Imperial College, London
Department of Computing

# Synthesization and Reconstruction of 3D Faces by Deep Neural Networks

Baris Gecer

August 2020

Supervised by Dr. Stefanos Zafeiriou

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of the Imperial College, London

This thesis is entirely my own work, and, except where otherwise indicated, describes my own research

# Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Abstract

The past few decades have witnessed substantial progress towards 3D facial modelling and reconstruction as it is high importance for many computer vision and graphics applications including Augmented/Virtual Reality (AR/VR), computer games, movie post-production, image/video editing, medical applications, etc. In the traditional approaches, facial texture and shape are represented as triangle mesh that can cover identity and expression variation with non-rigid deformation. A dataset of 3D face scans is then densely registered into a common topology in order to construct a linear statistical model. Such models are called 3D Morphable Models (3DMMs) and can be used for 3D face synthesization or reconstruction by a single or few 2D face images. The works presented in this thesis focus on the modernization of these traditional techniques in the light of recent advances of deep learning and thanks to the availability of large-scale datasets.

Ever since the introduction of 3DMMs by over two decades, there has been a lot of progress on it and they are still considered as one of the best methodologies to model 3D faces. Nevertheless, there are still several aspects of it that need to be upgraded to the *"deep era"*. Firstly, the conventional 3DMMs are built by linear statistical approaches such as Principal Component Analysis (PCA) which omits high-frequency information by its nature. While this does not curtail shape, which is often smooth in the original data, texture models are heavily afflicted by losing high-frequency details and photorealism. Secondly, the existing 3DMM fitting approaches rely on very primitive (i.e. RGB values, sparse landmarks) or hand-crafted features (i.e. HOG, SIFT) as supervision that are sensitive to "in-the-wild" images (i.e. lighting, pose, occlusion), or somewhat missing identity/expression resemblance with the target image. Finally, shape, texture, and expression modalities are separately modelled by ignoring the correlation among them, placing a fundamental limit to the synthesization of semantically meaningful 3D faces. Moreover, photorealistic 3D face synthesis has not been studied thoroughly in the literature.

This thesis attempts to address the above-mentioned issues by harnessing the power of deep neural network and generative adversarial networks as explained below:

Due to the linear texture models, many of the state-of-the-art methods are still not capable of reconstructing facial textures with high-frequency details. For this, we take a radically different approach and build a high-quality texture model by Generative Adversarial Networks (GANs) that preserves details. That is, we utilize GANs to train a very powerful generator of facial texture in the UV space.

And then show that it is possible to employ this generator network as a statistical texture prior to 3DMM fitting. The resulting texture reconstructions are plausible and photorealistic as GANs are faithful to the real-data distribution in both low- and high- frequency domains.

Then, we revisit the conventional 3DMM fitting approaches making use of non-linear optimization to find the optimal latent parameters that best reconstruct the test image but under a new perspective. We propose to optimize the parameters with the supervision of pretrained deep identity features through our end-to-end differentiable framework. In order to be robust towards initialization and expedite the fitting process, we also propose a novel self-supervised regression-based approach. We demonstrate excellent 3D face reconstructions that are photorealistic and identity preserving and achieve for the first time, to the best of our knowledge, facial texture reconstruction with high-frequency details.

In order to extend the non-linear texture model for photo-realistic 3D face synthesis, we present a methodology that generates high-quality texture, shape, and normals jointly. To do so, we propose a novel GAN that can generate data from different modalities while exploiting their correlations. Furthermore, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions. Additionally, we study another approach for photo-realistic face synthesis by 3D guidance. This study proposes to generate 3D faces by linear 3DMM and then augment their 2D rendering by an image-to-image translation network to the photorealistic face domain. Both works demonstrate excellent photorealistic face synthesis and show that the generated faces are improving face recognition benchmarks as synthetic training data.

Finally, we study expression reconstruction for personalized 3D face models where we improve generalization and robustness of expression encoding. First, we propose a 3D augmentation approach on 2D head-mounted camera images to increase robustness to perspective changes. And, we also propose to train generic expression encoder network by populating the number of identities with a novel multi-id personalized model training architecture in a self-supervised manner. Both approaches show promising results in both qualitative and quantitative experiments.

# Acknowledgements

This Ph.D. thesis contains a summary and conclusions of a unique academic journey that contains every bit of the emotional spectrum. During this journey, I often found myself in the middle of emotional roller coaster rides, different states of survival, and never-ending deadlines which would be impossible to overcome without the support of my colleagues, family, and friends. I would like to express my sincere thanks to these special individuals.

I cannot express my gratitude to my supervisor, Dr. Stefanos Zafeiriou, who is not only an excellent academic in the field but also is keen to share his unique expertise with others. With his sound guidance at my back, I felt motivated and encouraged at every step of this ladder and blessed to have such an amazing mentor with the warmth of a brother.

My sincere thanks also go to Dr. Tae-Kyun Kim, Dr. Fernando de la Torre, and Dr. Josef Kittler, for their gentle support and constructive comments. Likewise, I appreciate thorough review of this thesis by my examiners, Dr. Abhijeet Ghosh and Dr. William A. P. Smith.

I am also thankful to the the Turkish Ministry of Education for their financial support.

I feel fortunate for the enjoyable time at the office with my wonderful colleagues: Jiankang, Stylianos P., Stylianos M., Alexandros, Caner, Ju-il, Binod, Vasileios, Athanasios, Evangelos, Anil, Guillermo, and many others. Life would be very difficult without the exceptional companionship and wise counseling of my very best friends: Enver, Husameddin, Aziz, Kemal, Mehmed, Tugay, Yasin, Burak, Ismail, Mustafa, Alim, Volkan, Engin, Cagatay, Abdulhamid, Birol, Enes, and others. I feel fortunate and honored for having their warm companionship, great entertainment, and compassionate support.

While having very little understanding, I thank my family for always keeping their incomparable support at my back. Without a doubt, I am deeply indebted to my dear mother, father, and sister. Thank you!

This journey required more sacrifice from my dear wife, Esra Nur, than myself. Never she has shown a single hesitation for continuously moving out, having a humble lifestyle, and taking more responsibility at home. All came with her heartwarming smile, indefatigable patience, and previous love. The secret ingredient of this thesis was the support of my 7-month old son, Emir Taha, who attempted to recompense the time he borrowed by the following: *"uhyh n gbhj bjhjbbyikjkulkkkkkkkjkjyihlvebxtfb gvbl gk gjhkfl gkh lfkhjhgl"*. Thank you both for standing beside me throughout this work!

