | 8.2% | 0.0 |
| Baseline | 50.4% | 96.7% | 94.0% | 11% | 2.97 |

Table 3.1. Quantitative results for attribute prediction (discrete binary attributes, landmarks and image histogram) from ID vectors (row 1 and 2) and images (row 3). In row 4 we show baseline performance in which we simply always predict the most common class, the mean landmarks or the mean histogram respectively. The attribute prediction results show accuracy (higher is better), the landmark prediction is measured in percentage of interocular distance (lower is better), and for the histograms we measure Earth Mover's disance (EMD) to ground truth (lower is better).

results, we repeat the experiments with ID vectors generated by the ArcFace network.

For discrete binary attributes, we show the percentage classified correctly. Our result regressed from the ID vector is shown in the first and second rows. This shows that non-ID leakage exists in different networks. For comparison in the third row we show the result from [77] computed *from the original image*. In the fourth row, we show the baseline performance obtained by always guessing the more common class for the binary attribute prediction, the mean landmarks for the landmark prediction, and the mean histogram for the histogram prediction. We can see that we significantly outperform that baseline and, remarkably, match or even exceed the performance of an image-based method despite only having access to an ID vector that should be independent of these non-ID attributes. In Fig. 3.2 and Fig. 3.3 we show some examples of correctly and incorrectly classified samples. It is interesting to note that quite subtle smiles are encoded in the ID vectors such that we correctly classify them and, even in the case of the false positives shown, there are still smile-like features in the wrinkles around the mouth. Similarly, the


Figure 3.2. Examples of correctly classified samples. We show the original images but note that the classification is done *only on the ID vectors derived from these images*.

false positives for wearing hat are in fact wearing headgear.

We now evaluate image-based attributes. In Table 3.1 we show quantitative results for landmark prediction in the fifth column and histogram prediction in the sixth column. For the landmark error, we show Euclidean distance averaged over landmarks expressed as a percentage of the interocular distance. For histogram error, we show the Earth Mover's distance to ground truth. Our prediction from ID vector outperforms the baseline and is only marginally worse than prediction from images (in the case of landmarks). In the case of histogram, the prediction from images is exact. In Fig. 3.4 we show qualitative results for landmark and histogram estimation

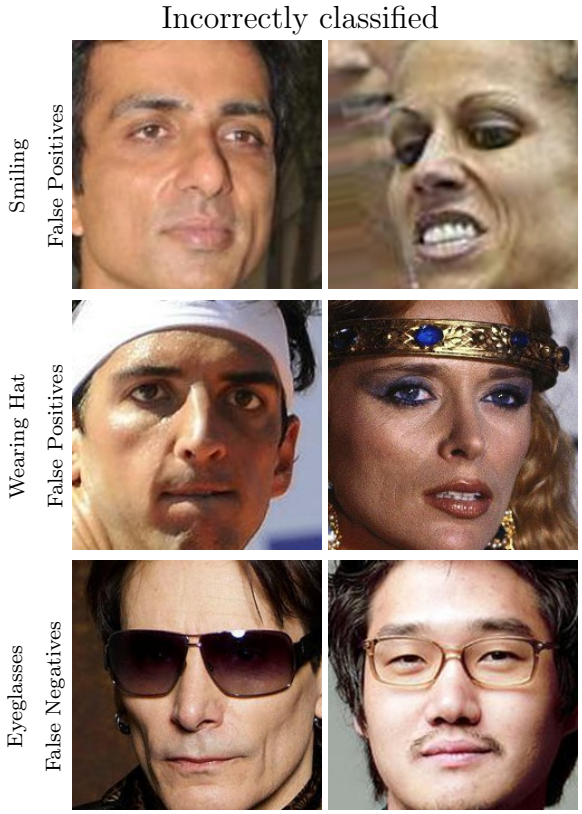



Figure 3.3. Examples of incorrectly classified samples. The classification errors include both false positives and false negatives, based *only on the ID vectors derived from these images*.

from ID vectors. In the first row we show the original image with ground truth (dlib) landmarks overlaid. In the second row we show the original images with the landmarks regressed from the ID vector overlaid. The landmarks are qualitatively convincing and clearly reconstruct pose - an entirely non-ID related property. In the third row we show the ground truth (dotted lines) and estimated (solid lines) RGB image histograms.

This section of our work demonstrates the capability of our approach to accurately predict non-ID information attributes, such as landmarks and color histograms, from face embedding. Notably, our method's performance in landmark prediction closely approaches that achieved by direct image anal-

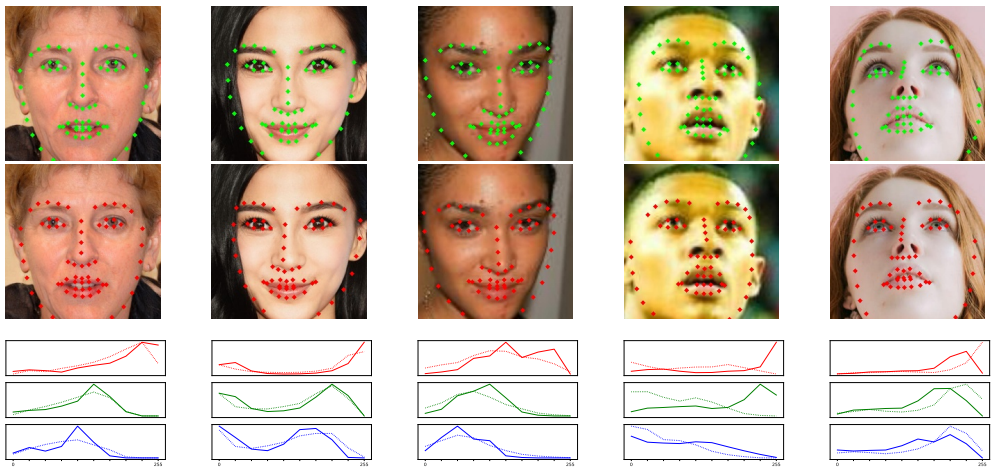


Figure 3.4. Qualitative results for histogram and landmark regression. row 1: input image with ground truth landmarks, row 2: landmarks regressed from ID vector, row 3: ground truth image histograms (dotted) and histograms regressed from ID vector (solid).

ysis, underscoring the depth of information encoded within ID embedding. Furthermore, the reconstruction of image histograms from ID embedding underscores the potential for non-ID information leakage in face recognition systems. The reconstruction of image histograms from ID embeddings unveils the potential for recuperating environmental information, such as background elements, from face embeddings, thereby amplifying concerns related to privacy. Our work not only investigate the extent to which non-ID attributes can be inferred from identity embedding but also lays the groundwork for future works aimed at improving the privacy and integrity of face recognition network through the mitigation of such leakage.

## 3.3 Mitigating non-ID leakage

In this section, we aim to suppress the leakage of non-ID information into facial descriptors. Preliminary findings, discussed in Section 3.2, reveal that

even state-of-the-art ID embeddings inadvertently retain elements of expression information, such as the ability to predict a subject's smile from their ID embedding. This issue is corroborated by recent studies which show that encoding non-ID attributes like race, gender, and age within facial descriptors can compromise the effectiveness of facial verification and matching systems [31, 36, 69]. Our objective is to refine facial descriptors to minimize the retention of non-ID information, thereby enhancing the performance of facial recognition systems in verification tasks. To achieve this, we propose an adversarial training method designed to selectively mitigate non-ID attributes from the descriptors while ensuring that the identity information crucial for accurate verification remains robust. We hypothesize that this approach will not only reduce bias but also improve the performance of face recognition technologies.

### 3.3.1 Adversarial Debiasing

**Method overview** The key idea in our proposed approach is to train a model to classify identities while discouraging it from predicting a specific protected attribute.

Let $\mathcal{I}_i$ denote the $i$th training face image and $F(\cdot)$ represent the feature extraction function performed by the backbone network $F$. The corresponding feature vector, i.e. face descriptor, $\mathbf{d}_i$ for each image is obtained as follows:

$$\mathbf{d}_i = F(\mathcal{I}_i). \tag{3.1}$$

We present our network architecture in Fig. 3.5. The proposed architecture consists of three main components:

1. **Face embedding network (backbone)** $F$: The face embedding net-

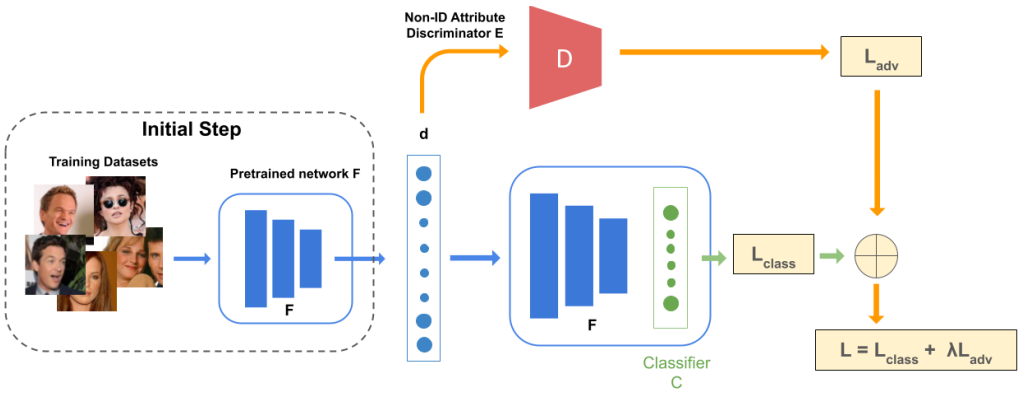



Figure 3.5. Overall Pipeline of Our Method. In stage 1, face descriptors $\mathbf{d}$ are extracted from a pre-trained network $F$. In stage 2, these extracted descriptors are input into a non-ID attribute classifier $D$, which is trained to classify specific non-ID attributes using the gradients of the loss function $L_{adv}$. In stage 3, an identity classifier $C$ is trained to utilize the gradients of both $L_{adv}$ and $L_{class}$. The gradients from $L_{adv}$ are employed to debias $\mathbf{d}$ with respect to the target attribute while also enabling the classification of identity.

work is responsible for extracting identity features $\mathbf{d}_i$ from face images $\mathcal{I}_i$. In the context of adversarial learning, we can also think of this as the *generator* network, which generates face descriptors.

2. **Classifier** $C$: A classifier that takes in $\mathbf{d}_i$ and generates a prediction vector for identity classification.

3. **Non-ID attribute classifier** $D$: An attribute prediction model that takes $\mathbf{d}_i$ as input and attempts to classify non-ID attributes. In the context of adversarial learning, this network plays the role of the discriminator.

We now explain our network as an adversarial approach. Backbone $F$ is used to generate descriptors $\mathbf{d}_i$. $\mathbf{d}_i$ is fed to the prediction model $D$ which acts as a discriminator and aims to classify the non-ID attribute. The goal is to train $F$ and $C$ to maximise face classification performance while minimising

the ability of the adversary $D$ to extract non-ID information from the face embeddings.

### 3.3.2 Losses

**Training discriminators** The non-ID discriminator $D$ aims to classify non-ID attributes based on the feature descriptors $\mathbf{d}_i$. Let $D(\cdot)$ denote the discrimination function of non-ID discriminator $D$, and $y_i$ represent the true non-ID attribute labels. The loss function $L_D$ for updating $D$ is defined as the cross-entropy loss between the predicted and true labels:

$$L_D(\mathbf{d}_i, y_i) = -\sum_{c=1}^{C} y_{i,c} \log(D(\mathbf{d}_i)_c), \tag{3.2}$$

where $C$ is the number of non-ID attribute classes, and $y_{i,c}$ is a binary indicator of whether class $c$ is the correct classification for observation $i$.

**Adversarial Training to Fool discriminator** During adversarial training, the goal is to update the face embedding network to generate features that discriminator $D$ cannot easily classify, by minimizing the negative of the loss function used for $D$. The adversarial loss $L_{adv}$ for updating the backbone is defined as:

$$L_{adv}(\mathbf{d}_i, y_i) = \sum_{c=1}^{C} y_{i,c} \log(D(\mathbf{d}_i)_c). \tag{3.3}$$

**Identity Recognition Training** The identity recognition component of the model is trained using a modified ArcFace loss [28], which enhances the discriminability of the identity features.

Let $C(\cdot)$ represent the identity recognition function, and $z_i$ the identity

label for image $\mathcal{I}_i$. The ArcFace loss $L_{ArcFace}$ is defined as:

$$L_{class}(\mathbf{d}_i, z_i) = -\log \frac{e^{C(\mathbf{d}_i)_{z_i}}}{\sum_{j=1}^{N} e^{C(\mathbf{d}_i)_j}} \tag{3.4}$$

where $N$ is the number of identity classes, and $C(\mathbf{d}_i)_j$ represents the score of class $j$ for the feature vector $\mathbf{d}_i$.

**Overall Training Objective** The overall training objective combines the adversarial loss with the identity recognition loss, aiming to simultaneously improve the identity recognition capability of the backbone while reducing its sensitivity to non-ID attributes. The combined loss $L$ for a given input image $x_i$ and its corresponding identity and non-ID labels $z_i$ and $y_i$ is:

$$L(\mathbf{d}_i, z_i, y_i) = L_{class}(\mathbf{d}_i, z_i) + \lambda L_{adv}(\mathbf{d}_i, y_i) \tag{3.5}$$

where $\lambda$ is used to balance between the two losses (a high value of lambda will sacrifice recognition accuracy in order to remove more non-ID information from the descriptor).

### 3.3.3 Training Steps

We now discuss the training process of our network, which utilizes an adversarial training mechanism. This approach systematically updates both the identity recognition network and the non-ID discriminator in a specific sequence, with the aim of mitigating non-ID information from the facial descriptors. The identity recognition network consists of the face embedding network $F$, which serves as the feature extraction backbone, and the classifier $C$, which maps the extracted features to identity labels. The training process can be broken down into the following steps:

**Step 1 - Feature Extraction:** The face embedding network $F$ extracts descriptors $\mathbf{d}_i$ from the input face images $\mathcal{I}_i$.

**Step 2 - Update non-ID Discriminator:** $\mathbf{d}_i$ are then fed into the discriminator network $D$, which attempts to classify non-ID attributes based on the features. The loss from the discriminator is used to update the weight of $D$ denoted as $\emptyset_D$, encouraging it to accurately distinguish between different non-ID attributes.

**Step 3 - Adversarial Training:** Next, the face embedding network $F$ is updated to encourage the removal of non-ID information from the descriptors. The adversarial training involves calculating the gradient of the adversarial loss $L_{adv}$ with respect to the descriptors $\mathbf{d}_i$ and backpropagating this into the face embedding network to update the weights $\omega_F$. This encourages the backbone $F$ to generate features that are challenging for the discriminator $D$ to classify, effectively purging non-identity information from the descriptors. This step is pivotal in debiasing the feature representations towards pure identity information.

**Step 4 - Identity Recognition Training:** Finally, classifier $C$ and face embedding network $F$ are updated to reduce the ArcFace loss to enhance the discriminability of identity features.

### 3.3.4 Implementation Details

We employ VGGFace2 [17] and MS1MV3 [43] as our training datasets to conduct comparisons with other methods. For the embedding network, we adopt the **ResNet-50** architecture as described in [28].

The **VGGFace2** dataset is widely used in deep face recognition. It comprises over 3.3 million images across more than 9,131 identities. Sourced from Google, the images are characterized by their diversity, encompassing

variations in pose, age, lighting, and ethnicity. This diversity makes it an ideal choice for training face recognition networks due to its extensive scope and scale.

The **MS1MV3** dataset, a cleaned version of MS1MV0 using a semi-automatic approach, has been instrumental in the development and benchmarking of face recognition algorithms. It includes approximately 5.2 million images of 93,000 individuals, offering a diverse collection of facial images sourced from the internet. This diversity covers a wide range of variations in pose, lighting, expression, and occlusion, presenting a challenging yet realistic scenario for training robust face recognition systems.

During training, we employ the LFW [49], CFP-FP [98], and AgeDB-30 [84] datasets as validation sets to monitor the convergence status of the model. For evaluation, we utilize aligned faces from the IJB-C dataset [79]. The IJB-C dataset encompasses 3,531 subjects with 31,300 still images and 117,500 frames from 11,779 videos. There are two evaluation protocols employed on the IJB-B and IJB-C datasets: 1:1 verification and 1:N identification. In our study, we adhere to the 1:1 face verification protocol, which includes 12,115 templates featuring 10,270 genuine matches and 8 million impostor matches.

**Pretraining Non-ID Attribute Classifier: Expression MLP** We employ the methodology outlined in Section 3.2 to train a Multi-Layer Perceptron (MLP) classifier, referred to here as the Non-ID Attribute Classifier $D$, for categorizing facial expressions encoded in ArcFace descriptors, which are extracted from the MS1M dataset [43]. The expression classifier $D$ is trained in a supervised manner, utilizing expression labels predicted by DeepFace [99], which outputs prediction scores for seven expressions: dis-

gust, fear, happiness, sadness, surprise, anger, and neutral. The dominant expression for each image is determined by selecting the emotion with the highest score among these categories. While our approach could use any attribute or combination of attributes, we chose expressions as an exemplar and obvious non-ID feature that can be reliably labelled from images.