Above all, the most valuable benefit of this journey to me is the observation of the secret manifestations of Allah's divine names and attributes, which can be seen more closely with science and open hearth. Certainly, I too should say *All praise be to Him*, with the definition of Bediuzzaman: "*Since the perfections found in all beings which are the cause of acclaim and tribute are His, praise too belongs to Him. Acclaim and laudation, from whomever to whomever it has come and will come, from pre-eternity to post-eternity, all of it belongs to Him.*"

.

*Dedicated to my beautiful son Emir Taha and my beloved wife Esra Nur ...*

'*Whoever gave me this face and these features is the Maker of all human beings which resemble me in regard to basic features, and He is the Maker too of all living beings.*'

*Bediuzzaman Said Nursi*

# Contents

**Abstract** **4**

**Acknowledgements** **7**

**1 Introduction** **25**

1.1 Challenges . . . . . 27

1.2 Contributions . . . . . 30

1.3 Publications . . . . . 32

**2 Background** **34**

2.1 3D Face Modeling . . . . . 35

2.1.1 Linear Models . . . . . 36

2.1.2 Non-Linear Generative Models . . . . . 38

2.1.3 Person-specific face models . . . . . 41

2.2 2D & 3D Face Synthesis . . . . . 42

2.2.1 Image Formation . . . . . 42

2.2.2 Differentiable Renderers . . . . . 45

2.3 3D Face Reconstruction from a Single Image . . . . . 47

2.3.1 Energy Functions . . . 48
2.3.2 Optimization-based Approaches . . . 51
2.3.3 Regression-based Approaches . . . 52
2.4 Conclusion . . . 54

**3 Generative Adversarial Network Regression and Fitting for 3D Face Reconstruction 55**

3.1 Introduction . . . 56
3.2 History of 3DMM Fitting . . . 59
3.2.1 3DMM representation . . . 59
3.2.2 Fitting . . . 61
3.3 Approach . . . 62
3.3.1 GAN Texture Model . . . 63
3.3.2 Differentiable Renderer . . . 64
3.3.3 Cost Functions . . . 65
3.3.4 Model Fitting . . . 67
3.4 Drawbacks and Improvements on GANFit . . . 68
3.4.1 Fitting with Multiple Images (i.e. Video): . . . 68
3.4.2 Regressing parameters by Encoder . . . 68
3.5 Experiments . . . 71
3.5.1 Implementation Details . . . 71
3.5.2 Qualitative Results . . . 71
3.5.3 Quantitative Experiments . . . 74
3.5.4 Ablation Study . . . 79

3.6 Conclusion . . . . . 80

**4 Synthesizing Coupled 3D Face Modalities by Trunk-Branch GAN 82**

4.1 Introduction . . . . . 83

4.2 Related Work . . . . . 86

4.2.1 3D face modeling . . . . . 86

4.2.2 Photorealistic face synthesis . . . . . 87

4.2.3 Boosting face recognition by synthetic training data . . . . . 88

4.3 Approach . . . . . 89

4.3.1 UV Maps for Shape, Texture and Normals . . . . . 89

4.3.2 Trunk-Branch GAN to Generate Coupled Texture, Shape and Normals . . . . 90

4.3.3 Expression Augmentation by Conditional GAN . . . . . 91

4.3.4 Photorealistic Rendering with Generated UV maps . . . . . 92

4.4 Results . . . . . 93

4.4.1 Qualitative Results . . . . . 94

4.4.2 Pose-invariant Face Recognition . . . . . 98

4.5 Conclusion . . . . . 101

**5 Generating Photorealistic Face Images from 3D Morphable Model 103**

5.1 Introduction . . . . . 104

5.2 Related Works . . . . . 107

5.3 Adversarial Identity Generation . . . . . 110

5.3.1 Unsupervised Domain Adaptation . . . . . 110

5.3.2 Adversarial Pair Matching . . . . . 111

5.3.3 Identity Preservation . . . 113
5.4 Implementation Details . . . 115
5.5 Results and Discussions . . . 116
5.5.1 Qualitative Evaluation . . . 117
5.5.2 The Added Realism and Identity Preservation . . . 118
5.5.3 Face Recognition with GANFaces dataset . . . 119
5.6 Conclusions . . . 121

**6 Robustifing 3D Expression Reconstruction of Codec Avatars 123**
6.1 Introduction . . . 124
6.2 Methodologies . . . 126
6.2.1 Baseline . . . 126
6.2.2 3D Data Augmentation . . . 127
6.2.3 Generic Encoders . . . 129
6.3 Experiments . . . 130
6.4 Conclusion . . . 134

**7 Conclusion 137**
7.1 Summary of Thesis Achievements . . . 137
7.2 Limitations & Future Work . . . 138
7.2.1 Higher Resolution and Full-head Completion . . . 138
7.2.2 Semi-personalized 3D Face Models . . . 139

**Bibliography 140**

# List of Tables

3.1 Accuracy results for the meshes on the MICC Dataset using point-to-plane distance. The table reports the mean error (Mean), the standard deviation (Std.). . . . . 78

4.1 Verification performance (%) of different models on CFP-FP and AgeDB-30. . . . . 101

4.2 The angles between face pairs from CFP-FP predicted by different models trained from the CASIA and combined data. The generated data can obviously enhance the pose-invariant feature embedding. . . . . 102

5.1 Comparison with state-of-the-art studies on LFW performances . . . . . 120

6.1 Mean Absolute Errors of different settings for different parts of the face. The comparison is done on UV Texture and mesh geometry compared to Ground truth provided by [WSS+19] correspondence. Best accuracy of each line is typed in bold. . . . . 131

# List of Figures

1.1 Top: UV textures with high-frequency details. Bottom: Projection of top images to a PCA space built by 10,000 high resolution images [BRZ+16] . . . . . 27

1.2 Some cherry picked failure cases of Progressive Growing GAN [KALL18] . . . . . 30

2.1 Illustration of 3D mesh pre-processing steps. In (a) we present a raw mesh, in (b) the registered mesh using the Large Scale Face Model (LSFM) template [BRZ+16], in (c) the unwrapped 3D mesh in the 2D UV space, and finally in (d) the interpolated 2D UV map. Interpolation is carried out using the barycentric coordinates of each pixel in the registered 3D mesh. . . . . 35

2.2 Toy experiment by GAN Lab [KTC+19] to illustrate capabilities of GAN for capturing sophisticated distributions. (a) Distribution of the target distribution. (b) Generated samples by GAN after training. (c) The generator's transformation manifold from noise space. Darker the areas, more likely to generate at that area. (d) The discriminator's decision space: Green for real, Purple for generated. (e) All of them overlayyed. . . . . 39