The expression classifier $D$ comprises 13 hidden layers, each containing 16 neurons with ReLU activation functions. The output layer is equipped with a number of neurons corresponding to the different facial expression categories. We conduct the training with a learning rate of 0.01 across 400 epochs and the classifier is subsequently evaluated on descriptors derived from the IJB-C dataset. This procedure allows $D$ to serve as an adversary, estimating non-ID characteristics, specifically expressions, from an ArcFace ID embedding.

**Training Setting.** For data preprocessing, we follow the paper [28] to generate the normalized faces ($112 \times 112$) with five landmarks predicted by RetinaFace [27]. For the embedding network, we adopt Resnet50. We set the batch size to 512 and trained models on 4 NVIDIA TESLA V100 GPUs. The models are trained with SGD, with momentum to 0.9 and weight decay $5e^{-4}$. The learning rate for the feature embedding network starts from $1e^{-2}$ and is divided by 10 at $10, 16, 22, 28$ epochs. The learning rate for the discriminator is fixed at $1e^{-3}$. The training process is finished at 35 epochs. the value of $\lambda$ for adversarial loss is set to 5.0.

### 3.3.5 Experiments

**Experimental Settings** As outlined in Section 3.3.4, we utilize the MS1MV3 and VGGFace2 datasets separately as our training data to facilitate a fair comparison with current state-of-the-art face recognition methods. For the

| Datasets | IJB-C(↑) |
|---|---|
| MS1MV3 | 95.55 |
| Vggface2 | 90.9 |
| MS1MV3 (ours) | **97.70** |
| Vggface2 (ours) | **91.93** |

Table 3.2. The $1:1$ verification accuracy (TPR@FPR=$1e-3$) on IJB-C datasets. All methods use a deep network ResNet-50 as its backbone. The previous results of Arcface and Vggface2 are from [138]

embedding network, we employ a modified ResNet-50 architecture, similar to that used in ArcFace [28]. This configuration is designed to produce 512-dimensional discriminative features for each image.

Arcface [28] achieves state-of-the-art performance in face verification and identification. Hence, we construct the baselines and our framework based on the top of Arcface descriptors, i.e. we initialise with a conventionally trained Arcface network and finetune from there. We also perform similar experiments with VGGface2 datasets with ResNet50 [17]. For evaluation, we use aligned faces from IJB-C [79] and follow the same 1:1 face verification protocol. The 1:1 verification protocol is specifically designed to assess the efficacy of face recognition algorithms by calculating the likelihood that two facial images represent the same individual. This involves computing similarity scores derived from facial features extracted by the recognition model. In the IJB-C dataset, there are 23,124 templates, encompassing 19,557 genuine matches and 15,639K impostor matches. Performance metrics for this protocol primarily involve the True Positive Rate (TPR) at various False Positive Rate (FPR) thresholds. This comprehensive approach offers a detailed evaluation of an algorithm's precision in confirming identities under the stringent conditions set forth by the IJB-C framework.

In Table 3.2, we present our 1:1 verification results alongside those obtained using the ArcFace and VGGFace2 methods on the IJB-C dataset,

respectively. It can be seen that by integrating our adversarial learning approach with the ArcFace and VGGFace2 frameworks, our methods are capable of further enhancing performance, achieving an accuracy of 97.70% on ArcFace and 91.93% on VGGFace2. Notably, our adversarial technique yields more substantial performance improvements when applied to ArcFace trained on the MS1MV3 dataset compared to its application on the VGGFace2 dataset. This suggests that our approach is particularly effective when deployed on larger-scale datasets.

The MS1MV3 dataset, which is sourced mainly from the internet, includes about one million images of 100,000 individuals. It features a mix of non-celebrity images and a higher degree of noise and variability. In stark contrast, the VGGFace2 dataset contains over 3.31 million high-quality images from 9,131 celebrities, characterized by diversity in age, ethnicity, and pose. Consequently, the presence of non-ID information—encompassing background, lighting conditions, expressions, and more—is markedly more pronounced in MS1MV3, posing greater challenges to facial recognition algorithms. The application of adversarial learning to remove non-ID information on the MS1M dataset enables models to focus more on discriminative features, thereby leading to more significant performance improvements. This outcome emphasizes the necessity of eliminating non-ID information from facial features, particularly in an era where training datasets are increasingly expansive.

## 3.4 Conclusion

Our investigation into the leakage of non-ID information into facial descriptors has unveiled a critical vulnerability in the current face recognition tech-

nologies. This leakage, characterized by the unintended encoding of extraneous attributes such as environmental conditions, facial expressions, and accessories within identity vectors, poses a substantial challenge to the integrity and reliability of these systems. Non-ID information is a nuisance factor for face recognition. It means that some of the capacity of the embedding space is wasted on useless information and that distance measures incorrectly observe identity dissimilarity when in fact the difference is due to non-ID factors.

Through a methodical exploration, employing both predictive modelling and adversarial learning techniques, we have quantitatively assessed the extent of this leakage and its implications on the performance of face recognition models. It shows that the presence of non-ID information can introduce a potential bias, thereby compromising the accuracy and fairness of face recognition tasks. The application of adversarial learning in our work provides a route to improving face recognition performance while also alleviating privacy concerns. By effectively penalizing the inclusion of non-ID information, this approach not only reduces leakage but also improves face verification performance.

# 4 ID2image: Inversion from face descriptors to images with a generative model

## 4.1 Introduction

In this chapter, we first ask whether it is possible to create an image from an identity (ID) vector that correctly recreates the identity of the person. Conventional text-based passwords are usually passed through a cryptographic hash function for storage [5]. Since these are pseudo-one-way functions, it is extremely difficult to invert an encrypted password to cleartext, with brute force or dictionary attacks as the only options. For password verification, only the encrypted version needs to be stored and a leak is not critical due to the difficulty of inversion.

It is tempting to assume that ID embeddings from face images possess similar characteristics. For example, the Face ID facial recognition system developed by Apple Inc. makes this argument in its advertising to reassure users:

> *Face ID doesn't store an image of your face. Instead of storing an image, Face ID saves a mathematical value created from the characteristics of your features. It's impossible for anyone else to recreate your likeness from this [52].*

However, since the embedding of a face image to an ID vector is a noisy pro-

cess, it cannot be assumed that an identical embedding will arise from any image of the same person's face. Therefore, the ID vector cannot be passed through a hash function for storage since the hashed value will change dramatically with small changes in the ID vector. From a security perspective, this means that raw ID vectors must be stored for future identification. Additionally, from an inversion perspective, similar images tend to map to similar ID vectors. Since embeddings are computed using a differentiable function in deep neural networks, it is feasible to optimize an image representation to minimize ID vector loss. This means that it is possible to optimise an image representation to minimise an ID vector loss, thereby reconstructing an image with the desired identity (subject to suitable regularisation, which we achieve via a generative model).

We connect this line of investigation to the work in Chapter 3 by asking whether it is possible to recover an image from an ID vector that not only captures the correct face identity *but also non-ID characteristics of the actual image that was used to compute the ID vector*. The possibility of ID vector to image inversion raises privacy and security questions.

For the reasons mentioned above, ID descriptor vectors cannot be stored securely hashed. This means that any third party with whom identity must be verified is receiving an encoding of a face image from which an image can be recovered. Where non-ID information leaks into this representation, it means the image itself can potentially be recovered. This could, for example, leak unintended information in the background of the image or maybe an unflattering image that the user would not wish to be made public.

Against this backdrop, we utilize advanced generative models for ID-to-image inversion to explore non-ID information leakage in face embeddings, aiming to investigate the privacy leakage problem in current facial recognition

systems.

GANs have proven effective in generating photorealistic images, sparking significant interest in reversing the GAN generation process—a technique commonly referred to as GAN inversion [9, 94, 126]. Particularly, StyleGAN, developed by Karras *et al.* [57, 58], excels in creating high-resolution, lifelike facial images. Several studies have successfully demonstrated the reconstruction of target images from latent codes using StyleGAN [101, 117, 2, 3, 135]. These methods typically perform *image-to-latent* inversion, where a given image is mapped back to the latent code of a GAN through an inversion process. In contrast, our research focuses on *ID-descriptor to image* inversion, where we use StyleGAN as the generative model to map ID descriptors directly to StyleGAN codes and subsequently to images.

Few works have attempted to invert face descriptors back to face images. Genova *et al.* [39] tried inverting face descriptors by using unsupervised training to convert face descriptors into face images, training a regressor to map ID descriptors to 3DMM parameters and minimizing the ID loss between the ID descriptor of the original image and the ID descriptor of a rendered 3DMM image. Similarly, Gecer *et al.* [38] and Cole *et al.* [23] utilized an identity loss in 3D face model fitting for face inversion; however, their methods, which do not leverage a GAN, produce outcomes lacking photorealism.

To date, the potential for recovering non-ID information from face descriptors remains largely unexplored. An early investigation by Kumar *et al.* [65] into the use of an "inverse crop" technique—where the face is removed from an image—revealed surprisingly high face recognition rates on the relatively straightforward LFW dataset, despite the inclusion of only peripheral features such as hair and parts of the ears and chin. This suggests that background or other non-ID information may be encoded in face descrip-

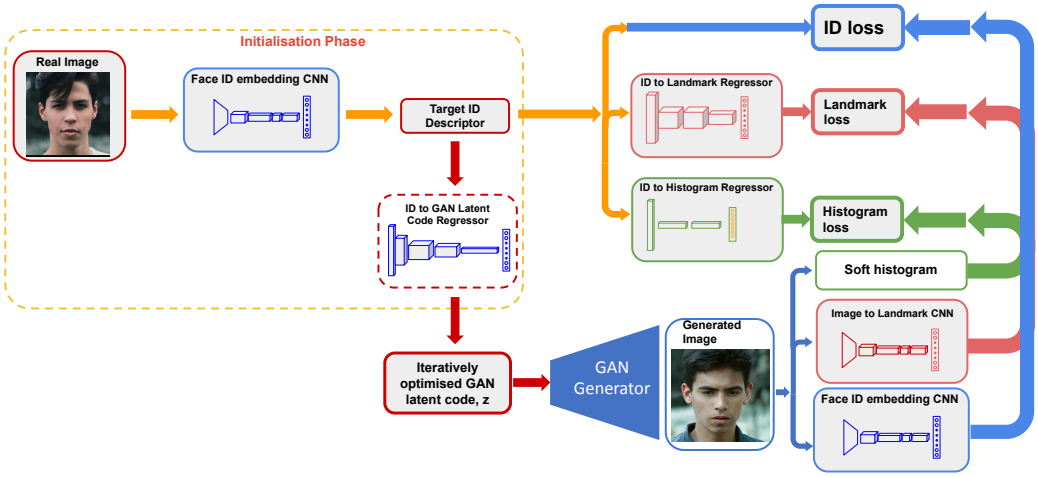



Figure 4.1. Reconstructing an image from an ID descriptor, including preservation of non-ID properties (pose, expression, and color distribution). We assume we only have access to the ID descriptor of a real image. We initialise the optimisation using a regression network to predict GAN latent code from ID descriptor. We then iteratively optimise the GAN latent code in order to produce an image that matches the ID, landmarks and histogram predicted from the target ID descriptor using pretrained networks.

tors, a hypothesis that remains underexplored. To the best of our knowledge, no study has yet investigated the recovery of non-ID properties from identity descriptors thoroughly.

## 4.2 Image from ID with a generative model

We now set out to investigate if, and how well, an image of a person can be recovered from an ID descriptor. Having shown that non-ID attributes can be estimated, or extracted, from a face descriptor, we also explore to what extent the original image itself, including non-ID information, can be reconstructed from an ID descriptor.

### 4.2.1 ID-only inversion

We begin by attempting to create an image that is recognisably the same *identity* as the person in the original image but not necessarily similar in other ways to the original image. We pose this as an optimisation problem and use a generative face model to constrain the problem. Specifically, we restrict the solution to the space of images represented by the StyleGAN2 face model [58]. We denote by $\mathcal{I} = G(\mathbf{z})$ the face image that arises via the generator, $G$ from the latent code $\mathbf{z}$. At inference time, the only optimisation variable is the GAN latent code which is iteratively optimised. The generator itself is never trained (we use a pre-trained StyleGAN2 generator which we keep fixed).

Through this adversarial training mechanism and structured latent representation, StyleGAN2 effectively explores the complex diversity of facial images, achieving the transformation from the latent space $\mathbf{z}$ to high-fidelity facial images. It also provides more nuanced control over the image generation process, resulting in visually coherent and diverse outcomes. A key feature of the StyleGAN architecture is the introduction of two latent spaces: the original latent space $Z$ and the mapped latent space $W$. The $Z$ space is typically a high-dimensional, randomly sampled vector space, following a standard distribution such as Gaussian. In contrast, the $W$ space is derived from the $Z$ space through a mapping network. $W$ is designed to produce a more disentangled representation, where variations in the $W$ space have more linear and interpretable effects on the generated images. This disentanglement effectively separates high-level attributes and stochastic variations (e.g., freckles, hair) from the structural aspects of the image, enabling precise control over the synthesis process. While the $W$ space enhances image quality and control, its complexity introduces uncertainty in inverse tasks such as

face restoration. Therefore, in our work, we directly optimize the latent code in the $Z$ space to reconstruct the input image $\mathcal{I}$, thereby simplifying the inverse process and reducing uncertainties associated with the transformation step to the $W$ space.

Suppose we are given a target ID descriptor, $\mathbf{d}$, then we wish to solve the following optimisation problem:

$$\min_{\mathbf{z}} L_{\text{ID}}(\mathbf{z}), \quad \text{where } L_{\text{ID}}(\mathbf{z}) = \|F(G(\mathbf{z})) - \mathbf{d}\|_2^2, \tag{4.1}$$

where $F$ is the same pretrained face encoder network as in the previous chapter. $F(G(\mathbf{z}))$ is therefore the descriptor extracted from the face image synthesised by the GAN generator according to latent code $\mathbf{z}$. We can solve this optimisation problem with gradient descent over the unknown latent code parameters.

In practice, this optimisation is prone to convergence on local minima and sensitive to initialisation. For this reason, we train a network that we use for initialisation that regresses a StyleGAN2 latent code directly from an ID descriptor.

We introduce the ID-to-GAN Latent Regressor, denoted as $N$, which is trained to regress the GAN latent code $\mathbf{z}$ from a given target ID descriptor $\mathbf{d}$. We can train this network in a supervised manner by generating training data as follows. A random latent code is drawn from a Gaussian distribution $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The corresponding descriptor is computed as $\mathbf{d}_i = F(G(\mathbf{z}_i))$. Now, the regressor is trained to predict a GAN latent from $\mathbf{d}_i$ that is close to the target by minimising the following loss:

$$L_{\text{inv}} = \|N(\mathbf{d}_i) - \mathbf{z}_i\|_2^2. \tag{4.2}$$

This loss function is designed to minimize the Euclidean distance between the predicted latent codes and the actual latent codes. This minimization aids the network in effectively optimizing $\mathbf{z}$ to align with the target identity features represented in $\mathbf{d}$.

The regressor network architecture consists of a Multilayer Perceptron (MLP) with three hidden layers, each equipped with 2,048 ReLU-activated units, and it outputs a StyleGAN2 latent vector. We optimize the network using the Adam optimizer with a learning rate of 0.001 as specified in Equation 4.1. Some of the results from our ID to GAN latent Regressor are displayed in the second column of Fig. 4.2. The regressed results can be subsequently refined by nonlinear optimisation of (4.1).

### 4.2.2 Image reconstruction

The results of the above process successfully produce an image with the correct identity, as shown in Table 4.1 and Fig. 4.2, which demonstrate the alignment between the reconstructed images and target identities. However, the reconstructed images often fail to accurately capture certain features of the original image, such as the pose and expression of the face, the lighting in the image, the background, and the presence of apparel.

The results of the above process successfully produce an image with the correct identity. However, they often fail to reconstruct certain features of the original image, for example, the pose and expression of the face, the lighting in the image, the background and the presence of apparel. We have shown that, with suitable supervision and training, it is possible to extract some of these properties from weak signals that find their way into the ID descriptor. Once reconstructed, we now show that these can be used to provide additional, direct supervision to the inversion problem. Essentially,

we ask that not only the ID be reconstructed but that additional, non-ID, features estimated from the ID descriptor also be reconstructed (specifically pose and image histogram).

To clarify, we treat landmarks as a non-ID factor in this study. While landmarks can encode intrinsic identity-related features, such as facial shape, their variations are primarily influenced by extrinsic factors, including pose and expression. These extrinsic factors are our primary focus in the context of image reconstruction. Consequently, the landmarks are designed to capture pose and expression information rather than identity. In subsequent discussions, we consistently refer to "landmarks" as representing non-ID properties to ensure terminological consistency and emphasize the disentangling of identity and non-ID features.

**Overview of Inversion** Fig. 4.1 provides an overview of the proposed ID-descriptor to image inversion framework. The process begins with an ID descriptor, $\mathbf{d}$, which encodes identity-specific features extracted from face images. Instead of directly using face images as inputs, the ID descriptor serves as the starting point for inversion. A generative model, specifically StyleGAN, is employed to map $\mathbf{d}$ to its corresponding latent code in the GAN's latent space, which is then used to synthesize realistic face images.