2.3 Progressive renderings by *Blinn-Phong Reflectance Model*. (a-e) BRDF components of a subject scanned in a light stage. (f-j) Full face and zoom-in rendering results by adding one component at a time. (k) Per-vertex diffuse albedo rendering. (l) Per-vertex diffuse normals rendering. (m) Diffuse normal rendering from UV map. All images rendered by a differentiable renderer implemented within the scope of this thesis. . . . . 44

2.4 An overview of traditional linear 3D Morphable Model Fitting. . . . . 46

3.1 The proposed deep fitting approach can reconstruct high quality texture and geometry from a single image with precise identity recovery. The reconstructions in the figure and the rest of the chapter are represented by a vector of size 700 floating points and rendered without any special effects. We would like to highlight that the depicted texture is reconstructed by our model and none of the features taken directly from the image. . . . . 57

3.2 Detailed overview of the proposed approach. A 3D face reconstruction is rendered by a differentiable renderer (shown in purple). Cost functions are mainly formulated by means of identity features on a pretrained face recognition network (shown in gray) and they are optimized by flowing the error all the way back to the latent parameters ($p_s, p_e, p_t, c, i$, shown in green) with gradient descent optimization. End-to-end differentiable architecture enables us to use computationally cheap and reliable first order derivatives for optimization thus making it possible to employ deep networks as a generator (i.e. statistical model) or as a cost function. . . . . 62

3.3 Example fits of our approach for the images from various datasets. Please note that our fitting approach is robust to occlusion (e.g. glasses), low resolution and black-white in the photos and generalizes well with ethnicity, gender and age. The reconstructed textures are very well at capturing high frequency details of the identities; likewise, the reconstructed geometries from 3DMM are surprisingly good at identity preservation thanks to the identity features used, e.g. crooked nose at bottom-left, dull eyes at bottom-right and chin dimple at top-left . . . . . 66

3.4 The fitting results of the proposed approach for the input image (a) initialized by FastGANFit (b) or randomly resulting reconstructions given random initial textures (c-i) . . . . . 69

3.5 Overview of the approach with regression network. The network is end-to-end connected with the differentiable renderer and the lost functions of GANFit. It benefits from the activations of all layers of a pretrained face recognition network and detection of a hourglass landmark detector. The network is trained similar to GANFit optimization: 1) alignment 2) full objective. The only difference is that now the regression network is being optimized (i.e. trained) instead of the trainable latent parameters of GANFit. . . . . 70

3.6 Comparison of our qualitative results with other state-of-the-art methods in MoFA-Test dataset. Rows 2-4 show comparison of textured geometry with original expression, rows 5-8 show comparison of textured geometry with neutral expression and rows 9-11 compare only shapes. . . . . 72

3.7 Qualitative comparison with [YSN+18, SYH+17] by overlaying the reconstructions on the input images. Our method can generate high fidelity texture with accurate shape, camera and illumination fitting. . . . . 74

3.8 Qualitative comparison with [SWH+17] by means of texture maps, whole and partial face renderings. Please note that while our method does not require any particular renderer for special effects, e.g. lighting, [SWH+17] produce these renderings with a commercial renderer called Arnold. . . . . 75

3.9 Results under more challenging conditions, i.e. strong illuminations, self-occlusions and facial hair. (a) Input image. (b) Estimated fitting overlayyed including illumination estimation. (c) Overlayyed fitting without illumination. (d) Normalized pixel-wise intensity ratio of (b) to (c). (e) Estimated shape mesh . . . . . 76

3.10 Our results on BAM dataset[WFJ+17] compared to [GCM+18]. Our method is robust to many image deformations and even capable of recovering identities from paintings thanks to strong identity features. . . . . 77

3.11 Cosine similarity distributions of rendered and real images LFW based on activations at the embedding layer of VGG-Face network[PVZ15]. Our method achieves more than 0.5 similarity on average which [GCM+18] has 0.35 average similarity and [THMM17] 0.16 average similarity. Camera and lighting parameters are fixed for all renderings. . . . . 79

3.12 Our method successfully preserve identity so that distribution of cosine similarity of same/different pairs is separable by thresholding. Camera and lighting parameters are fixed for all renderings. . . . . 80

3.13 Contributions of the components or loss terms of the proposed approach with an leave-one-out ablation study. . . . . 81

4.1 We propose a novel GAN that can synthesize high-quality texture, shape, and normals jointly for realistic and coherent 3D faces of novel identities. The separation of branch networks allows the specialization of the characteristic of each one of the modalities while the trunk network maintains the local correspondences among them. Moreover, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions. We annotate the training dataset automatically by an expression recognition network to couple those expression encodings to the texture, shape, and normals UV maps. . . . . 84

4.2 UV extraction process. In (a) we present a raw mesh, in (b) the registered mesh using the Large Scale Face Model (LSFM) template [BRZ+16], in (c) the unwrapped 3D mesh in the 2D UV space, and (d) the interpolated 2D UV map. Interpolation is carried out using the barycentric coordinates of each pixel in the registered 3D mesh. 89

4.3 Zoom-in on rendering results with (a) only the shape, (b) adding the albedo texture, (c) adding the generated normals, and (d) using identity-generic detail normal, specular albedo, roughness, scatter and translucency maps. . . . . 93

4.4 Generated UV representations and their corresponding additive renderings. Please note the strong correlation between UV maps, high fidelity and photorealistic renderings. The figure is best viewed in zoom. . . . . 95

4.5 Variation of generated 3D faces by our model. Each block shows diversity in a different aspect. Readers are encouraged to zoom in on a digital version. . . . . 96

4.6 First and forth rows shows generations of six universal expressions. The other rows show texture and normals maps are used to generate the corresponding 3D faces. Please note how expressions are represented and correlated in the texture and normals space. . . . . 97

4.7 The figure shows the interpolation between pair of identities. Smooth transition indicates generalization of our GAN model. . . . . 98

4.8 Complete full head representations in association with the facial topology corresponding in Fig.4.7. Even in the full head topology our generation methodology ensures a smooth transition during interpolation. . . . . 98

4.9 Comparison with seperate GAN models and PCA model. (a) Generation by our model. (b) Same texture with random shape and normals. (c) Same texture and shape with random normals (i.e. beard). (d) Generation by a PCA model constructed by the same training data and the same identity-generic rendering tools as explained in Sec. 3.4. . . . . 99

4.10 Examples of generated data ('Gen') by the proposed method. . . . . 100

4.11 Angle distributions of CFP-FP positive (red) and negative (blue) pairs in the $512$-$D$ feature space. Red area indicates positive pairs while blue indicates negative pairs. . 101