The inversion framework is guided by a combination of ID and non-ID supervisory signals. The ID supervision ensures that the reconstructed images preserve identity-relevant features embedded in $\mathbf{d}$, ensuring fidelity to the original descriptor. Non-ID supervision focuses on reconstructing additional attributes, including pose and expression (represented by facial landmarks) and overall color distribution (represented by the image histogram). These non-ID attributes are estimated from $\mathbf{d}$ through pretrained auxiliary net-

works $f_{\text{ID}\to\text{landmarks}}$ and $f_{\text{ID}\to\text{histogram}}$. Dedicated loss functions are introduced to enforce consistency between the reconstructed image and the estimated non-ID attributes, encouraging the framework to model pose, expression, and overall image appearance accurately.

**Landmarks** From the target ID descriptor, $\mathbf{d}$, we use the pretrained regression network described in Section 3.2.3 to compute approximate target landmarks, $f_{\text{ID}\to\text{landmarks}}(\mathbf{d})$. During reconstruction, we compare the target landmarks with those extracted from the current image reconstruction using the pretrained image to landmark regression CNN, $f_{\text{image}\to\text{landmarks}}(\mathbf{i})$:

$$L_{\text{landmarks}}(\mathbf{z}) = \|f_{\text{image}\to\text{landmarks}}(G(\mathbf{z})) - f_{\text{ID}\to\text{landmarks}}(\mathbf{d})\|_2^2. \tag{4.3}$$

**Soft histogram** For the histogram reconstruction loss, we follow a similar strategy. We use the pretrained regression network described in Section 3.2.2 to compute an approximate target histogram, $f_{\text{ID}\to\text{histogram}}(\mathbf{d})$. The exact histogram of the reconstructed image is discrete and therefore not differentiable. For this reason, we use a differentiable soft approximation of the image histogram.

The idea is to use sigmoid to softly assign values to bins. Consider a vector $\mathbf{x} \in \mathbb{R}^M$ of $M$ values. We wish to compute a soft histogram $H(\mathbf{x}) \in \mathbb{R}^N$ which softly assigns all values in $\mathbf{x}$ to $N \in \mathbb{Z}^+$ histogram bins. We specify minimum and maximum values (we use min = 0 and max = 255 for image histograms) and the bin width by $\delta = \frac{\max - \min}{N}$. The $i$th bin centre is given by $c_i = \min + \delta(i - 0.5)$. Then, the value of the $k$th bin in $H$ is:

$$H(\mathbf{x})_k = \sum_{j=1}^{M} f(\mathbf{x}_j - c_k + \delta/2) - f(\mathbf{x}_j - c_k - \delta/2), \tag{4.4}$$

where $f$ is an assignment function. In a hard (non-differentiable) histogram, $f$ is the Heaviside step function. In our soft histogram, we use sigmoid, $f(x) = \textbf{Sigmoid}(\sigma z)$, with parameter $\sigma$ which controls the softness of the bins. When $\sigma$ is very large, the soft histogram approaches the hard histogram but the gradient vanishes, while small $\sigma$ yields a very soft histogram that badly approximates the true histogram. We use $\sigma = 1.85$ in our experiments. To compute a soft image histogram, we apply (4.4) to all values in one colour channel of an image, yielding three histograms. Now we can write the histogram loss as:

$$L_{\text{histogram}}(\mathbf{z}) = \|H(G(\mathbf{z})) - f_{\text{ID}\to\text{histogram}}(\mathbf{d})\|_2^2 . \tag{4.5}$$

**Image reconstruction** We now pose the image reconstruction problem as optimising the weighted sum of the ID, landmark and histogram losses:

$$\min_{\mathbf{z}} w_1 L_{\text{ID}}(\mathbf{z}) + w_2 L_{\text{landmarks}}(\mathbf{z}) + w_3 L_{\text{histogram}}(\mathbf{z}), \tag{4.6}$$

where we use $w_1 = 1$, $w_2 = 0.0006$ and $w_3 = 0.01$ in our experiments.

### 4.2.3 Qualitative results

We now present results of inversion from ID to image. We begin with an ablation study of ID-only inversion in Fig. 4.2. The results show that ID loss optimisation significantly improves over direct regression. The identities in the third column are clearly a better visual match to those in the first column. We then follow with an ablation study of our full inversion pipeline in Fig. 4.3. We show input images in the first column and results with various combinations of losses in columns 2-4. We initialise with our ID to GAN latent code regressor. Then we iteratively optimise only ID loss (column 2

- as in Section 4.2.1), ID loss and landmark loss (column 3) and all of ID, histogram and landmark losses (column 4). The result in column 2 convincingly reconstructs the ID of the original person but the pose and lighting are wrong. Introducing landmark loss largely corrects the pose (though note StyleGAN2 is biased towards frontal poses which means large pose angles are often underestimated). Introducing histogram loss yields similar lighting and skin tone producing an image similar to the original.

Next, we illustrate that our approach is capable of reconstructing different images of the same person under different conditions. In Fig. 4.4 we show pairs of real images of the same person in column one. From left to right, these exhibit different lighting, expression and pose. We show our full inversion result in column two. Even though both original images should yield the same ID descriptor, there is enough leaked information that we are able to convincingly reconstruct lighting, expression and pose.

Finally, we show additional inversion results in Fig. 4.5. The last row shows a failure case in which the pose is incorrectly reconstructed. This occurs when estimated landmark accuracy is low and is further compounded by the StyleGAN2 bias towards frontal faces.

### 4.2.4 Quantitative results

We quantitatively evaluate the reconstructed images, comparing them to the original images. To facilitate that, we calculate the mean squared error (MSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [125] between each reconstructed image and the original image, on the MoFA-Test dataset [118], containing 84 images and 78 identities. Table 4.1 shows an ablation study of reconstructing only with ID loss, with ID and landmark loss, and with all losses. It can be seen that overall, the extra

| | ID only | ID+LMs | ID+LMs+Hist |
|---|---|---|---|
| MSE (lower better) | 2.05 | 2.14 | 2.03 |
| PSNR (higher better) | 12.98 | 13.21 | 13.35 |
| SSIM (higher better) | 0.14 | 0.14 | 0.15 |

Table 4.1. Quantitative evaluation on MOFA-test, comparing reconstruction with only ID loss, ID and landmark loss, and ID, landmark and histogram loss.

losses help to recreate the actual image and not just the identity of a person.

Second, we test how well the reconstructed images (using all losses) preserve identity on the MoFA-Test dataset. We use cosine similarity on VGGFace [86] (as opposed to VGGFace2, which is used for our inversion) embeddings as a measure of how well a method was able to reconstruct the identity. Fig. 4.6 shows the distribution of similarity scores of our method, compared with Genova *et al.* [39], Tran *et al.* [121], and MoFA [118]. Note that these three methods solve a different problem: reconstruction with a 3D morphable model *given the original image*. However, Genova *et al.* [39] do this via an ID bottleneck meaning the comparison is meaningful. With an average similarity score of 0.77, we significantly outperform all other methods (0.40 for Genova *et al.*, 0.22 for Tran *et al.*, and 0.18 for MoFA). This is particularly notable given that we reconstruct the image *only from an ID vector*. The difference is likely partly down to using a generative model (StyleGAN2) that is much more powerful than a 3DMM.

## 4.3 Conclusion

In this chapter, we have demonstrated the feasibility of reconstructing facial images from ID descriptors using a generative model, specifically through the

process of inversion from ID descriptors to images. Our approach successfully reconstructs the identities of individuals from their face embeddings and also uncovers the potential to recover non-ID attributes, such as facial expressions and image histograms, that are encoded within these descriptors. The ability to reconstruct not only an image that matches the identity but also the actual original input image carries significant privacy and security implications.

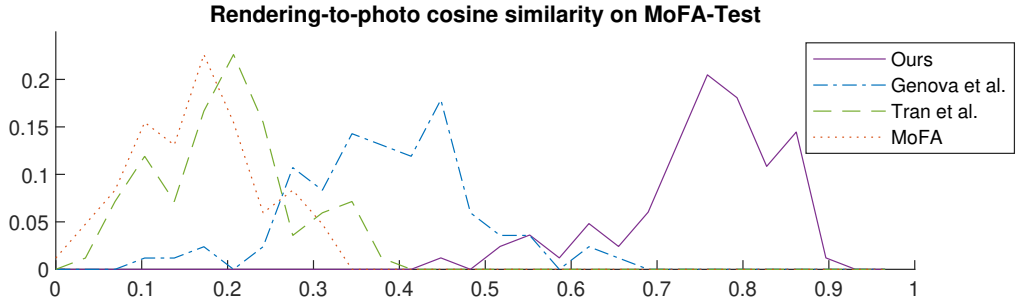



Figure 4.6. Distribution of VGGFace cosine similarity for MoFA-Test. We show the distribution of similarity scores of our method, Genova *et al.* [39], Tran *et al.* [121], and MoFA [118] for the original images and their corresponding reconstruction.

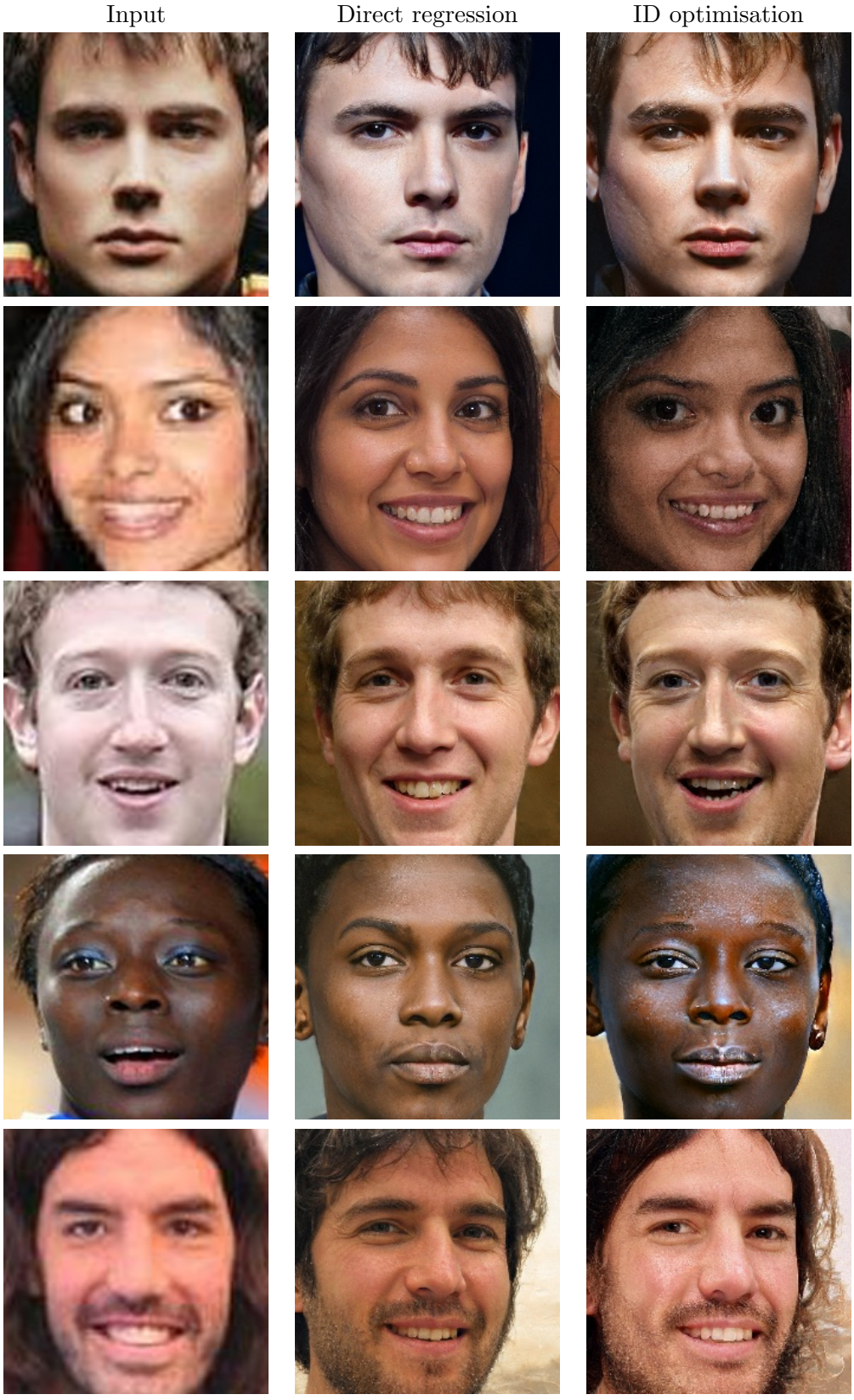



Figure 4.2. Comparison between direct regression and ID loss optimisation for StyleGAN2 latent code. Left column: Input images. Middle column: Output of ID to StyleGAN2 latent code regression network. Right column: After subsequent optimisation of StyleGAN2 latent code to minimise $L_{\text{ID}}$.

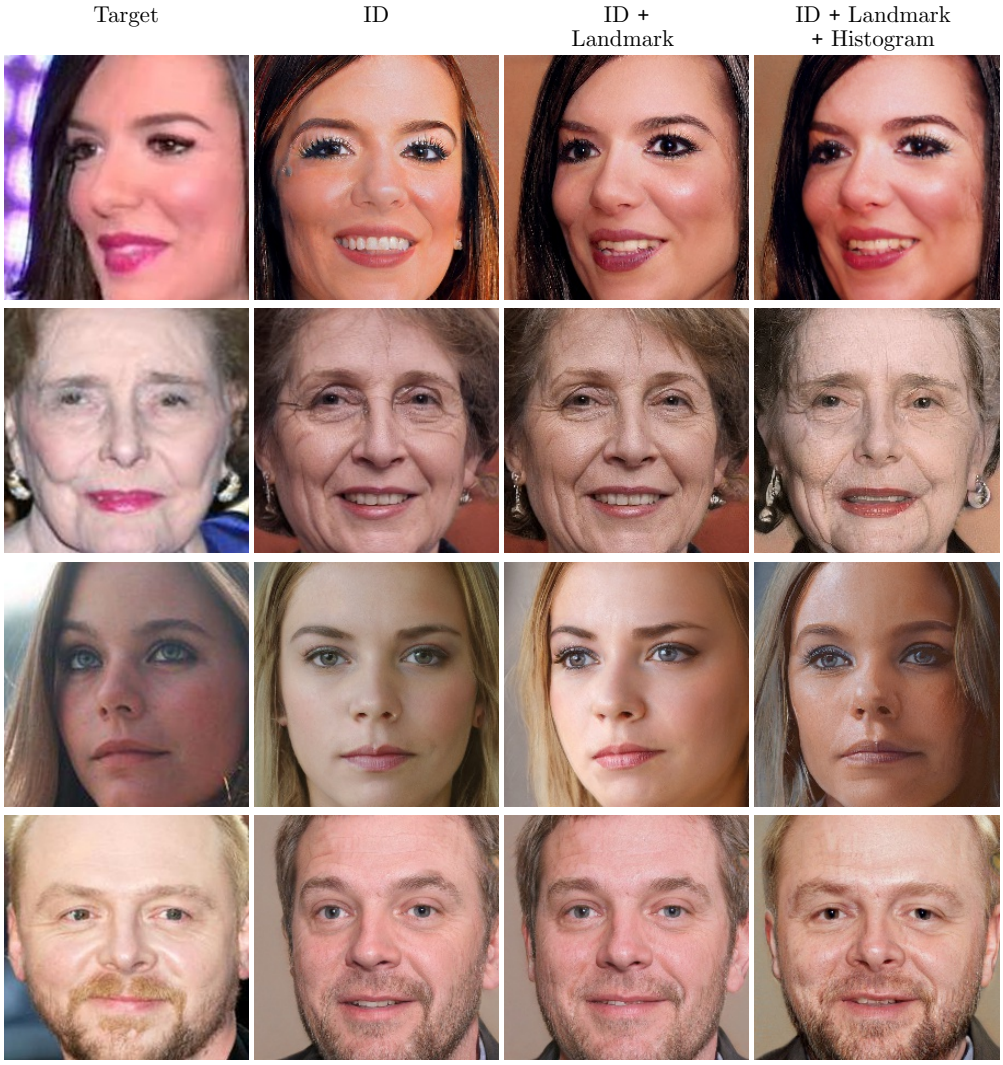



Figure 4.3. Ablation study. We show inversion results with only ID loss, ID and landmark losses and all three proposed losses.


Figure 4.4. Qualitative results for reconstruction for the same person under very different conditions (lighting, pose, expression). The first row labels indicate the type of image, and the left column labels describe the condition being varied for each set of images.

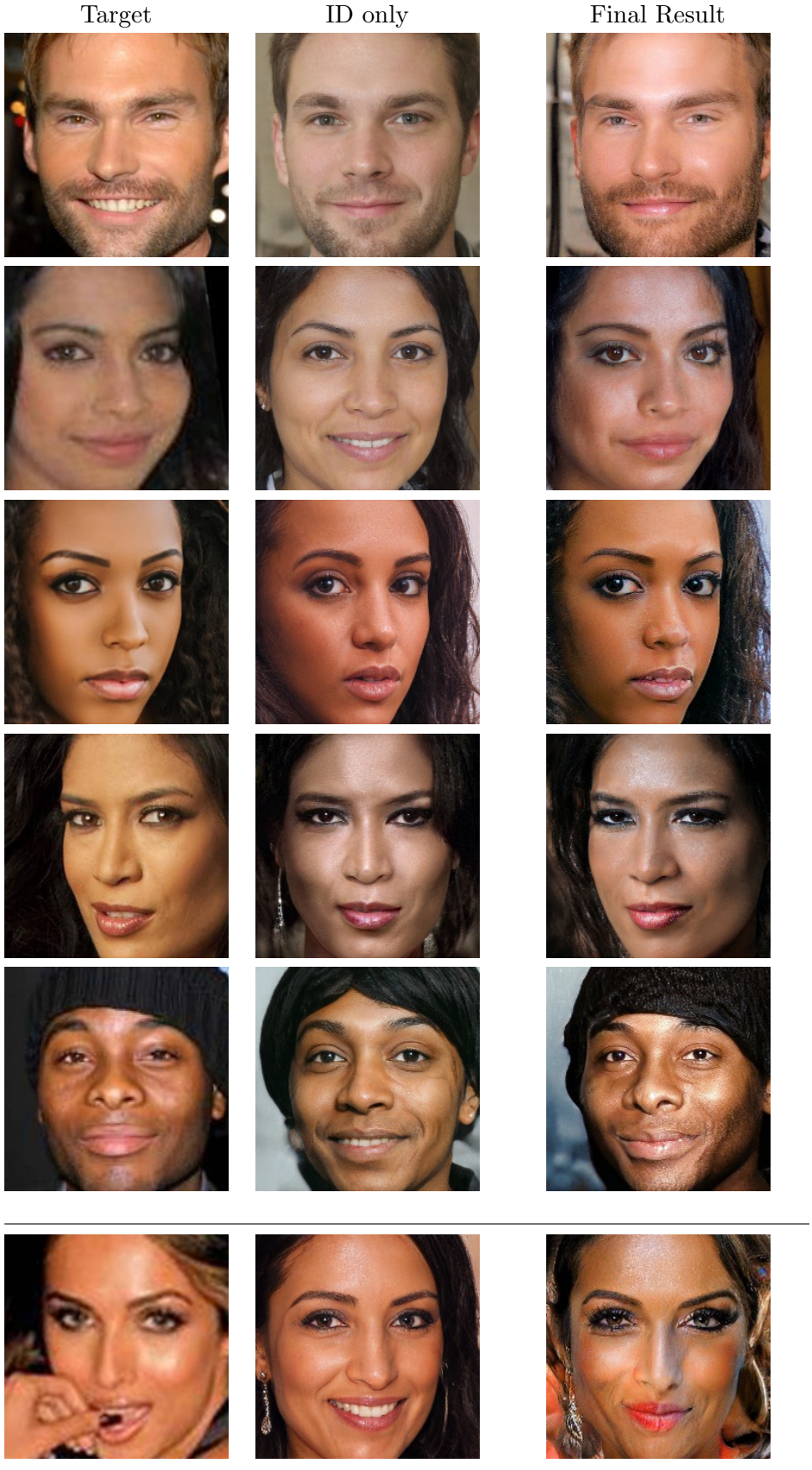



Figure 4.5. Additional inversion results. We show the original target image (left), reconstructions using only ID loss (middle), and full reconstruction result (right) for each set. The section below the separator line highlights a failure case.

# 5 Learning 3D alignment from recognition supervision

## 5.1 Introduction

The success of face recognition over the past 10 years has been dominated by data-driven approaches, relying on large CNN models and very large training datasets exhibiting large variations in pose, illumination, expression, background and so on. The assumption is that the CNN is able to disentangle these non-identity related factors from identity purely by volume of data and a powerful enough learning architecture. However, it is still interesting to explore whether explicit disentanglement of one or more of these features can assist face recognition and whether they can be separated using only a recognition signal. In this chapter, we consider the disentanglement of 3D geometry and pose from 2D appearance.

Spatial Transformer Networks (STNs), as introduced by Jaderberg *et al.* [53], offer a novel mechanism to perform spatial manipulations of data within the network. STNs can be incorporated into a neural network architecture, giving the network the ability to explicitly account for the effects of pose and nonrigid deformations, thus providing a robust framework for pose-invariant face recognition. Building upon this capability, this work proposes an innovative approach that extends the utility of STNs by investigating

whether 3D face alignment and reconstruction can be learned solely from a recognition signal, without the need for additional labels. This method aims to streamline the face fitting process and enhance the discriminative capacity of facial embeddings, even in pose-varied scenarios.

By focusing on spatial variations within identity embeddings, we seek to develop a more effective pose-invariant face recognition network. In this chapter, we will investigate two pivotal questions: *1. Can we learn to reconstruct the 3D shape of a face using only identity supervision? 2. Does reconstructing the 3D shape of a face enable recognition without spatial invariance?* Both of these questions are tackled by constructing a network in two parts. The first predicts a 3D shape and then uses this prediction to normalise the image to a pose-free space. The second performs recognition on the normalised image using a customised CNN architecture that does not have spatial invariance. The idea is that the lack of spatial invariance encourages the normalisation to align facial features to the same position in the normalised space. The only way to achieve this is via an accurate fit of the 3D model. Hence, we expect the recognition loss from the recognition network to provide a training signal to better align the 3D model to the 2D image. The answers to these questions could profoundly impact the future of face recognition technology, the amount of training data and the required complexity of the network can be significantly reduced.

Briefly stated, our main contributions in this chapter are as follows: First, we propose a 3D extension of the spatial transformer network that integrates a 3D morphable model and appropriate parameterisation of pose and shape. Second, we introduce a differentiable approximation of visibility so that we can correctly handle self-occlusion. Third, we propose a CNN architecture that does not exhibit spatial invariance. Finally, we combine these ingredients

in a phased training procedure and show that our model is able to learn 3D spatial alignment with only recognition supervision.

## 5.2 Method

### 5.2.1 Overview

In this section, we present a novel approach to face recognition that employs a 3DMM as an STN to achieve face alignment from a single image. The essence of our approach is to first estimate the 3D shape and orientation of a face using the 3DMM-STN framework, which then projects the 3D facial geometry onto a UV texture map. This texture map is subsequently utilized as the input for a face recognition network, thereby facilitating the extraction of robust facial features essential for recognition tasks.

A critical aspect of our method is the implementation of a visibility regressor designed to compute occlusion masks, which addresses the challenges of inaccuracies and artefacts in texture mapping due to self-occlusion commonly observed in profile views of the face. Fig. 5.2 illustrates a UV image that has been generated by projecting an input image using a 3DMM into a 2D UV map. Notably, distortion around the side facial areas results from these parts being invisible in the 2D source image, which leads to inaccuracies in the UV texture mapping. By accurately predicting which vertices of the 3D face model are occluded, the regressor facilitates selective masking of these areas on the UV map. This enhancement significantly improves the fidelity and accuracy of the resultant texture mapping.

Our approach leverages end-to-end training, optimizing STN parameters via backpropagation from the face embedding loss. Unlike many methods that rely on a landmark loss, which requires either costly landmark labels

from datasets or manual landmark detection prone to inaccuracies [35, 87, 16, 72], our process eliminates the need for such manual interventions, thereby simplifying the learning process. Requiring less data and exhibiting lower complexity than traditional methods, our system offers a streamlined and efficient solution for learning pose invariance.

Fig. 5.1 summarises our pipeline. The steps in the pipeline are as follows:

1. **3D Modeling and Texture Mapping:**

    - Employ a 3DMM-STN to estimate and project the 3D shape and orientation of a face from a 2D image onto a UV texture map.

2. **Visibility Handling with Regressor:**

    - Implement a visibility regressor to calculate occlusion masks and enhance the UV map by employing facial symmetry to address inaccuracies in texture mapping and fill in occluded or invisible areas.

3. **Face Recognition on UV Textures:**

    - Optimize STN parameters through backpropagation of face embedding loss, eliminating the need for explicit facial landmark detection.

    - Train the face recognition network directly on these enhanced UV facial textures, inherently achieving pose invariance and focusing on identity features rather than pose variability.

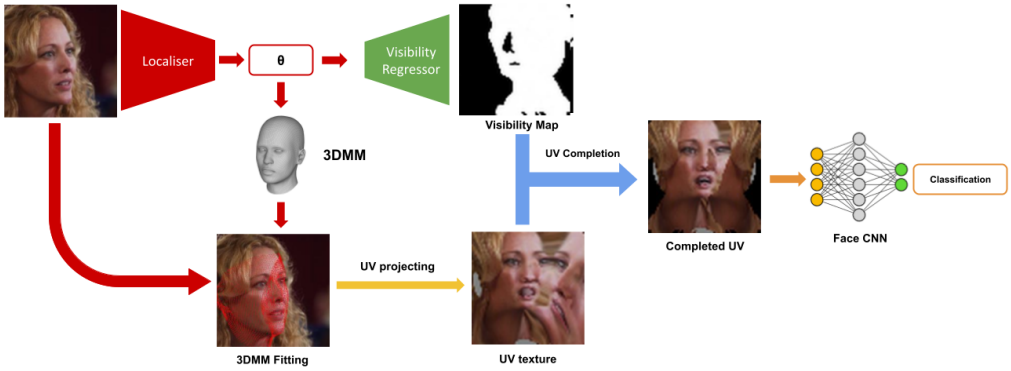



Figure 5.1. Overview of Our Face Recognition Pipeline. The localiser predicts 3DMM shape parameters and pose. Then 3D geometry is projected to 2D. A bilinear sampler then resamples the input image onto a regular output grid, which undergoes processing by our UV completion method to fill in pixels missing due to self-occlusion. Finally, this newly resampled image is fed into the face recognition network.

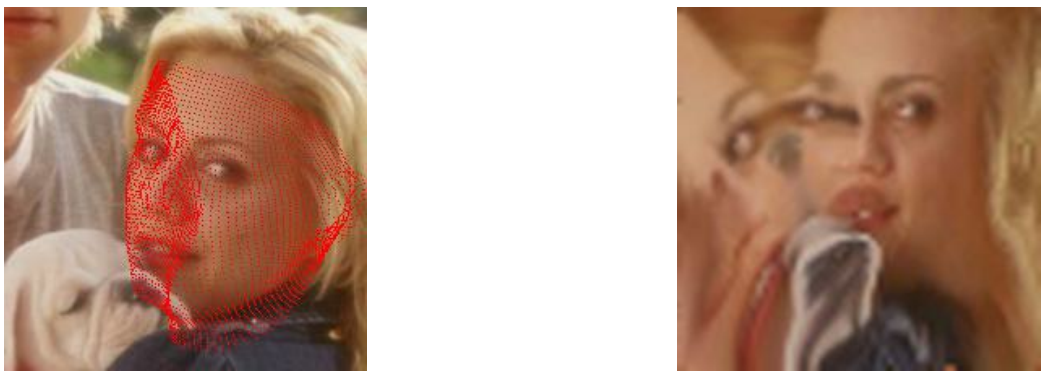
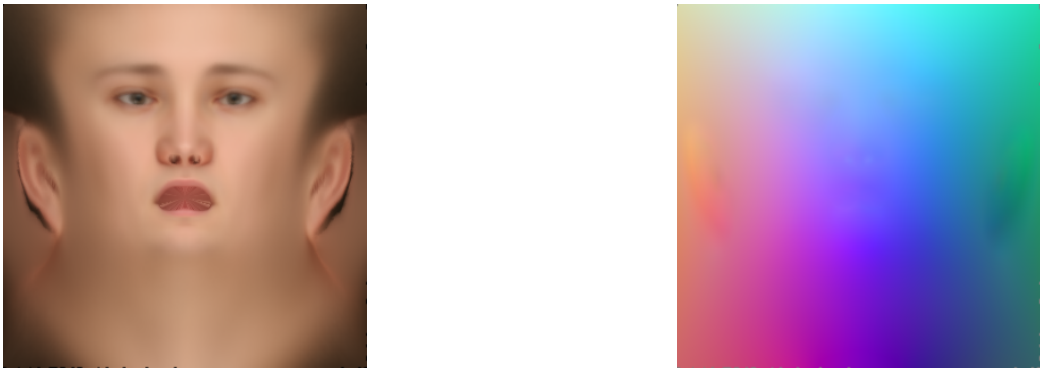


Figure 5.2. From left to right: Overlay of input image and aligned shape, (b) Sampled image in UV space



(a) (b)

Figure 5.3. (a) is the RGB albedo at UV coordinate, (b) is the 3D position of the model interpolated at UV coordinate

### 5.2.2 Projecting a 3DMM to UV and pixel space

The original STN introduced by Jaderberg *et al.* [53] enables spatial transformations on images or feature maps through a (sub-)differentiable module that takes 2D pixel or 3D voxel data as input, applies 2D or 3D transformations such as rotations and translations to a regular grid, and outputs a resampling of the input data onto the grid. The parameters of the transformation are estimated from the original input by a network called the *localiser*, the regular grid computed by a component known as the *grid generator* and the final output computed by a *bilinear sampler*.

Our proposed 3DMM-STN includes the same components with some differences. Similar to the original STN, we estimate transformation parameters with a localiser network. However, the localiser estimates a 3D transformation from 2D input data. The grid generator is replaced by the geometric model from a 3DMM. The sampler remains the same and is 2D since we sample the input 2D image data. Since we must also handle occlusion due to the projection from 3D to 2D, we introduce an extra component to compute per-vertex visibility. To enable the resampled output to be processed by a

conventional CNN, we flatten the 3DMM into a 2D UV space. In this way, the sampled 2D colours can be transferred via the 3D model from the 2D input into a 2D UV space output with aligned features.

This strategy ensures that the output image maintains an approximately uniform area with respect to the face shape, thereby obviating the need for the face recognition network to learn the complex invariants associated with different head poses. Consequently, this allows for the employment of a shallower MLP as the face CNN, as opposed to a deeper architecture. The UV mapping provides a consistent representation of facial features irrespective of the original head pose in the image, significantly reducing the network's need to learn pose variations. This consistency is crucial as it allows the network to focus more on distinguishing between different identities, potentially leading to faster convergence and improved generalization.

**3DMM** A face, as represented by a 3DMM, is represented as a 3D triangle mesh, yet the face surface itself is a 2D manifold embedded in 3D space. This dual nature allows us to define a mapping—commonly referred to as an embedding, flattening, or parameterization—of the face surface into 2D space, designated as UV space. One significant advantage of this approach is that both the 3D face and any textures on the face can be treated as 2D images, thereby facilitating processing using standard methods, such as 2D CNNs. We employ this UV space as a canonical, pose-free representation for faces, focusing solely on the geometric component of the 3DMM. The 3D morphable model layer generates a shape $\mathbf{X}$ composed of $N$ 3D vertices, achieved through a linear combination of $D$ basis shapes stored within the matrix $\mathbf{P}$, and the mean shape $\boldsymbol{\mu}$. This configuration is influenced by shape

parameters $\boldsymbol{\alpha}$, as described by:

$$X(\boldsymbol{\alpha})_{i,j} = x(\boldsymbol{\alpha})_{3(j-1)+i} \tag{5.1}$$

with $i$ ranging from 1 to 3, and $j$ ranging from 1 to $N$. Here, the vector $\mathbf{x}(\boldsymbol{\alpha})$ is defined as:

$$\mathbf{x}(\alpha) = \boldsymbol{\mu} + \mathbf{P}\boldsymbol{\alpha}. \tag{5.2}$$

### 5.2.3 3D localiser

The localiser network is a CNN that takes a face image as input and regresses the geometric parameters of a face. Specifically, the 3DMM shape (intrinsic) and 3D pose (extrinsic) parameters, which we refer to as $\theta$:

$$\theta = \left( \underbrace{R, t, s}_{\text{pose}}, \underbrace{\boldsymbol{\alpha}}_{shape} \right). \tag{5.3}$$

Here, $R \in SO(3)$ is a rotation, $t \in \mathbb{R}^2$ is the translation vector, $s$ is an orthographic scale and $\boldsymbol{\alpha}$ the 3DMM shape parameters.

To ensure the scale parameter $s$ remains positive, it is modelled as the exponent of its logarithmic estimate. Hence, the localiser output is treated as $\log s$ which is subsequently exponentiated to guarantee a positive scale. For the rotation, we regress the 9 values unconstrained and post-process it as described below.

For our localiser network, we initialise with pretrained ResNet-18 architecture, delete the classification layer and add a new fully connected layer with $9 + D$ outputs. Here the value of $D = 10$ corresponds to the weight of the first ten 3D basis shapes of the 3DMM.

**Mapping unconstrained matrices to rotation matrices** Given that directly regressed $\hat{R}$ provided by the localiser may not fulfil the orthogonality constraints required of a valid rotation matrix, a subsequent mapping step is required to project $\hat{R}$ onto the space of rotation matrices, $SO(3)$. We deal with this problem as the orthogonal Procrustes problem. This method finds the nearest orthogonal matrix to the input matrix, ensuring that the resultant matrix $R$ is a valid rotation matrix. Mathematically, this is achieved by solving the optimization problem:

$$\min_{R} \|R - \hat{R}\|_F. \tag{5.4}$$

subject to the constraint $R \in SO(3)$, where $\|\cdot\|_F$ denotes the Frobenius norm, and $SO(3)$ represents the set of $3 \times 3$ special orthogonal matrices.