5.1 Our approach aims to synthesize photorealistic images conditioned by a given synthetic image by 3DMM. It regularizes cycle consistency [ZPIE17] by introducing an additional adversarial game between the two generator networks in an unsupervised fashion. Thus the under-constrained cycle loss is supervised to have correct matching between the two domains by the help of a limited number of paired data. We also encourage the generator to preserve face identity by a set-based supervision through a pretrained classification network. . . . . 105

5.2 Comparison of our pair matching method to the related work. (a) In the traditional GAN approach, the discriminator module aligns the distribution of real and synthetic images by means of a classification network. (b) BEGAN[BSM17] and many others showed that the alignment of reconstruction error distributions offers a more stable training. (c) We propose to utilize this autoencoder approach to align the distribution of pairs to encourage generated images to be transformed to the realistic domain with a game between real and synthetic pairs. (d) An alternative to our method is to introduce wrongly labeled images to the discriminator to teach pair-wise matching as proposed by [RAY+16] for text to images synthesis. . . . . 112

5.3 Quality of 9 images of 3 identities (per row) during the training. Red plot shows the proposed identity preservation loss ($x$ axis) over the iterations ($y$ axis). Note the changes of fine-details on the faces which is the main motivation of set-based identity preservation. . . . . 113

5.4 Random samples from GANFaces dataset. Each row belongs to the same identity. Notice the variation in pose, expression and lighting. . . . . 117

5.5 Comparison to (a) input 3DMM images, (b) results with simGAN settings [BSD+17],(c) results with cycleGAN settings [ZPIE17] and (d) results with additional reconstruction loss. Figures correspond to left half of the Fig. 5.4 and each row belongs to the same identity. . . . . 118

5.6 Images generated by the proposed approach conditioned by identity variation in the vertical axis, normalized and mouth open expression in left and right blocks and pose variation in the horizontal axis. Images in this figure are not included in the training 119

5.7 Distances of 1000 positive and 1000 negative pairs from three different datasets (GANFaces, 3DMM synthetic images, Oxford VGG) embedded on a NN4 network that is trained with CASIA Face dataset . . . . . 120

5.8 Face recognition benchmark experiments. (Left) Number of images used from the two datasets in the experiments. The total number of images in the VGG data set is 1.8M since some images were removed from the URL (Middle) Performance on the LFW dataset with (solid) and without (dashed) the GANFaces-500K dataset. (Right) True Positive Rates on the IJB-A verification task with (solid) and without (dashed) the GANFaces-500K dataset. . . . . 121

6.1 Once the correspondence is established by [WSS+19], one can re-train the HMC encoder with different augmentation techniques such as brightness, gamma, 2D translation, rotation, distortion and cropping. . . . . 126

6.2 In order to 3D augment a given HMC image, we first project 3D mesh on the image with known geometry and camera parameters. After filling triangles with HMC image with the projected 2D vertex coordinates, we render the mesh with a small changes to rotation and translation of the camera parameters. We repeat this procedure randomly during the training. . . . . 128

6.3 Training architecture of the Generic HMC Encoder by the data of multiple subjects. The key insight is to encode all HMC images into a common expression space and then reconstruct them in the corresponding personalized Codec spaces. Aggregation of all subjects provides better generalization to the enviromental conditions that appear in all capture sessions. . . . . 129

6.4 Qualitative results of the test capture of subject A in condition 1. Each row shows results of a different expression reconstruction of HMC images by the Codec Avatars. The experiment setting of each column is displayed at the top. . . . . 133

6.5 Qualitative results of the test capture of subject A in condition 2. Each row shows results of a different expression reconstruction of HMC images by the Codec Avatars. The experiment setting of each column is displayed at the top. . . . . 134

6.6 Qualitative results of the test capture of subject B in condition 1. Each row shows results of a different expression reconstruction of HMC images by the Codec Avatars. The experiment setting of each column is displayed at the top. . . . . 135

7.1 3D face reconstruction from a single in-the-wild image to high-resolution BRDF renderable full-head reconstruction. Initial reconstruction is done by [GPKZ19, GPKZ20] as shown in **Chapter 3**. Then the reconstruction is enhanced by [LMG+20] to higher resolution BRDF renderable texture and normals maps. Finally, its full-head is completed by [PVOS+20]. . . . . 139

# Chapter 1

# Introduction


1.1 Challenges . . . . . 27

1.2 Contributions . . . . . 30

1.3 Publications . . . . . 32


One of the most ubiquitous class of data in computer vision is, without doubt, human face with more than a seven billion real unique samples. Having such variation and deformation within very similar characteristic patterns distinct from other objects, makes it scientifically interesting for many. Thus, human faces have always been featuring in many computer vision and machine learning applications such as detection [ZZLQ16, DGZ+19], identity/expression recognition [DGXZ19, SKP15, GBK17, PVZ15, VA19], 3D modeling/reconstruction [EST+20, BV99, BV03, RBSB18, CWZ+14, BAP+17, JZD+18, CKPZ18, TZK+17, RSK16, LMG+20, THM+18, SRK17], 2D/3D generation [BSM17, KALL18, KLA19, KLA+20, SSK19b, SSK19a, GLP+20], person reenactment/swapping [TZS+19, NKH19, TZN+15, KRZ20] and so on. This thesis will cover a series of work around photorealistic deep 3D face reconstruction and synthesis.

The first evidence of proper 3D face reconstruction date back 3rd millennium B.C. of antique Egypt and Greece where artists use *sandstone, marble, clay* and other materials to make a sculpture of political leaders or commanders. Even if it is moved to digital platforms these days, the way that artists shape their artworks still enlightens the approaches pursued automated digital face reconstruction.

One key aspect of this artistry is that the reconstruction has to be distinguishable from other faces and should resemble the targeted subject. This can only be achieved by careful examination of identity-related characteristics of the facial features.

One of the earliest and still very popular approach for digital 3D face modeling and reconstruction is called 3D Morphable Models (3DMMs) [BV99]. 3DMMs are basically linear statistical models of shape and appearance that are constructed by a set of 3D faces. Such a model provides analytic definition of human face which later becomes *a priori* to synthesize novel faces in 3D or to solve 3D face reconstruction from a single 2D image, which is otherwise an ill-posed problem. The success of its descendants [EST+20] that builds upon it, validates the effectiveness of this approach and indicates there is a lot more to explore.

In the past two years, a lot of work has been conducted on how to harness Deep Convolutional Neural Networks (DCNNs) for 3D shape and texture reconstruction. The first such methods either trained regression DCNNs from image to the parameters of a 3DMM [THMM17] or used a 3DMM to synthesize images [RSK16, GZC+19] and formulate an image-to-image translation problem using DCNNs to estimate the depth[1] [SRK17]. The more recent unsupervised DCNN-based methods are trained to regress 3DMM parameters from identity features by making use of differentiable image formation architectures [CBK+17] and differentiable renderers [GCM+18, TZK+17, RSOK17]. The most recent methods such as [TZG+18, TL18, GZC+16] use both the 3DMM model, as well as additional network structures (called correctives) in order to extend the shape and texture representation.