The solution involves the use of Singular Value Decomposition (SVD). The SVD of $\hat{R}$ is given by:

$$\hat{R} = U\Sigma V^T \tag{5.5}$$

where $U$ and $V$ are orthogonal matrices, and $\Sigma$ is a diagonal matrix with singular values.

To project $\hat{R}$ onto $SO(3)$, the resulting matrix $R$ should also be orthogonal and $\det(R) = 1$ (indicating a proper rotation without reflection). The closest orthogonal matrix $R$ can be computed as follows:

$$R = UV^T$$

This mapping process is differentiable, which allows the backpropagation of gradients from loss functions defined on the rotation matrix $R$. This approach to rotation matrix regression has been shown to be the most stable [70].

### 5.2.4 Grid generator and sampler

The grid generator and sampling networks are engineered to transform an input image into a canonical, pose-normalized view by outputting sampled textures into a 2D embedding. The intensities extracted from this source image, based on the 3DMM, are then mapped to corresponding points within a flattened 2D grid. We utilize an architecture similar to that described in [8], combining a linear statistical model with a scaled orthographic projection as illustrated in Fig. 5.1. This type of projection is chosen because it preserves the spatial relationships by projecting 3D points onto a 2D plane without introducing perspective-related distortion, though our method could easily be extended to a perspective model. The transformation parameters $\theta$, estimated by the localiser network, dictate the alignment and scaling of the 3D model to the 2D grid. The orthographic projection of a 3D mesh vertex $\mathbf{p}_{3D}$ to a point in 2D image space is given by:

$$\mathbf{p}_{2D} = s \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \mathbf{R} \cdot \mathbf{p}_{3D} + \mathbf{t}, \tag{5.6}$$

where $s$ is the scale and $\mathbf{R}$ the rotation respectively, and $\mathbf{t}$ is the translation vector. This equation succinctly describes the transformation from 3D space to a normalized 2D plane.

For sampling image intensities from the source to the grid, we employ differentiable bilinear sampling. The intensity at a target location $(x_{ti}, y_{ti})$ in a color channel $c$ is calculated using:

$$V_i^c = \sum_{j=1}^{H} \sum_{k=1}^{W} I_{jk}^c \cdot \max(0, 1 - |x_i^s - k|) \cdot \max(0, 1 - |y_i^s - j|) \tag{5.7}$$

Here, $I_{jk}^c$ is the intensity at the $(j, k)$ coordinate of the input image, with $H$

and $W$ being the height and width of the image. The bilinear sampling is differentiable and thus suitable for gradient-based optimization techniques [53]. The result of this process is a 2D image, retaining the original dimensions of the input image, which captures the detailed textures mapped from the 3D mesh.

### 5.2.5 Model-based differentiable visibility

As implemented above, the bilinear sampler assigns a colour to every vertex in the 3D morphable model (and therefore every pixel in the UV warped image) regardless of whether that vertex is self-occluded. This introduces spurious content into the UV images that is unhelpful for recognition. For example, if a face is turned to the left so that its right side is visible, then pixels from the right-hand side of the face would be copied to occluded vertices on the left-hand side of the face. See for example Fig. 5.2.

Vertex visibility is a binary function. A vertex is either occluded or not occluded by another part of the face. This means that the visibility function is discontinuous and not differentiable at changes in occlusion. If we were to compute exact binary visibility (for example by rasterisation or ray casting) and use this within our framework, the recognition loss could not backpropagate through the visibility function and the network could not learn how changes in visibility influence the sampling and therefore recognition result. For this reason, we propose a novel, differentiable approximation to visibility using a neural field. The method is also very efficient requiring only a single forward pass through a small MLP (see Fig. 5.4).

**Neural field representation** The shape of the face is determined entirely by the estimated shape parameter, $\boldsymbol{\alpha}$. Visibility is also affected by the pose

of the face, captured by the estimated rotation matrix, $R$. However, only two of the three components of rotation affect visibility (in-plane rotations about the optical axis do not change vertex visibility). We therefore start by extracting the pitch and yaw rotation angles from $R$ as follows:

$$\theta = \arcsin(R_{21}) \tag{5.8}$$

$$\phi = \arctan\left(\frac{-R_{20}}{R_{22}}\right) \tag{5.9}$$

Vertex visibility can then be expressed as a function solely of $(\boldsymbol{\alpha}, \theta, \phi)$. We choose to represent this as a conditional neural field, i.e. a function $f_{\boldsymbol{\alpha},\theta,\phi} : [-1,1]^2 \rightarrow [0,1]$ that maps a position in UV space, $(u,v) \in [-1,1]^2$ to a soft visibility value in the range $0 \ldots 1$ (where a value of 0 represents full occlusion and 1 full visibility).

At inference time, we can pass the estimated shape parameters and rotation angles as conditioning and the UV coordinates of every point in our regular grid in order to estimate a UV space 2D visibility map.

**Neural field architecture** We implement the neural field using a MLP, adopting the 'conditioning-by-concatenation' approach [132]. This method involves stacking the input UV coordinates, shape parameters, and rotation angles into a $D+4$ dimensional vector, which is then fed into the MLP. The architecture of our MLP is based on SIREN [103], utilizing a sine activation function instead of ReLU. This design allows for processing all pixel coordinates for a visibility map in a single training iteration. During inference, a complete visibility map is generated in one forward pass, with the output being transformed via a sigmoid activation function to produce a visibility probability between 0 and 1.

**Visibility neural field training** We pretrain the neural field on synthetic data, subsequently freezing its weights during the training of the localizer and recognition networks. Despite the weights being frozen, gradients can still backpropagate through the visibility neural field, enabling the network to adapt based on changes in visibility. Practically, this adaptability is facilitated by softening the boundaries of the binary visibility masks. This softening creates a smooth transition between visible and non-visible areas within the neural field, enhancing the model's ability to handle varying visibility conditions effectively.

For the creation of our training dataset, we produced a sequence of headshot renders using a 3D renderer [63]. This process began with arbitrary values for pitch, yaw, and facial expressions to ensure a diverse range of facial perspectives. To determine the ground truth visibility of each vertex on the 3D model, we defined a ray in the view direction for each vertex and checked for intersections against all triangles of the mesh. A vertex was considered non-occluded if its corresponding ray did not intersect with any part of the mesh.

This dataset comprises 10,000 rendered facial images with varying angles, randomly generated to represent a diverse range of 3D facial shapes. Each image is annotated with the rotation angle of the 3D shape, which ranges from -90 to 90 degrees, and the visibility status of each 3D vertex. We trained our network using BCE loss, which involves comparing the network's predicted visibility map with the baseline visibility map. Examples of faces generated in our synthesized dataset are illustrated in Fig. 5.5.

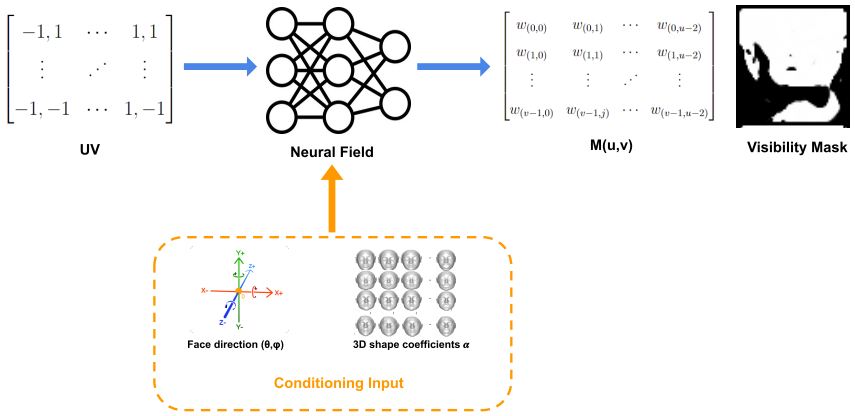
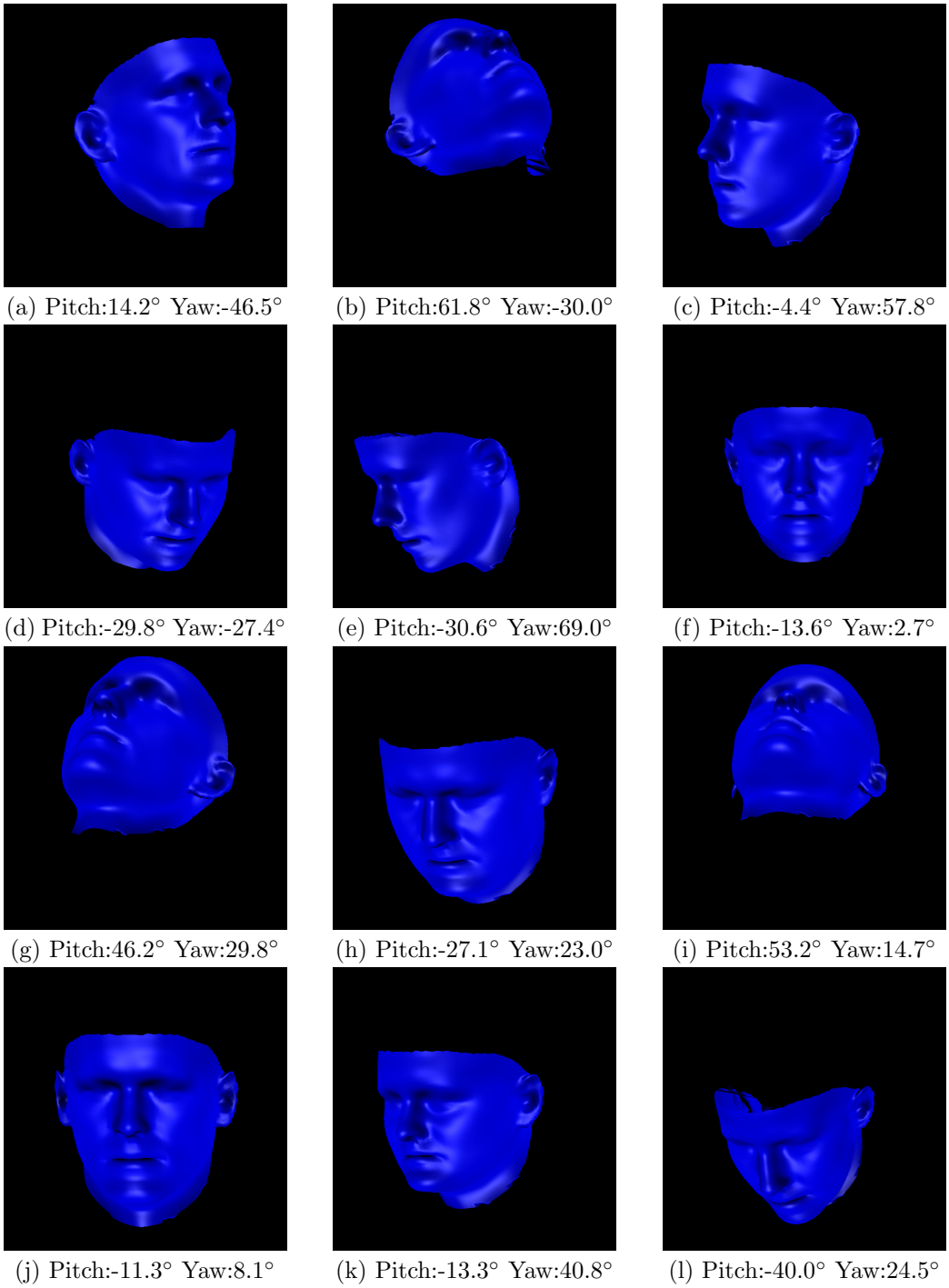



Figure 5.4. An overview of our neural field network. We decompose the output of the localiser to extract the pitch and yaw angles of the aligned 3DMM shape, as well as the coefficients of the 3DMM shape basis and the 2D coordinates of the UV map. These extracted parameters serve as inputs to our neural field. The synthesized visibility map $M$ is then compared with the corresponding ground truth map of the image to compute the Binary Cross-Entropy (BCE) loss. Note that our visibility map is differentiable, allowing the computed gradients to be utilized for back-propagation.

### 5.2.6 UV image completion

Once we have the estimated 3DMM shape and visibility map of the facial image, we can start to unwrap them to the UV coordinate to obtain a corresponding UV-map. Directly unwarped face textures lack face information in the invisible facial area owing to self-occlusion. Many recent work utilize deep encoder-decoder architectures to recover the facial texture from partial and masked facial images [26, 7, 35]. Our work aligns with this objective, aiming to preserve identity information in the unwrapped texture. However, we adopt a more straightforward approach by mirroring the face to fill in the texture of the occluded parts.

Given the visibility map predicted by the visibility regressor and the sampled UV texture, we first apply symmetry operations to both, creating a mirrored version of the face through horizontal flipping. Then, the visibility



(a) Pitch:14.2° Yaw:-46.5° (b) Pitch:61.8° Yaw:-30.0° (c) Pitch:-4.4° Yaw:57.8°

(d) Pitch:-29.8° Yaw:-27.4° (e) Pitch:-30.6° Yaw:69.0° (f) Pitch:-13.6° Yaw:2.7°

(g) Pitch:46.2° Yaw:29.8° (h) Pitch:-27.1° Yaw:23.0° (i) Pitch:53.2° Yaw:14.7°

(j) Pitch:-11.3° Yaw:8.1° (k) Pitch:-13.3° Yaw:40.8° (l) Pitch:-40.0° Yaw:24.5°

Figure 5.5. Examples of faces with 3D render generated at random angles

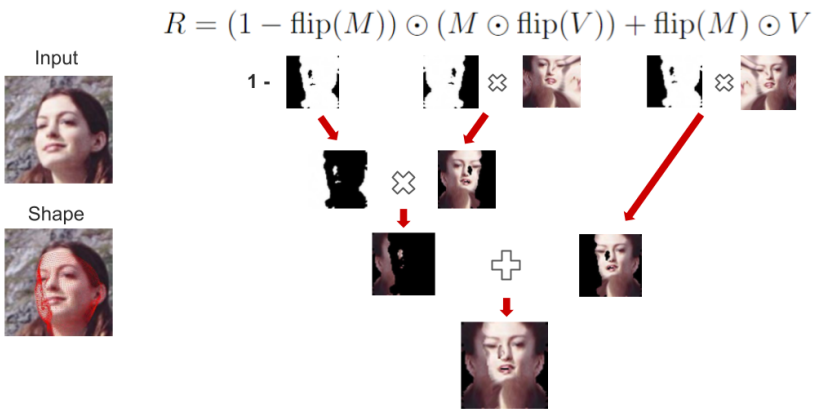



Figure 5.6. The illustration of UV image completion. Input image and the predicted shape and pose by the localizer shown in left. On the right, the results of our step-by-step visualization are shown. The final resampled image is presented at the bottom.

map is used to determine which parts are visible in the original image and which parts require completion using the mirrored image. Thus, the occluded or invisible parts can be pasted directly through their visible symmetrical counterparts. Given a resampled image $V$ and a visibility map $M$, the process of facial reconstruction in UV space can be formalized as follows:

$$R = (1 - \text{flip}(M)) \odot (M \odot \text{flip}(V)) + \text{flip}(M) \odot V, \tag{5.10}$$

where $R$ denotes the final completion UV re-sampled image. flip$(\cdot)$ represents the horizontal flipping operation, and $\odot$ signifies element-wise multiplication. Our pipeline is depicted in Fig. 5.6.

### 5.2.7 UV face recognition CNN

To provide a useful training signal to the localiser network, we must design our recognition network to not exhibit spatial invariance. In other words, if the 3D alignment (and therefore the image-to-UV space warping) changes

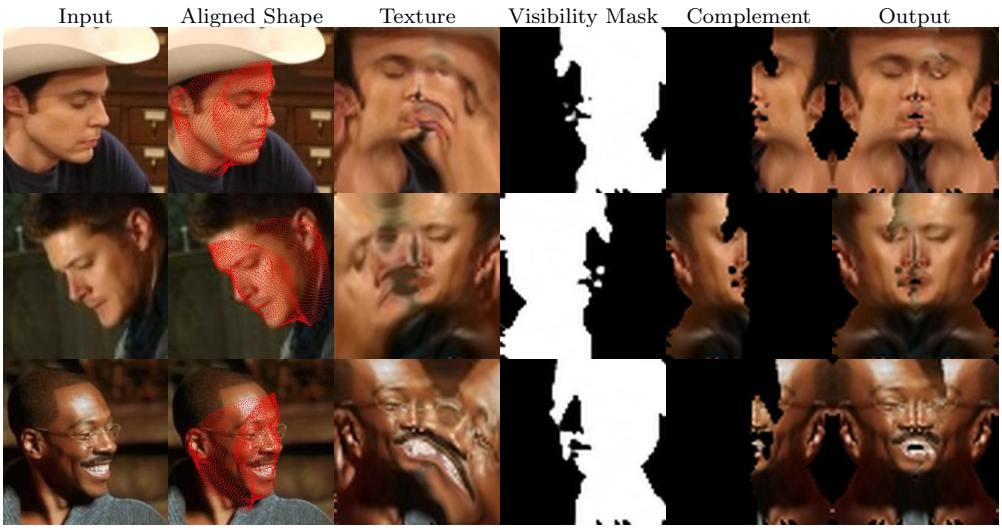



Figure 5.7. More examples for our UV completion. From left to right: input image; alignment shape; re-sample images; visibility masks; Complements used to fill in invisible areas; final re-sampled textures.

then the features extracted by the recognition network must change. This is in stark contrast to typical recognition CNNs that aim to achieve invariance to translations and perhaps other geometric transformations. This allows the network to robustly recognize faces irrespective of their position or scale within an image. In traditional CNNs, max pooling is used to reduce the spatial dimensions of feature maps [102]. This has a side effect of introducing spatial invariance: so long as the max feature lies within the same max pooling window, the output will be the same.