3D face reconstruction and modeling empower many useful applications such as Augmented Reality/Virtual Reality (AR/VR) [LSSS18, WSS+19], computer games [HSW+17], movie post-production [TZS+19, CHZ14, GVS+15, BWP13], medical applications such as plastic surgery [PCKO15], personalized facial mask [SPH20], etc. Photorealistic face modeling can also advance computer vision algorithms by providing synthetically generating training data with accurate annotation [SRK17].

That said, 3D face reconstruction and modeling is still an open research topic that contains many challenges to this date, including maintaining photorealism and generalization, preserving identity, expression, and other facial features. This thesis studies novel techniques for high-quality digital-

[1] The depth was afterwards refined by fitting a 3DMM and then changing the normals by using image features.



Figure 1.1: Top: UV textures with high-frequency details. Bottom: Projection of top images to a PCA space built by 10,000 high resolution images [BRZ+16]

ization of Faces into a 3D virtual world (i.e. AR/VR) while preserving identity-, expression- and motion-related attributes. For that, it explores photorealistic 3D Face reconstruction, modeling, and synthesis by Generative Adversarial Networks (GANs) and Deep Neural Networks (DNNs).

The remainder of the introductory chapter is organized as follows. Firstly, we summarize the challenges and the problem scope in Section 1.1, and then we introduce the proposed contributions of this thesis in Section 1.2. Finally, in Section 1.3 we list relevant publications that appeared in peer-reviewed conferences and journals. Additionally, in **Chapter** 2, we explain the concepts and scope briefly on which this thesis will be based. Specifically, Section 2.1 gives a smooth introduction to conventional 3D face models, i.e. 3DMMs, and non-linear generative models, i.e. GANs, and subsequently, Section 2.2 explains the traditional steps necessary for 2D and 3D face synthesis. And finally, Section 2.3 briefly formulates 3D face reconstruction from a single image by putting the pieces together.

## 1.1 Challenges

The past few decades have witnessed a lot of progress towards 3D face modeling and reconstruction from single images by traditional approaches, such as 3D Morphable Models (3DMMs). Recently, there have been quite a few works that improve these approaches significantly by harnessing deep neu-

ral networks. Nevertheless, there are still a number of aspects that can be advanced by discriminative and generative deep approaches. In the following, we mention a few of these aspects.

Firstly, the appearance component of a 3DMM often modeled by components of a linear texture space such as PCA which omits high-frequency information by its nature as shown in Fig. 1.1. Some others proposed to learn non-linear texture models by auto-encoders directly from in-the-wild images which unfortunately include severe noise due to various illumination and occlusions. In all cases, the quality of the facial texture reconstruction of the state-of-the-art methods is still limited for modeling textures with high-frequency details and photorealism. On the other hand, there is increasing literature on generating high-resolution photorealistic images by Generative Adversarial Networks (GANs). Within just a few years, there have been many approaches to generate up to $1024 \times 1024$ 2D face images that are fairly indistinguishable from real images. Nevertheless, as shown in Fig. 1.2, these approaches often generates images that are globally inconsistent as it is not possible to fully align images, i.e. by non-rigid registration, without a proper 3D prior. Although recent latent space manipulation techniques [KLA19, KLA+20] improves this aspect, its reconstruction ability is still limited due to limited network capacity (i.e. the network has to learn the pose, expression, illumination variation along with identity variation).

The energy function is another key aspect of accurate 3D reconstruction by 3DMM and learning-based approaches. The existing approaches rely on very primitive (i.e. pixel values, sparse landmarks, edges) [BV03, RV05, BSBW17, TZK+17] or hand-crafted features (i.e. HOG, SIFT) [BAP+17, BRV+18]. The primitive features are quite sensitive to "in-the-wild" conditions (i.e. lighting, pose, occlusion) and either too local or too global to maintain identity/expression resemblance with the target image. Although hand-crafted features offer more robust and mid-level supervision, deep learning is shown to be offering superior features for many other tasks. Thus, there is an unexplored potential to leverage 3D reconstruction by deep features specialized for facial patterns which are, in fact, well-provided by face recognition CNNs.

Another direction covered in this thesis is to generate realistic faces by 3D which is of high importance for many computer graphics and computer vision applications. This can be accomplished either by improving the realism of the 3D models directly or by augmenting rendered images from 3D

models to be more photorealistic. For 3D model photorealism, recent studies [SSK19b, GPKZ19] demonstrated that GANs can be used for generating high-quality textures of faces. Nevertheless, the generation process either omits the geometry and normals, or independent processes are used to produce 3D shape information. Moreover, shape, texture, and expression modalities are separately modeled by ignoring the correlation among them, placing a fundamental limit to the synthesization of semantically meaningful 3D faces. An alternative approach would be formulating this problem as an image-to-image translation from 3D rendered images to real images in order to bridge the gap between the two domains.

Finally, we are interested in robust and generalized expression reconstruction for personalized 3D models. Recently introduced *Codec Avatars* [LSSS18, WSS+19] have shown promising results for creating authentic 3D avatars from a multi-camera capture system. These personalized avatars can be animated by using headset mounted cameras (HMC) that are commercially available on VR headsets. The current reconstruction pipeline of [LSSS18, WSS+19] contains strong reliance on the training data and is sensitive to slight changes in perspective, lighting, background, and day-to-day facial attribute changes (e.g. make-up, facial hair, etc.). In a dataset of $\sim$200 subjects, where each subject's model is trained individually with a fixed background, lighting, pose, etc., there is still a room for improvement in robustness and generalization to these conditions by accumulating the variation in the whole dataset.

To sum up, we can summarize the challenges this thesis is concerned with the following keywords:

- Non-linear 3D Face Models
- Preserving identity and expression for 3D face reconstruction
- Photorealistic 3D face synthesis
- Robust and generalized reconstruction



Figure 1.2: Some cherry picked failure cases of Progressive Growing GAN [KALL18]

## 1.2 Contributions

In this section, we summarize the contribution of this thesis in line with the aforementioned scope and challenges. The main objective of the works presented in this thesis is to modernize multiple aspects of 3D face modeling and reconstruction with the recent deep generative and discriminative techniques.

In **Chapter 3**, we propose to replace linear texture models by a GAN trained on high-resolution UV maps as our statistical representation of the facial texture. Later, we show that it is possible to replace PCA-based texture models by this GAN model and reconstruct textures with high-frequency details and precise facial features.