However, since retaining spatial features in face embeddings is pivotal in our method, we leverage spatial signals for face alignment. To preserve spatial features, our network adopts a different downsampling strategy. We use stridden convolution layers to replace traditional pooling layers and do not use fully connected layers at the end of the network (which can learn some transformation invariance) creating a purely convolutional architecture.

Our face recognition network processes a $64 \times 64$ UV texture map through

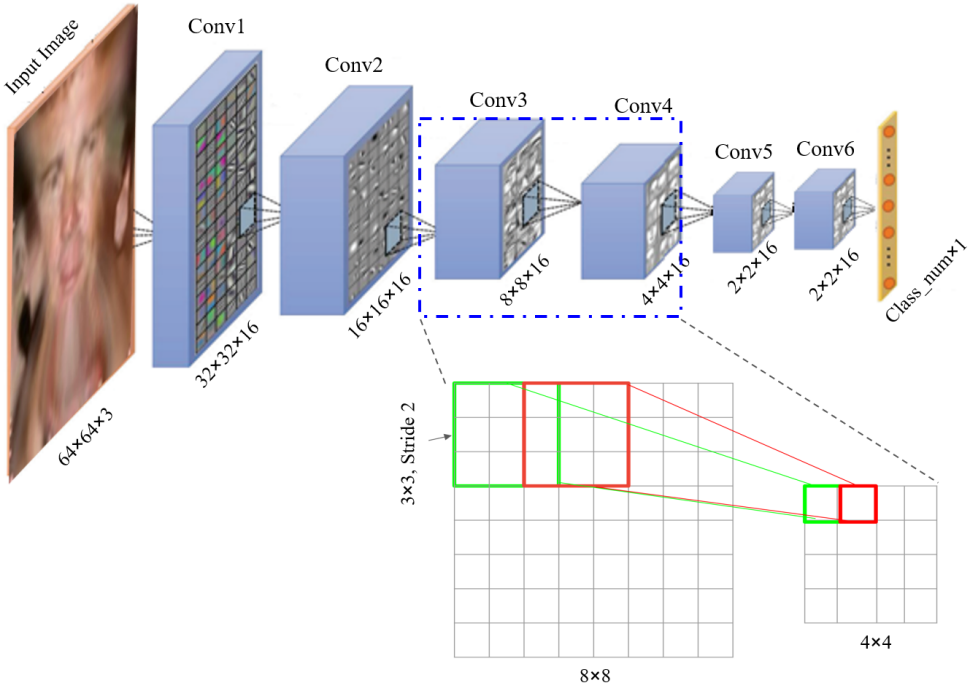



Figure 5.8. The architecture of our purely convolutional face recognition network. It intentionally removes spatial invariance that is conventionally introduced by max pooling layers. Features retain the spatial position until the final layer.

six convolutional layers. Each layer uses a $3 \times 3$ kernel with a stride of 2, successively reducing dimensions from $64 \times 64$ to $32 \times 32$, $16 \times 16$, $8 \times 8$, $4 \times 4$, and finally $2 \times 2$. A final convolution layer then collapses the spatial dimension to $1 \times 1$ to produce classification logits. Each layer uses a combination of stridden convolutions, batch normalization, and ReLU activations. By using stridden convolutions instead of traditional pooling layers, our network design retains more spatial information within the image while still reducing dimensionality. The full architecture of our face CNN is depicted in Fig. 5.8, which also illustrates the utilization of convolutional layers with a stride of 2 to reduce the size of features throughout the network.

**Initialisation** In our experiments, we observed that the training of the localizer is extremely sensitive to its initialization. Therefore, we pretrained the localizer with fixed $\theta_{\text{fixed}}$ parameters to standardize outputs to the mean face shape ($\boldsymbol{\alpha} = \mathbf{0}$), frontal views and a fixed scale, irrespective of the face angles in the input images (see in Fig. 5.9). The settings for $\theta_{\text{fixed}}$ include $R_{\text{fixed}} = I_3$, indicating no rotation; $t_{\text{fixed}} = [0, \frac{10}{112}]$; $s_{\text{fixed}} = \frac{\text{W}_{\text{input}}}{276}$, where $\text{W}_{\text{input}}$ is the width of the input image; and $\boldsymbol{\alpha}_{\text{fixed}} = \mathbf{0}$, ensuring the 3DMM initial shape. These configurations ensure that the initial face shape aligns well with the majority of faces in the dataset, providing a solid foundation for further training.

The initial weights, $\mathbf{W}_{\text{init}}$, are optimised using a loss function $L(\mathbf{y}, \hat{\mathbf{y}})$ designed to minimize the error between the target label $\mathbf{y}$ and predicted label $\hat{\mathbf{y}}$:

$$L(\mathbf{y}_x, \hat{\mathbf{y}}_x) = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{y}_x^i - \hat{\mathbf{y}}_x^i| \tag{5.11}$$

where $x$ can be $r, t, s, \alpha$ representing the respective parameter sets.

Another way to initialise the localiser to always output these standard parameters would be to initialise the bias of the final layer with the desired values. However, we found that this made the network very difficult to train. Instead training the network with real input images and the target pseudo-label ensures all layers of the network are suitably initialised to predict sensible parameters for any input.

The pretrained weights $\mathbf{W}_{\text{init}}$ were utilized as the initial weights for the localiser network, ensuring to stabilize the initial training. The whole localiser is then fine-tuned as part of the subsequent training.

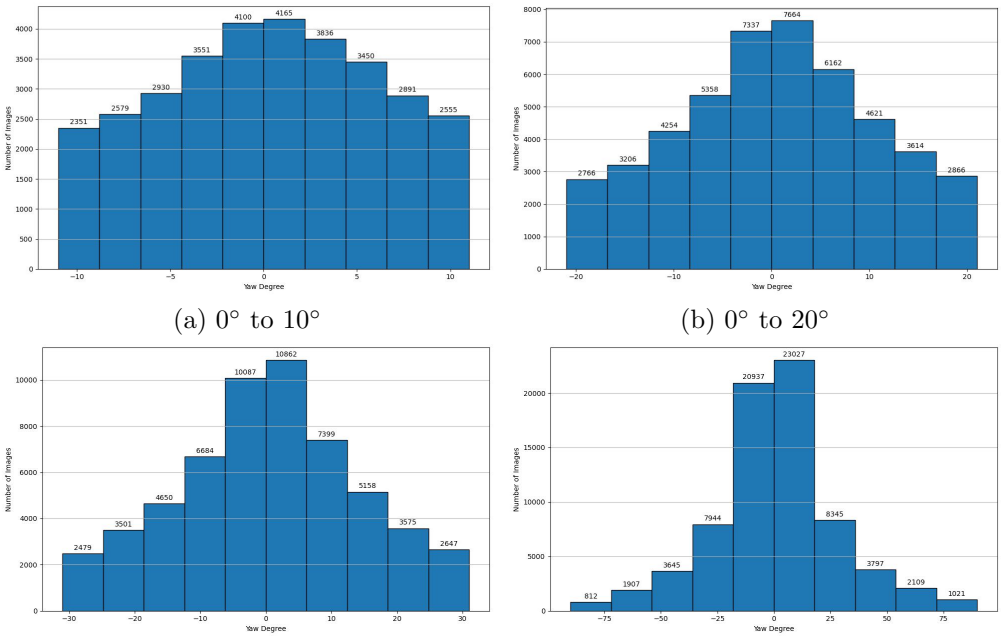
This formulation allows the training process to emphasize different aspects of the facial geometry by adjusting the weights for each parameter type,

Figure 5.9. Some examples of our adaptive initialization method.

thus tailoring the loss function to the specific requirements of the model's application.

### 5.2.8 Training Strategy

For training, we selected the CASIA-WebFace dataset [134] due to its established role as a benchmark in face recognition and its specific advantages for our task. CASIA-WebFace consists of 494,414 images from 10,575 individu-




(a) 0° to 10° (b) 0° to 20°

(c) 0° to 30° (d) Full dataset

Figure 5.10. Yaw angles distributions on CASIA-TOP200 with yaw distributing at: (a) 0° to 10°, (b) 0° to 20° , (c) 0° to 30° and (d) Full dataset

als, offering a large-scale dataset suitable for deep learning-based methods. In addition to its extensive size, the dataset provides a significant number of images per identity with considerable diversity in pose, expression, and viewing angles. This diversity is particularly beneficial for our method, as it facilitates the training of the 3D localiser by ensuring robust coverage of spatial variations. By leveraging these characteristics, the CASIA-WebFace dataset enables the proposed framework to effectively learn alignment features across different poses and expressions, which is critical for subsequent 3D alignment tasks.

However, when the estimated parameters from the localiser are highly inaccurate, sampling to UV space becomes ineffective, providing no meaningful supervision through the recognition loss to improve parameter estimation. Despite the diversity of CASIA-WebFace, training directly on a dataset

with varied facial angles often leads to tricky convergence for several reasons: (1) The amount of training required for facial recognition and alignment to converge differs. (2) Training the entire network from scratch can cause oscillations in the learning process. (3) Challenging facial poses made the network prone to falling into local minima without proper initialization.

Therefore, it was essential to warm-up the network and incrementally train it with increasingly challenging samples. To ensure a balanced ratio of data for facial recognition and alignment, we first selected the top 200 subjects with the highest sample volume from the CASIA-WebFace as our new dataset. We then use 6DRepNet [47] to determine the yaw angles of each facial image in the CASIA-WebFace dataset, categorizing it into subsets of [0°,10°], [0°,20°], [0°,30°]. In the warm-up phase, we initially trained our facial recognition module on the [0°,10°] subset. Subsequently, we trained our complete network progressively on [0°,10°], [0°,20°] and [0°,30°] subset. Our dataset comprises 7,3545 images from 200 subjects, 3,2409 images in [0°,10°], 4,7849 images in [0°,20°] and 5,7043 images in [0°,30°]. Fig. 5.10 illustrates the pose distribution of our datasets.

### 5.2.9 Losses

Our loss function contains two terms: a classification loss to provide the recognition signal and a shape regularization term to ensure the predicted 3DMM geometry remains plausible.

**Recognition loss** The flattened UV map of a face is passed to the recognition CNN. This computes logits for each of the face identities in the training set. We compute the cross entropy loss for these logits against the identity

labels as:

$$L_{id} = -\sum_{c=1}^{C} y_c \log(\hat{p}_c) \tag{5.12}$$

where $C$ denotes the total number of identity classes in the dataset. $y_c$ is a binary indicator (0 or 1) signifying whether class label $c$ is the correct classification. $\hat{p}_c$ represents the output of the final convolutional layer for class $c$.

**Statistical regularisation loss** To prevent unreasonable deformations of the 3D Morphable Model during training, we incorporate a shape regularization loss function. This loss function is designed to constrain the magnitude of shape basis parameters $\alpha$, thereby the structural integrity of the 3D model within reasonable bounds. Specifically, the shape regularization loss is mathematically formulated as follows:

$$L_{reg} = \lambda \cdot \|\alpha\|_2^2 \tag{5.13}$$

where, $\lambda$ represents the weight that adjusts the strength of the regularization effect; shape_basis_params denotes the shape basis parameters of the 3DMM; By penalizing the square of the L2 norm of the 3DMM coefficients, this loss function encourages the network to maintain reasonable estimations of the 3DMM shape during training.

**Gradient flow** The recognition loss is backpropagated through the recognition CNN and into the output of the STN. Via the differentiable sampler, the gradient is backpropagated through the estimated 3DMM parameters and into the localiser. In this way, the localiser seeks to modify its estimated parameters such that the recognition CNN produces a higher probability for the correct identity. Our assumption is that it achieves this by encourag-

Figure 5.11. Qualitative Results of the Visibility Regressor. The left grids of images show the input face images with the overlaid 3D face shapes rendered using the estimated shape and pose. The right grid displays the corresponding visibility masks predicted by the visibility neural field.

ing better spatial alignment of the image features. The recognition loss also backpropagates into the visibility network. Points that are close to a visibility boundary have intermediate visibility which provides a gradient to either hide or uncover a feature in the image as deemed useful by the recognition network.

The regularisation loss directly backpropagates into the localiser and encourages conservative shape parameter estimates, discouraging large, unlikely values.

## 5.3 Experimental Results

### 5.3.1 Quantitative Results

**Landmark error** To evaluate the face alignment performance of our 3D Morphable Model Spatial Transformer Network (3DMM-STN), we utilize the landmark error, which serves as a proxy for the quality of fit. We use a

face landmark detection [10] to locate 68 landmarks from the input image, where we select 14 landmarks that most comprehensively summarize the facial features. These provide our ground truth. The predicted landmarks, on the other hand, are derived from the 3DMM by selecting the 14 vertices that best correspond to these 14 selected facial landmarks (see Fig. 5.14). The loss is calculated using the mean Euclidean distance between the two sets of landmarks:

$$L_{\text{landmark}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{v}_i - \mathbf{P}_i\|_2, \tag{5.14}$$

where $N = 14$ represents the number of selected landmarks. Here, $\mathbf{v}_i$ denotes the coordinates of the $i$-th predicted vertex within the 3DMM, designed to correspond to the facial landmarks. $\mathbf{P}_i$ represents the coordinates of the $i$-th ground truth landmark obtained from the image.

It is important to note that this landmark loss is utilized exclusively for performance evaluation and is not incorporated into the training process of the network. When the value is low (or reducing) we assume that this corresponds to a good (or improving) fit. If it increases, we assume this corresponds to the fit becoming worse.

**Average image sharpness metric** Besides utilizing landmark loss to assess the face alignment performance of our method, we also introduce the Average Image Sharpness Metric (AISM). This metric evaluates the sharpness of the mean average of all of the UV re-sampled images, acting as an indicator for determining if the 3DMM-STN is effectively learning to predict the correspondence between the model and the images throughout training. An increase in the sharpness of the average face—evidenced by rising gradient magnitude values across epochs—indicates an improvement in the localizer's ability to accurately align faces within the images. When faces are

misaligned, features appear at different locations in UV space and become blurred, leading to a less sharp image.

Given the average UV image $\overline{V}$, the gradient components $G_x$ and $G_y$ are computed using the Sobel operators $S_x$ and $S_y$:

$$G_x = S_x * \overline{V}, \quad G_y = S_y * \overline{V} \tag{5.15}$$

The gradient magnitude at each pixel, indicating the sharpness of edges, is calculated as:

$$G = \sqrt{G_x^2 + G_y^2} \tag{5.16}$$

The AISM is quantified as the mean of the gradient magnitudes across the entire image:

$$\text{AISM} = \frac{1}{N} \sum_{p=1}^{h \cdot w} G_p \tag{5.17}$$

where $N$ is the total number of pixels in $\overline{V}$. A higher AISM represents more precisely aligned.

In Fig. 5.15a, we present the recognition loss and accuracy curves observed during our phased training on three subsets of the CASIA dataset, divided according to their absolute yaw angles: [0°-10°], [0°-20°], and [0°-30°]. Initially, we train our model from scratch on the [0°-10°] subset, subsequently utilizing the learned weights for further training on the [0°-20°] and [0°-30°] subsets. Fig. 5.15b and Fig. 5.15c display the sharpness and landmark accuracy metric curves, respectively. We also show the final loss and metrics values for each phase in Table 3.1. In Fig. 5.12, we present a comparison of mean resampled images before and after the training phases. This figure illustrates the changes in UV-resampled images across the three subsets.

In [0°-10°] and [0°-20°] stages, a notable correlation is observed between

the reduction in loss and improvements in sharpness metrics, while landmark errors concurrently decrease. However, this trend shifts when introducing more challenging samples with larger pose variations in the [0°-30°] stage. Here, the progression of loss reduction, increase in sharpness metrics and decrease in landmark errors noticeably slows down. This behaviour is visually corroborated in Fig. 5.12, where the average resampled face images transition from initially blurred to progressively clearer facial features during the [0°-10°] and [0°-20°] phases, with the greatest enhancement in facial clarity evident within the [0°-10°] phase. In the [0°-30°] phase, improvements in clarity become minimal, suggesting that the model's ability to learn face alignment faces challenges with the incorporation of samples featuring extensive pose variations.

### 5.3.2 Qualitative results

Fig. 5.16 shows qualitative fitting results. In the first row, we show input images. In the second we show the overlay of the estimated 3D morphable model geometry. Note the good alignment between the model and image and that the 3D rotation is well estimated. In the third row, we show the UV resampled images before visibility masking. Note that facial features are consistently mapped to UV space (e.g. the eyes, nose and mouth always map to approximately the same positions). In the fourth row, we show the prediction of the visibility regressor (please note that they are horizontally flipped relative to the UV images). These maps correctly predict occlusions on the side of the face when it is rotated out of the plane. Finally, in the bottom row, we show the completed UV maps using symmetry information. These normalised images are suitable for recognition with a spatially-dependent representation (since all features are aligned).

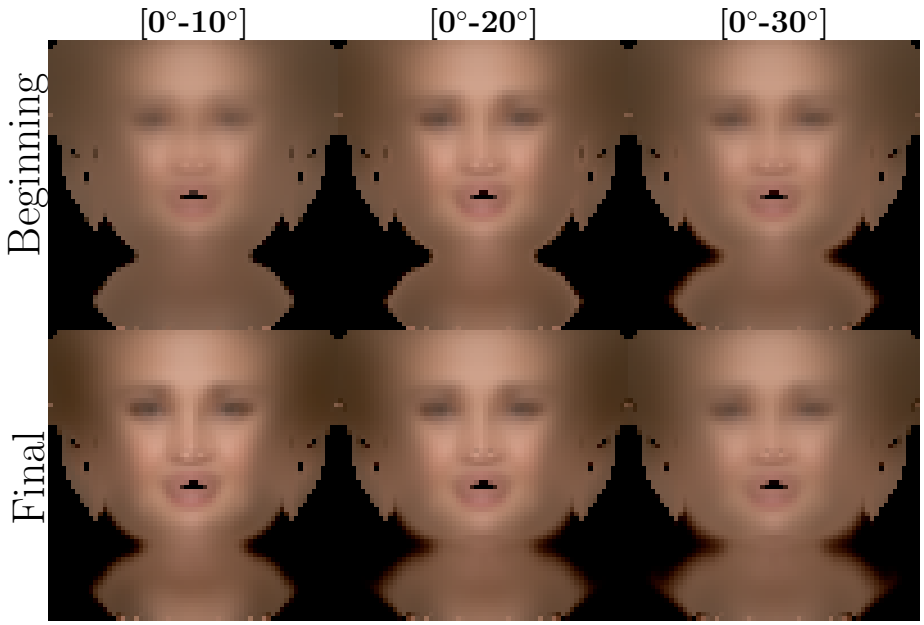



Figure 5.12. Comparison of mean resampled images before and after training across different phases of the proposed phased training approach. The columns represent training phases corresponding to subsets of the CASIA dataset, divided by absolute yaw angles: [0°-10°], [0°-20°], and [0°-30°]. Each phase shows the mean UV-resampled face images before training (top row) and the corresponding mean images after training (bottom row). This figure illustrates the progressive improvements in face clarity and alignment across the different training phases.

In Fig. 5.18, we further investigate our facial alignment accuracy by comparing the ground truth of 14 selected landmarks with those predicted by our 3DMM-STN. There is in general a good agreement, however, accuracy degrades with larger pose angles. Fig. 5.17 illustrates failure cases where misalignments have caused distortions in the resampled images, leading to training instabilities. Our method still faces challenges in situations involving complex shadows and occlusions, extreme poses, and extreme lighting conditions.

**Phased training** Fig. 5.13 presents an ablation study that examines the effectiveness of phased training strategies. In scenario (a), a model trained

(a) (b)

Figure 5.13. Ablation Study: (a) Results from directly training a pretrained model on a 0°-30° dataset after initial training on a 0°-10° set. (b) Results from progressively training the same model on 0°-10°, 0°-20°, and 0°-30° datasets.

| | CASIA-TOP200 Dataset | | | |
|---|---|---|---|---|
| Metrics | initial | $[0° - 10°]$ | $[0° - 20°]$ | $[0° - 30°]$ |
| Loss (↓) | 5.3910 | 0.1562 | 0.1850 | 0.1984 |
| Landmark Loss (↓) | 10.0850 | 7.6871 | 7.3666 | 7.5357 |
| Gradient Magnitude Loss(↑) | 0.1817 | 0.1845 | 0.1814 | 0.1791 |

Table 5.1. The training loss, landmark loss, and sharpness loss on CASIA-TOP200.

from scratch on the 0°-10° training set is directly transferred and further trained on the 0°-30° training set. Scenario (b) represents the results from a model that undergoes gradual training expansion, starting from the 0°-10° training set, then extending to the 0°-20°, and finally the 0°-30° training sets. It is evident that phased training leads to a model better able to fit to large pose angles whereas the model with reduced phasing fails to learn to fit in several cases. The findings emphasize the necessity of adopting a phased training strategy to achieve better results.

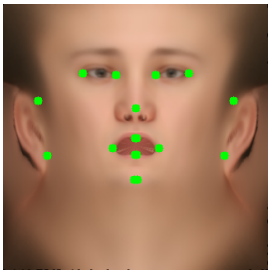


Figure 5.14. 14 selected landmarks, these points are crucial for effectively capturing the variations in key facial features. Each landmark is chosen to represent significant anatomical regions on the face

## 5.4 Conclusion

In this chapter, we have shown how to incorporate a 3DMM within an STN framework. We have shown that training the STN end-to-end with a recognition network that does not exhibit spatial invariance allows the network to learn 3D face alignment entirely from a recognition signal without any direct supervision on the 3D face geometry, or the 3D-2D alignment. Of particular interest is the balance between the size of the localiser and recognition networks. In typical face recognition CNNs (that comprise only a single network), the network must learn both feature extraction and alignment within a single network. This usually requires very large networks with 10s of millions of parameters. In our work, we have shown that the pre-alignment of the input UV images means the recognition network can be very small: only 22,888 for the architecture we use. Our localiser network is a ResNet-18 which contains 11 million parameters. This provides evidence that the 3D alignment is the harder part of the task. Once the features are well-aligned in a 2D space, the task becomes much easier and a relatively shallow CNN can extract discriminative, spatial features.

Our work not only enhances the understanding of pose-invariant face recognition but also proposes a potential framework for face recognition tech-

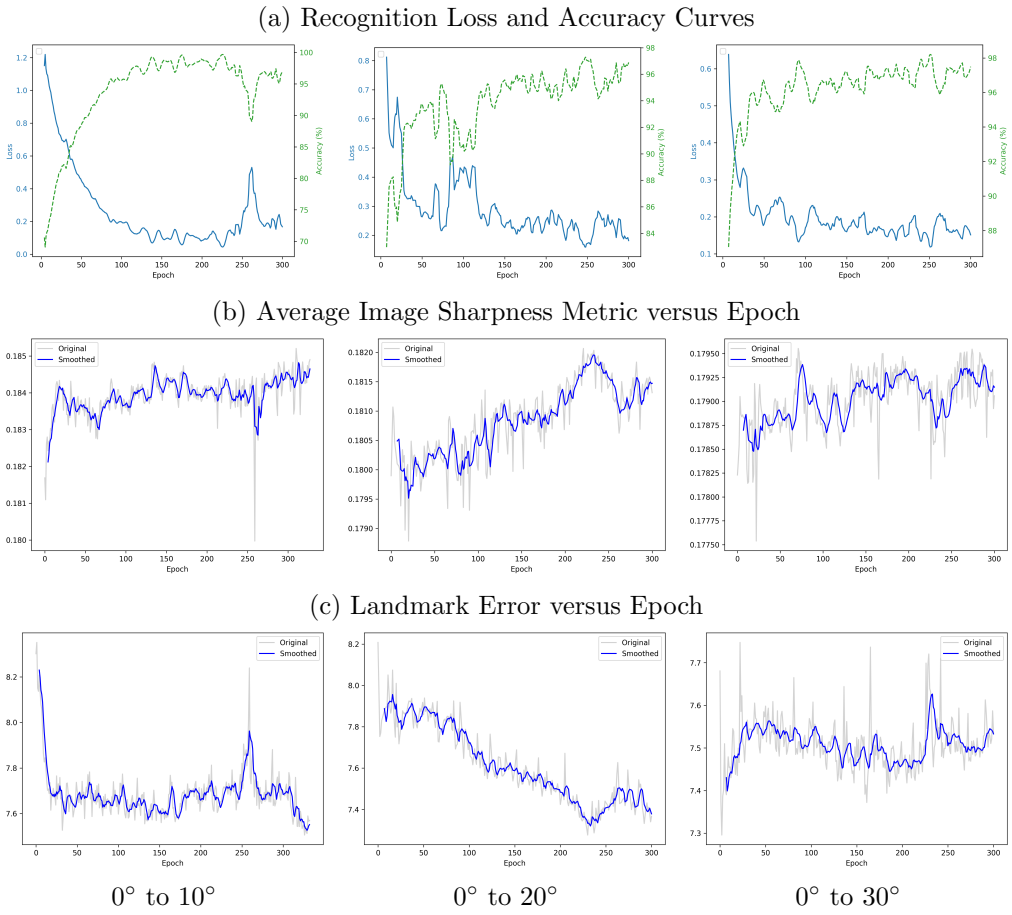



Figure 5.15. Training loss and evaluation metrics versus epoch on CASIA dataset with yaw distribution at: [0° - 10°], [0° - 20°] and [0° - 30°]. (a) Training Loss and Accuracy Curves, (b) the Average Image Sharpness Metric versus Epoch, and (c) Landmark Error versus Epoch.

nologies by reducing dependence on large-scale annotated datasets and complex network architectures. While our pipeline can achieve geometry label-free alignment, our system still encounters challenges in scenarios involving extreme poses, complex shadows, and severe lighting conditions.

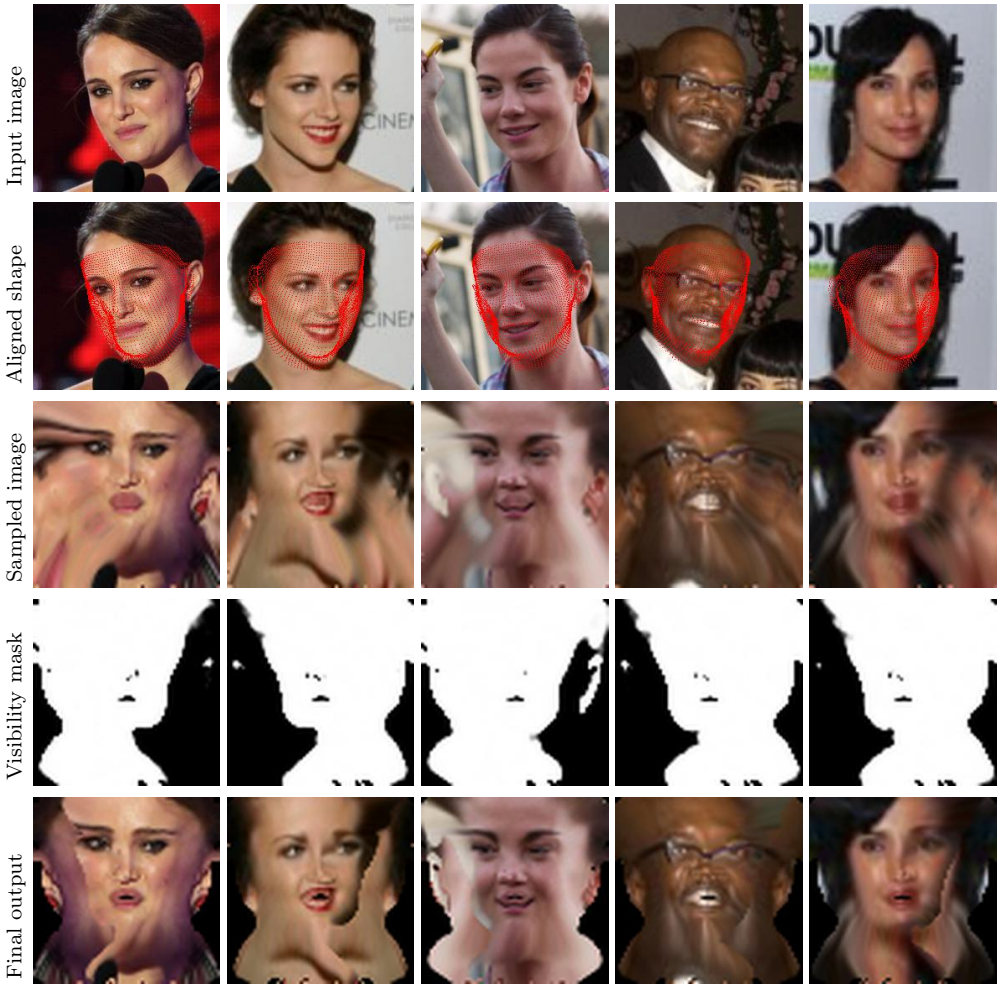



Figure 5.16. Result of 3DMM-STN

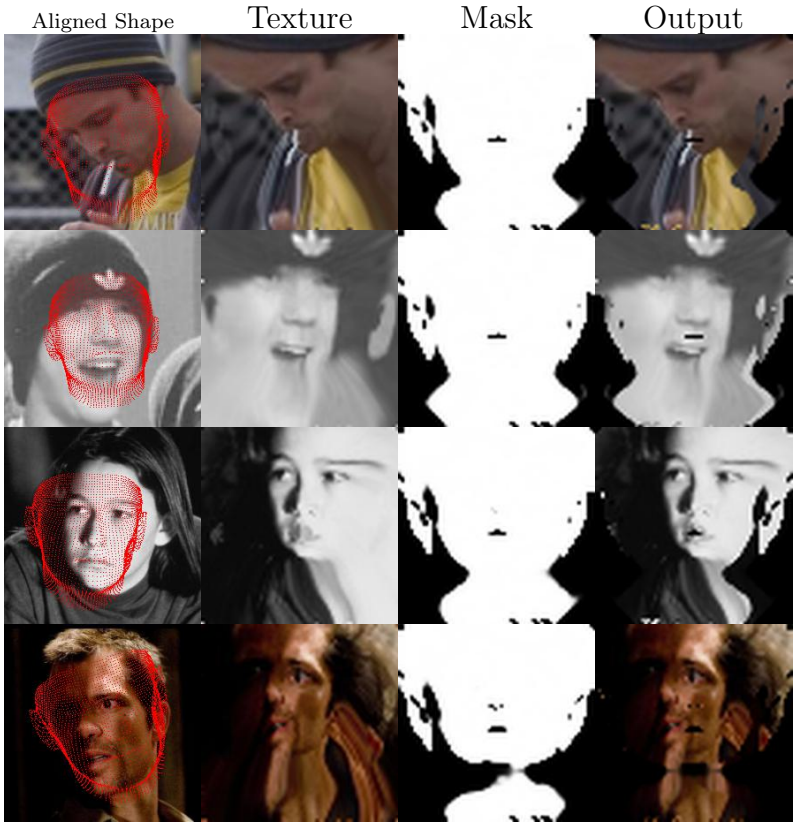



Figure 5.17. Failures cases. From left to right: alignment shape; re-sample images; visibility masks; final re-sampled textures.

Figure 5.18. Examples show that our face alignment accuracy via landmarks. Green: ground truth landmarks from [10]. Red: predicted landmarks by our method

# 6 Conclusions

In this chapter, we summarise what has been achieved and draw overarching conclusions that can be taken from the work conducted in this thesis. Finally, we discuss the potential future work, that can build upon the work that has been presented in this thesis.

This thesis explores the potential interplay between facial reconstruction and recognition, focusing on the distinction between identity-specific and non-identity information in deep face representations. Chapter 3 examines the extent to which non-ID information is unintentionally retained within ID embeddings. Contrary to the intended design, our findings reveal that non-ID attributes are encoded alongside identity features, posing potential risks to user privacy and system integrity. By training a specific MLP to map ID descriptors to target attributes, our method has demonstrated that non-identity information—including expressions, whether hats or glasses are worn—along with pose and lighting conditions, can be accurately predicted from face descriptors (Section 3.2). To mitigate these non-ID attributes from face descriptors, in Section 3.3, we introduced an adversarial training methodology aimed at mitigating non-ID attributes, facial expression, from the Arcface descriptors. By integrating our adversarial learning method with the ArcFace framework and training on the VGGFace2 and MS1MV3 datasets, our model achieves higher accuracy on the IJB-C dataset compared

to the original methods.

Chapter 4 examines the capabilities of generative models, with a specific focus on StyleGAN2, to reconstruct facial images from embeddings. Section 4.2.1 validates the feasibility of using StyleGAN2 for reconstructing facial images directly from ID embeddings. Further elaborated in Section 4.2.2, with adequate supervision and training, it is possible not only to reconstruct the identities of individuals from their face embeddings but also to capture potential non-ID attributes such as facial expressions and image histograms embedded within these descriptors.

In Chapter 5, we introduce an innovative approach to face recognition that involves reconstructing 3D facial geometry using only identity signals, without the need for additional geometric labels. This method, named 3DMM-STN and detailed in Section 5.2.2, integrates a 3DMM with a STN. Central to this approach is the localizer network, adept at learning 3D shape and pose parameters exclusively from identity-related data. In Section 5.2.6, we detail our facial UV completion method, specifically designed to address the challenge of missing pixels due to self-occlusion. Following this, Section 5.2.7 describes our face recognition network, which processes the UV-resampled images as inputs. This network is composed entirely of convolutional layers and eschews traditional pooling layers to preserve a greater amount of spatial information within the face descriptors.