GANs are shown to be very effective at capturing such details. However, they suffer from preserving 3D coherency [Goo16] of the target distribution when the training images are semi-aligned like 2D face images. At this point, we empirically observe that a GAN trained with UV representation of real textures addresses this problem and can generate realistic and coherent UVs almost all the time and generalize well. This is mainly thanks to the dense 3D registration of UV representations which nullify the need for the network to comprehend global structure. Therefore, we take advantage of this perfect harmony and train a progressive growing GAN [KALL18] to capture the distribution of UV representations of 10,000 high-resolution textures and use it as a texture model that replaces linear texture model.

**Chapter 3** also covers the novel deep cost functions and deep pipeline for 3D face reconstruction from a single image. Basically, instead of other cost functions used in the literature such as low-level $\ell_1$ or $\ell_2$ loss (e.g. RGB values [PB16], edges [RV05]) or hand-crafted features (e.g. SIFT [BAP+17]), we propose a novel cost function that is based on features extracted from the various layers of a face

recognition network. Unlike others, deep identity features are very powerful at preserving identity characteristics of the input image. This is because they are trained to be sensitive to the subtle details that are crucial to identify an individual while being robust to environmental conditions.

Moreover, we replace object-space rendering of 3DMM fitting approaches with an image-space differentiable renderer in a GPU accelerated deep learning framework [AAB+15]. The first benefit, GPU acceleration, provides cheaper derivatives through the above-mentioned deep architectures to keep running time reasonable. Also, to the best of our knowledge, image-space rendering is employed into 3DMM fitting for the first time in **Chapter 3**. This allows the usage of image-based deep feature extractors such as face recognition networks which we exploit as explained above. Although this concludes the modernization of optimization-based approaches by deep learning, we later extend it to regression-based approaches by a novel CNN architecture trained in a self-supervised manner.

In **Chapters 4 and 5**, we present two works to synthesize photorealistic faces by 3DMMs which are visually plausible and potentially useful as a synthetic training data. In **Chapter 4**, we propose to model and synthesize coherent 3D faces by jointly training a novel Trunk-branch based GAN architecture for shape, texture, and normals modalities. This architecture is specifically designed to maintain correlation while tolerating domain-specific differences of the three modalities and can be easily extended to other modalities. Additionally, the proposed network can be conditioned by expression to synthesize correlated shape, texture, and normals with expression.

In **Chapter 5**, we follow a different approach to synthesize; we randomly sample and render faces from a PCA-based 3DMM. Then, we bridge the gap between 3DMM and real images by a novel semi-supervised image-to-image translation network. The main novelty of this study is to exploit a small number of paired real and synthetic images, along with a large volume of unpaired data in a semi-supervised manner. As a result, the framework can generate photorealistic synthetic face images conditioned by samples rendered from a 3DMM. Both of these works are validated by synthesizing training data to boost face recognition performance.

In **Chapter 6**, we study several approaches to generalize *Codec Avatars* animation pipeline by head-mounted camera inputs. First, we augment the existing HMC training data in 3D space to simulate physical displacements of the VR headset. We change the perspective slightly by rotation and trans-

lation of the geometry and camera parameters optimized by [WSS+19]. Secondly, we combine the training data of multiple subjects to improve the robustness. This is done by designing a common expression encoder network trained by a self-supervised reconstruction loss of all subjects. The results show that the proposed approaches improve expression reconstruction under varying conditions meaning that the pipeline becomes more robust to the environment.

Finally, we draw conclusion and elaborate on future works in **Chapter 7**.

## 1.3 Publications

Some of the works in this thesis have been published in peer-reviewed conferences and journals. We aim to keep each chapter of this thesis as close to the original publication as possible to help the reader to refer to the relevant publications. For this reason, the chapters are narrated independently with their own introduction, related work, methodology, and experiments. Publications and their corresponding chapters are as follows:

- **Chapter 3:** Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, Long Beach, CA, USA, June 2019. IEEE

- **Chapter 3:** Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction. *Sumitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020

- **Chapter 4:** Baris Gecer, Alexander Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, 2020

- **Chapter 5:** Baris Gecer, Binod Bhattarai, Josef Kittler, Tae-Kyun Kim, Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised Adversarial Learning to Generate

Photorealistic Face Images of New Identities from 3D Morphable Model. In *European Conference on Computer Vision (ECCV)*, volume 11215, pages 230–248. Springer International Publishing, 2018

**Chapter 6:** contains the work that has been carried out during a *Research Internship at Facebook Reality Labs, Pittsburgh, USA*, in collaboration with Amin Jourabloo, Te-Li Wang, Fernando De la Torre, Yaser Sheikh.

# Chapter 2

# Background


**2.1 3D Face Modeling . . . . . 35**

2.1.1 Linear Models . . . . . 36

2.1.2 Non-Linear Generative Models . . . . . 38

2.1.3 Person-specific face models . . . . . 41

**2.2 2D & 3D Face Synthesis . . . . . 42**

2.2.1 Image Formation . . . . . 42

2.2.2 Differentiable Renderers . . . . . 45

**2.3 3D Face Reconstruction from a Single Image . . . . . 47**

2.3.1 Energy Functions . . . . . 48

2.3.2 Optimization-based Approaches . . . . . 51

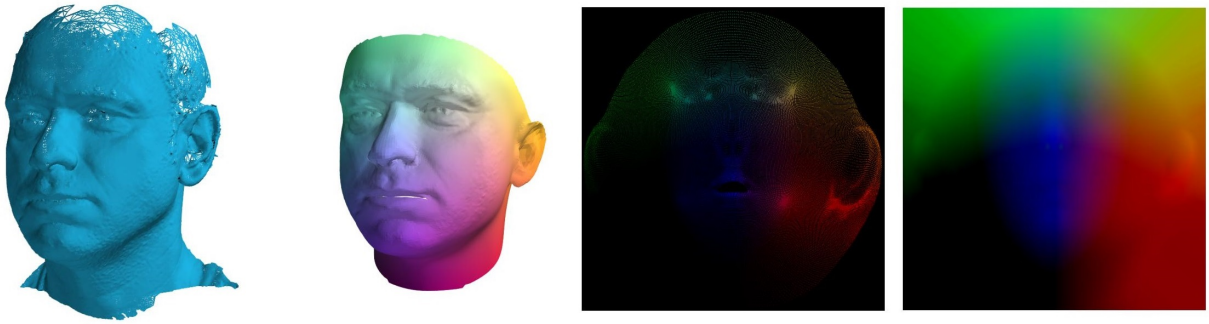
2.3.3 Regression-based Approaches . . . . . 52

**2.4 Conclusion . . . . . 54**


In this chapter, we briefly introduce definitions, concepts, and basic components regarding 3D face modeling and reconstruction for a smooth transition into the scope of this thesis. More specifically, we first define the conventional linear 3D face models in Section 2.1.1 and non-linear models in Section 2.1.2. then we outline the fundamentals of face synthesis by geometry and photometry properties

of mesh rendering as well as recent progress on differentiable renderers for 3D reconstruction. Finally, we explain 3D face reconstruction by optimization- and regression-based methods and give details about the commonly used energy functions in Section 2.3.