## 6.1 Conclusions

**Non-ID information leaks into face embeddings** We have presented an unexpected conclusion that contradicts common assumptions in face recognition techniques, revealing that not only identity but also non-identity at-

tributes can be recovered from ID embeddings created by deep CNNs. Despite the usual objective of deep CNN embeddings to capture solely identity-specific information, discarding changeable aspects of the face or environment, we found this is not the case. In-depth analysis of state-of-the-art face embedding networks, VGGFace2 and ArcFace, showed that non-identity attributes, such as landmark positions (which reflect pose and expression) and the image histogram, could be recovered from the ID embedding. Although landmarks also encode intrinsic identity-specific information like facial shape, our work primarily focus their role in capturing non-ID features. In fact, these attributes can be predicted from ID embeddings with comparable accuracy to predictions from the original image. This is a surprising conclusion that may have implications beyond faces and face recognition about unintentional leakage of information within learnt representations.

**Reducing non-ID information leakage improves face recognition performance** Non-ID information is a distraction for face recognition that can introduce unwanted bias. For example, if a person is smiling in all of their training images, then embedding the non-ID characteristic "smiling" into the ID vector appears useful at training time. However, if at test time the same person is not smiling, performance will be degraded. This is essentially the common observation that differences in the distribution of data between training and test sets will hamper performance. However, our conclusion is that this can be explicitly improved if we have knowledge about the distractor characteristics. Knowing a prior that expression is not useful for recognition means that we can supervise our encoder adversarially to avoid embedding such information. Again, this conclusion may extend to other objects besides faces and other tasks besides recognition.

**Input images can be reconstructed from ID embeddings** We fur-

ther conclude that photo-realistic images can be reconstructed from ID embeddings by utilising an optimization strategy using StyleGAN2, a generative model, enabling inversion back to the original image, including details like pose and lighting. These findings challenge existing paradigms and open up new considerations in the security and usage of facial recognition systems.

**ID supervision is sufficient to learn 3D alignment** Our final conclusion is that there is enough information in a recognition supervisory signal to learn to reconstruct the 3D shape of a face or, equivalently, to align a 3D model to an image. By posing the recognition problem as one of alignment followed by alignment-sensitive recognition, we can learn alignment without any explicit alignment supervision. Again, this is an unexpected finding. 3D reconstruction methods focus almost exclusively on image-based cues or supervision of geometric features. We are the first to show 3D alignment to an image without any geometric supervision or rendering self-supervision. This provides insight that there is sufficient information in ID to convey 3D geometric information.

## 6.2 Future Work

The works presented in this thesis can be expanded in multiple directions.

**Non-Identity information in face representation.** We successfully predicted non-ID information within VGGFace2 and ArcFace descriptors and proposed an adversarial training procedure to minimize leakage of protected attributes in face descriptors. There are many important avenues for future work. First, it is important to replicate these results on other face embedding networks (our initial experiments suggest that our findings indeed transfer between networks). Furthermore, including additional explicit non-ID fea-

tures could improve inversion performance, addressing both privacy concerns and the broader challenge of information leakage in identity vectors. Our investigation currently focuses on expression- and pose-related non-identity information, which is only the initial step in tackling the broader issue of non-identity information leakage. Non-identity attributes encompass a wide range, including lighting conditions, background elements, and transient facial changes like ageing or makeup. Future research will focus on developing specialized adversarial learning frameworks and techniques tailored for specific non-identity attributes.

In Chapter 4, We demonstrated the feasibility of reconstructing facial images and non-ID attributes from ID descriptors using StyleGAN2. An interesting future direction involves utilizing a broader array of training datasets, model architectures, and combinations of loss functions to more comprehensively test the leakage of non-identity information. Another promising direction is the refinement of inversion techniques to enhance the fidelity and accuracy of reconstructed images, particularly in replicating specific non-ID attributes and background elements. Another interesting idea would be to train our ID embedding to GAN latent regressor in a different way. The objective function we used measured the error in the predicted latent vector. However, if we wanted to encourage the model to reconstruct all scene elements including the background, it would be interesting to instead use the image reconstruction error as the training loss. In this case, we don't mind if the predicted GAN latents are different, so long as they reconstruct the whole image well. A variant of this idea could even mask out the face pixels from this image reconstruction error, forcing it to only seek to reconstruct the background.

Furthermore, developing network visualization tools for quantitatively

studying the types and extents of non-identity information leakage could provide valuable insights. Investigating the categories of identity information prone to leakage will enhance our understanding of the intrinsic relationships within facial features and the information they encode. Exploring the potential of non-ID leakage in terms of categories and contextual information could open new perspectives on privacy protection.

**Learning 3D alignment from recognition supervision.** We trained our 3DMM-STN solely with a classification loss on relatively small datasets. It would be intriguing to switch to a metric learning setup, such as triplet loss, and train on much larger datasets. Can this combination of 3D alignment, sampling, and recognition without spatial invariance match the performance of end-to-end face embedding networks?

Our approach discards information outside the 3DMM crop, including some identity-specific information like hair and parts of the neck, as well as external elements such as clothing and background that might provide useful context for recognition. A potential avenue for future work would be to integrate the descriptor extracted by our approach with another descriptor that observes the image regions outside the 3DMM crop.

While our primary goal was to demonstrate that 3D alignment could be learned solely from a recognition loss, in practice—if the objective is to maximize recognition performance—it may be beneficial to employ auxiliary losses to guide the 3D alignment. For example, landmark error could serve as a training loss. Other potential losses might include supervision of the 3DMM expression parameters using an expression classifier or shape consistency losses, where pairs of images with the same identity are encouraged to regress similar 3DMM shape parameters. Generally, these losses aim to assist the localizer in converging to the optimal alignment, which may not

be achievable using only the recognition loss, especially from a poor initial setup.

**Unsupervised disentanglement.** In our work, we rely on choosing attributes or features explicitly that we then seek to reconstruct or remove from the ID embedding. However, in practice, it is difficult or impossible to enumerate all possible non-ID factors that might unhelpfully correlate with identity. It would be interesting in future to try to solve this as an unsupervised disentanglement problem. i.e. given ID-labelled face images and their embeddings, to try to further factorise these into possible non-ID explanations. For example, could we discover the attribute 'wearing hat' without ever having hat labels or training a 'wearing hat' attribute predictor? This seems very challenging but an important direction for future work in face recognition to further improve generalisation beyond a training set which may contain these unhelpful correlations.

# Bibliography

[1] "Face recognition with learning-based descriptor," in *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2707–2714.

[2] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 4432–4441.

[3] ——, "Image2stylegan++: How to edit the embedded images?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8296–8305.

[4] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[5] S. Al-Kuwari, J. H. Davenport, and R. J. Bradford, "Cryptographic hash functions: Recent design trends and security notions." *IACR Cryptol. ePrint Arch.*, vol. 2011, p. 565, 2011.

[6] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[7] H. Bai, D. Kang, H. Zhang, J. Pan, and L. Bao, "Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 362–371.

[8] A. Bas, P. Huber, W. A. Smith, M. Awais, and J. Kittler, "3d morphable models as spatial transformer networks," in *Proceedings of the*

*IEEE International Conference on Computer Vision Workshops*, 2017, pp. 904–912.

[9] N. Bayat, V. R. Khazaie, and Y. Mohsenzadeh, "Inverse mapping of face gans," *arXiv preprint arXiv:2009.05671*, 2020.

[10] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint arXiv:1907.05047*, 2019.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[12] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," in *Compututer Graphics Forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 641–650.

[13] V. Blanz, P. Grother, P. J. Phillips, and T. Vetter, "Face recognition based on frontal views generated from non-frontal images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2005, pp. 454–461.

[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[15] Z. Bu, J. Dong, Q. Long, and W. J. Su, "Deep learning with gaussian differential privacy," *Harvard data science review*, vol. 2020, no. 23, 2020.

[16] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

[17] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[18] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International journal of computer vision*, vol. 107, pp. 177–190, 2014.

[19] M. A. P. Chamikara, P. Bertók, I. Khalil, D. Liu, and S. Camtepe, "Privacy preserving face recognition utilizing differential privacy," *Computers & Security*, vol. 97, p. 101951, 2020.

[20] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 122–129.

[21] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3025–3032.

[22] Q. Chen, C. Xiang, M. Xue, B. Li, N. Borisov, D. Kaarfar, and H. Zhu, "Differentially private data generative models," *arXiv preprint arXiv:1812.02274*, 2018.

[23] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *CVPR*. IEEE Computer Society, 2017, pp. 3386–3395.

[24] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[26] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7093–7102.

[27] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[28] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[29] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 60–68.

[30] P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and R. Chellappa, "How are attributes expressed in face dcnns?" in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 85–92.

[31] P. Dhar, J. Gleason, H. Souri, C. D. Castillo, and R. Chellappa, "Towards gender-neutral face descriptors for mitigating bias in face recognition," *arXiv preprint arXiv:2006.07845*, 2020.

[32] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 1002–1014, 2017.

[33] C. N. Duong, T.-D. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy, "Vec2face: Unveil human faces from their blackbox features in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6132–6141.

[34] C. Dwork, "Differential privacy," in *ICALP (2)*, ser. Lecture Notes in Computer Science, vol. 4052. Springer, 2006, pp. 1–12.

[35] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 534–551.

[36] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.

[37] C. Garvie and J. Frankle, "Facial-recognition software might have a racial bias problem," *The Atlantic*, vol. 7, no. 04, p. 2017, 2016.

[38] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3D morphable model regression," in

*Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8377–8386.

[40] S. Gong, X. Liu, and A. K. Jain, "Mitigating face recognition bias via group adaptive classifier," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3414–3424.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in neural information processing systems (NeurIPS)*, 2014, pp. 2672–2680.

[42] C. Gou, Y. Wu, F.-Y. Wang, and Q. Ji, "Shape augmented regression for 3d face alignment," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 604–615.

[43] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 87–102.

[44] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[45] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "von mises-fisher mixture model-based deep learning: Application to face verification," *arXiv preprint arXiv:1706.04264*, 2017.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6d rotation representation for unconstrained head pose estimation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2496–2500.

[48] M. Q. Hill, C. J. Parde, C. D. Castillo, Y. I. Colon, R. Ranjan, J.-C. Chen, V. Blanz, and A. J. O'Toole, "Deep convolutional neural

networks in the face of caricature," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 522–529, 2019.

[49] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[50] L. Huang, M. Wang, J. Liang, W. Deng, H. Shi, D. Wen, Y. Zhang, and J. Zhao, "Gradient attention balance network: Mitigating face recognition racial bias via gradient attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 38–47.

[51] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch, "Fitting 3d morphable face models using local features," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 1195–1199.

[52] A. Inc., "There's more to iPhone," https://www.apple.com/iphone/more, 2021, [Online; accessed 13-August-2021].

[53] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[54] A. Jourabloo and X. Liu, "Pose-invariant 3d face alignment," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3694–3702.

[55] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single cnn," in *Proceedings of the IEEE International Conference on computer vision*, 2017, pp. 3200–3209.

[56] T. Kanade, "Picture processing system by computer complex and recognition of human faces. in doctoral dissertation, kyoto university," 1973.

[57] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[58] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[59] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[60] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.

[61] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz, "Illumination-aware age progression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3334–3341.

[62] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939.

[63] T. Koizumi and W. A. Smith, ""look ma, no landmarks!"–unsupervised, model-based dense face alignment," in *European Conference on Computer Vision*. Springer, 2020, pp. 690–706.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[65] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*. IEEE Computer Society, 2009, pp. 365–372.

[66] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou, "Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild"," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 760–769.

[67] Y. Lee, T. Choi, H. Go, H. Lee, S. Cho, and J. Kim, "Exp-gan: 3d-aware facial image generation with expression control," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3812–3827.

[68] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 289–302, 2013.

[69] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[70] J. Levinson, C. Esteves, K. Chen, N. Snavely, A. Kanazawa, A. Rostamizadeh, and A. Makadia, "An analysis of SVD for deep rotation estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 554–22 565, 2020.

[71] A. Li, J. Guo, H. Yang, and Y. Chen, "Deepobfuscator: Adversarial training framework for privacy-preserving image classification," *arXiv preprint arXiv:1909.04126*, 2019.

[72] Z. Liang, S. Ding, and L. Lin, "Unconstrained facial landmark localization with backbone-branches fully-convolutional networks," *arXiv preprint arXiv:1507.03409*, 2015.

[73] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467–476, 2002.

[74] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[75] X. Liu, "Discriminative face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1941–1954, 2008.

[76] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," *arXiv preprint arXiv:1710.00870*, 2017.

[77] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[78] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *CVPR*. IEEE Computer Society, 2015, pp. 5188–5196.

[79] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 158–165.

[80] J. McDonagh and G. Tzimiropoulos, "Joint face detection and alignment with a deformable hough transform model," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 569–580.

[81] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 82–89.

[82] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[83] F. Mokhayeri, K. Kamali, and E. Granger, "Cross-domain face synthesis using a controllable gan," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 252–260.

[84] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.

[85] M. Parchami, S. Bashbaghi, and E. Granger, "Video-based face recognition using ensemble of haar-like deep convolutional neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 4625–4632.

[86] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2015.

[87] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The*

*Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 38–56.

[88] S. Ploumpis, H. Wang, N. Pears, W. A. Smith, and S. Zafeiriou, "Combining 3d morphable models: A large scale face-and-head model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 934–10 943.

[89] X. Qi and L. Zhang, "Face recognition via centralized coordinate learning," *arXiv preprint arXiv:1801.05678*, 2018.

[90] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[91] A. Rai, H. Gupta, A. Pandey, F. V. Carrasco, S. J. Takagi, A. Aubel, D. Kim, A. Prakash, and F. De la Torre, "Towards realistic generative 3d face models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3738–3748.

[92] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[93] A. Razzhigaev, K. Kireev, E. Kaziakhmedov, N. Tursynbek, and A. Petiushko, "Black-box face recovery from identity features," in *ECCV Workshops (5)*, ser. Lecture Notes in Computer Science, vol. 12539. Springer, 2020, pp. 462–475.

[94] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," *arXiv preprint arXiv:2008.00951*, 2020.

[95] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 460–469.

[96] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1259–1268.

[97] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE*

*conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[98] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[99] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9659697

[100] G. Shamai, R. Slossberg, and R. Kimmel, "Synthesizing facial photometries and corresponding geometries using generative adversarial networks," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3s, pp. 1–24, 2019.

[101] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.

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

[103] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[104] E. Smirnov, A. Melnikov, S. Novoselov, E. Luckyanets, and G. Lavrentyeva, "Doppelganger mining for face representation learning," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 1916–1923.

[105] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 991–21 002.

[106] K. Sun, S. Wu, Z. Huang, N. Zhang, Q. Wang, and H. Li, "Controllable 3d face synthesis with conditional generative occupancy fields," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 331–16 343, 2022.

[107] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Advances in neural information processing systems*, vol. 27, 2014.

[108] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[109] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[110] ——, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[111] ——, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2892–2900.

[112] ——, "Sparsifying neural network connections for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4856–4864.

[113] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[114] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 5, no. 6, 2014.

[115] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "Beyond identity: What information is stored in biometric face templates?" in *2020 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[116] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: Face model learning from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 812–10 822.

[117] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zöllhofer, and C. Theobalt, "StyleRig: Rigging StyleGAN for 3d control over portrait images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[118] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc. International Conference on Computer Vision (ICCV)*, Oct 2017.

[119] Y. Tie and L. Guan, "A deformable 3-d facial expression model for dynamic human emotional state recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 1, pp. 142–157, 2012.

[120] P. Tinsley, A. Czajka, and P. Flynn, "This face does not exist... but it might be yours! identity leakage in generative models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1320–1328.

[121] A. T. Tran, T. Hassner, I. Masi, and G. G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1493–1502.

[122] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[123] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[124] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu, "Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 333–20 342.

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

[126] T. Wei, D. Chen, W. Zhou, J. Liao, W. Zhang, L. Yuan, G. Hua, and N. Yu, "A simple baseline for stylegan inversion," *CoRR*, vol. abs/2104.07661, 2021.

[127] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*. Springer, 2016, pp. 499–515.

[128] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*. IEEE, 2011, pp. 529–534.

[129] H. Wu, X. Liu, and G. Doretto, "Face alignment via boosted ranking model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[130] J. Wulff and A. Torralba, "Improving inversion and generation diversity in stylegan using a gaussianized latent space," *arXiv preprint arXiv:2009.06529*, 2020.

[131] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.

[132] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*, vol. 41, no. 2. Wiley Online Library, 2022, pp. 641–676.

[133] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "Ganobfuscator: Mitigating information leakage under gan via differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2358–2371, 2019.

[134] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[135] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8715–8724.

[136] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 786–791.

[137] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5409–5418.

[138] H. Zhao, Y. Shi, X. Tong, X. Ying, and H. Zha, "Qamface: Quadratic additive angular margin loss for face recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1901–1905.

[139] Y. Zhao, X. Cao, S. Liu, J. Che, W. Ren, J. Cao, and J. Lin, "A facial expression transfer method based on 3dmm and diffusion models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 3145–3149.

[140] A. Zhmoginov and M. Sandler, "Inverting face embeddings with convolutional neural networks," *arXiv preprint arXiv:1606.04189*, 2016.

[141] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain gan inversion for real image editing," *arXiv preprint arXiv:2004.00049*, 2020.

[142] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[143] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li, "Discriminative 3d morphable model fitting," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.