## 2.1 3D Face Modeling



(a) (b) (c) (d)

Figure 2.1: Illustration of 3D mesh pre-processing steps. In (a) we present a raw mesh, in (b) the registered mesh using the Large Scale Face Model (LSFM) template [BRZ+16], in (c) the unwrapped 3D mesh in the 2D UV space, and finally in (d) the interpolated 2D UV map. Interpolation is carried out using the barycentric coordinates of each pixel in the registered 3D mesh.

Every childbirth is a natural face generation that forms a face in real 3D space by bone, flesh, and skin. The length and size of different bones, the formation of muscles, pigment density of the skin, the angles of the joints are the basic factors that shape the manifold of the real human face distribution. From the data science point of view, the human face is a single class of objects with local geometry and appearance variation and the same global structure. A 3D face model is essentially a digital data representation that aims to capture this variation while maintaining topological structure and fine-correspondence among different faces. It can also be viewed as a dimensionality reduction of real human face distribution to a semantically meaningful manifold. In the following subsections, we briefly describe and formulate a range of 2D and 3D face models studied in the literature.

### 2.1.1 Linear Models

The seminal work of 3D morphable models (3DMMs) [BV99] has shown, for the first time, that is possible to learn a linear statistical model from a population of 3D faces [PS09, BSBW14]. The first public 3DMM model for identity variation was the Basel Face Model built from 200 people [PKA+09] by Principal Component Analysis (PCA) [WEG87] followed by recent large scale statistical models of face and head [BRZ+16, DPSD17]. Facial expressions are also modeled by applying PCA [YMW+11, LBB+17, BBC11, AKV08] or are manually defined using linear blendshapes [LWP10, TZN+15, BWP13].

Since the introduction of 3DMMs, most of the related work preferred PCA [WEG87] as a linear statistical analysis to construct a generative model. Briefly speaking, PCA constructs a manifold where each dimension captures one orthonormal basis (i.e. *eigenvectors*) of the distribution and is ordered by the magnitude of respective variation (i.e. *eigenvalues*). Practically, the data is normalized by subtracting the mean ($\mathbf{m}$) before computing the covariance matrix. Now, the *eigenvectors* of the covariance matrix can be interpreted as the basis of 3DMM denoted by $\mathbf{U}$. Usually, the first $n$ dimensions are kept in order to omit high-frequency signals which are mostly caused the noise from 3D capture systems. The resulting model can sample novel instances from this new parametric space ($\mathbf{p} \sim \mathcal{N}(0, \mathbf{I}_n)$) by simply $\mathbf{m} + \mathbf{U}\mathbf{p}$.

A complete 3D face model typically contains a deformable shape and its corresponding dense texture. Based on that, usually, the following three models are built: 1) *shape model* to model the variation in the identity space, 2) *expression model* to capture the geometry variation across different facial expressions, and 3) *texture model* for the appearance dressed on the vertices. In the following, we give more depth of these representations by linear models.

#### Shape and Expression Model

A standard face model is built by a set of aligned 3D faces and should generalize to the unseen samples. Thus, the first step is to establish dense point-to-point correspondences of 3D facial meshes to a reference topology by a non-rigid registration method [ARV07, YDXZ20]. A mesh comprises

a fixed set of $N$ vertices, such that $\mathbf{S} = \left[\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_N^\mathsf{T}\right]^\mathsf{T}$ where $\mathbf{x}_i \in \mathbb{R}^3$ denotes a vertex in a 3-dimensional space. And the topology ($\mathcal{T}$) is defined by $M$ triangles such that $\mathcal{T} = \left[\mathbf{x}_{\mathbf{t}_1}^\mathsf{T}, \ldots, \mathbf{x}_{\mathbf{t}_M}^\mathsf{T}\right]$ where $\mathbf{t}_i = (t_i^1, t_i^2, t_i^3) \in \mathbb{N}^3, t_i^j < N$ denotes the indices of $i$-th triangle's vertices. An example of such registration can be seen in Fig. 4.2 (a-b) where we illustrate a raw 3D scan and its registered 3D mesh to the LSFM [BRZ+16] template.

Usually, two separate linear models are built to represent identity and expression variations. The former is learned from facial scans displaying only neutral expression to capture identity variation and the latter is learned from displacement vectors of scans with a certain set of expressions. After the models are built by PCA as explained above, a sample mesh ($\mathbf{S}(\mathbf{p}_{s,e})$) can be reconstructed from linear combination of PCA bases according to the latent parameters ($\mathbf{p}_{s,e}$) by the following:

$$\mathbf{S}(\mathbf{p}_s, \mathbf{p}_e) \approx \mathbf{m}_{s,e} + \mathbf{U}_s\mathbf{p}_s + \mathbf{U}_e\mathbf{p}_e \tag{2.1}$$

where $\mathbf{m}_{s,e}(= \mathbf{m}_s + \mathbf{m}_e)$ is the mean shape vectors, $\mathbf{U}_s \in \mathbb{R}^{3N \times n_s}$ are the bases that correspond to identity variation, and $\mathbf{U}_e \in \mathbb{R}^{3N \times n_e}$ to expression variation. Finally, $\mathbf{p}_s \in \mathbb{R}^{n_s}$ and $\mathbf{p}_e \in \mathbb{R}^{n_e}$ are the latent parameters of the identity and expression bases.

### Texture Model

A triangular mesh can be textured either by *per-vertex colors* or by a *texture map in UV space* like in Fig. 1.1 (top). The *per-vertex color* method is relatively easy to implement by assigning a single RGB value to each vertex or triangle. On the other hand, *UV texture maps* can offer higher resolution by filling triangles with barycentric interpolation. UV maps are obtained by registration of 3D texture data into 2D planes with universal per-pixel alignment for all textures. A commonly used UV map is generated by cylindrical unwrapping the mean shape into a 2D flat space, which we use to create an RGB image $\mathbf{I}_{UV}$. Each vertex in the 3D space has a texture coordinate ($t_{coord}$) in the UV image plane as an example of it can be seen in Fig. 4.2 (c). The texture information can be stored in these maps at any arbitrary resolution and one can sample from it to acquire per-vertex colors as by $\mathbf{T} = \mathcal{P}(\mathbf{I}_{UV}, t_{coord})$. Likewise, one can interpolate per-vertex colors to acquire the texture in UV

space by $\mathbf{I}_{UV} = \mathcal{P}^{-1}(\mathbf{T}, t_{coord})$.

In order to define a statistical texture representation, all the training textures are vectorized and PCA is applied. Under this model, an arbitrary texture $\mathbf{T}^0$ can be approximated as a linear combination of the mean texture $\mathbf{m}_t$ and a set of basis $\mathbf{U}_t$ as follows:

$$\mathbf{T}(\mathbf{p}_t) \approx \mathbf{m}_t + \mathbf{U}_t\mathbf{p}_t \tag{2.2}$$

where $\mathbf{p}_t$ is the texture parameters for the sample $\mathbf{T}^0$.

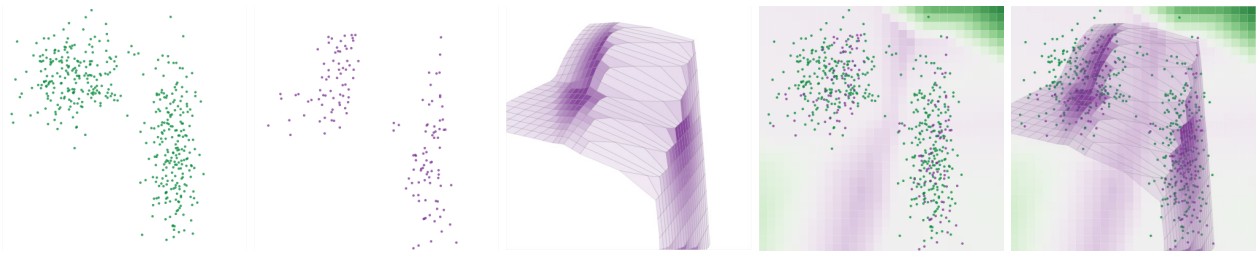
### 2.1.2 Non-Linear Generative Models

Most of the traditional 3D face works revolved around linear generative models that are constrained by a small number of anthropometric statistics (e.g. sparse face measurements in a population) [DMS98]. The recent 3D face fitting methods [TZG+18, TL18, GZC+16] still make use of similar statistical linear models. Linear models, despite their advantages, under-represents the high-frequency information and often result in overly-smoothed geometry and texture models. Hence, they can naturally represent only the low-frequency components of the facial texture.

In terms of preserving photorealism and high-frequency signals, non-linear generative models have been shown to be very successful in 2D image synthesis [MPK09, GPM+14, KW14, RMC16, Goo16, KALL18, KLA19, KLA+20, BSM17, GAA+17, ZXL+19, BDS19, VK20]. Therefore, after briefly mentioning traditional non-linear approaches, we introduce two families of works on deep non-linear generative models: Generative Adversarial Networks (GAN) [GPM+14] and Variational Auto-Encoders (VAE) [KW14].

#### Traditional Non-linear Models

There have been some bilinear and multilinear approaches for face modeling prior to deep learning [VT05, MMS08, BW15]. For example, [VT05] proposed a multifactor model to generalize Independent Component Analysis (ICA), called Multilinear-ICA (MICA). The study demonstrates that



(a) Target distribution (b) Generated (c) $\mathcal{G}$ manifold (d) $\mathcal{D}$ prediction (e) All

Figure 2.2: Toy experiment by GAN Lab [KTC+19] to illustrate capabilities of GAN for capturing sophisticated distributions. (a) Distribution of the target distribution. (b) Generated samples by GAN after training. (c) The generator's transformation manifold from noise space. Darker the areas, more likely to generate at that area. (d) The discriminator's decision space: Green for real, Purple for generated. (e) All of them overlayyed.

MICA can disentangle identity, pose, lighting, expression, etc. inherent to image formation and explicitly represents the higher-order statistics of each factor. Another study [MMS08] proposed to use bilinear models to jointly model identity and expression. The proposed approach can establish point correspondence based on an elastically deformable 3D model. Lastly, [BW15] jointly optimizes a multilinear model and the registration of 3D scans to build a high-quality face model. The study has shown that their method leads to more accurate correspondence and more efficient computation time than linear models.

### Generative Adversarial Networks

Generative Adversarial Networks (GANs) is a recently introduced family of techniques that train samplers of high-dimensional distributions by a minimax game between two networks [GPM+14]. It has been demonstrated that when a GAN is trained on facial images, it can generate photo-realistic high-resolution faces [KALL18, KLA19, BDS19, KLA+20].

GANs are designed to emulate real data distribution by training two competing networks simultaneously: *Generator* ($\mathcal{G}$) and *Discriminator* ($\mathcal{D}$). While the *Discriminator* is being trained to classify real samples ($x \sim p_{\text{data}}$) and the generated ones ($\mathcal{G}(\mathbf{z})$), the *Generator* is encouraged to produce samples that are indistinguishable by the *Discriminator*. Thus, a vanilla GAN training is often a zero-sum

game optimized by the following:

$$\theta_D = \arg\min_{\theta_D} \left[ -\mathbb{E}_{x \sim p_{\text{data}}} \log \mathcal{D}(x) - \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}))) \right] \tag{2.3}$$

$$\theta_G = \arg\min_{\theta_G} \left[ -\mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \log(\mathcal{D}(\mathcal{G}(\mathbf{z}))) \right] \tag{2.4}$$

where $\theta_D$ and $\theta_G$ are the parameters of the *Discriminator* and the *Generator* Networks, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is random noise as input to the *Generator* and $x \sim p_{\text{data}}$ is a random sample from the real data distribution.

Fig. 2.2 demonstrates a simple experiment on a toy dataset. Apparently, GANs are quite effective in terms of capturing sophisticated data distribution. In fact, linear models (i.e. PCA) would fail this experiment by generating samples mainly at the center of two clusters. Another such evidence is visible in Fig. 1.1 where we compare GAN generated (top) and PCA generated (bottom) UV textures of 3D faces.

GANs can be conditioned by a class, text, or image input as well [MO14, RAY+16, IZZE17]. Some other GAN-based studies allow to condition synthetic faces by rendered 3DMM images [SRK17], by landmarks [BJC18] or by another face image [BCW+18] (i.e. by disentangling identity and certain facial attributes). Similarly, the facial expression is also conditionally synthesized by an audio input [JCZ19], by action unit codes [PAM+18], by predefined 3D geometry [ZLGS03] or by expression of another face image [LDW+14].

Recently, GANs have been applied for generating 3D face components for various applications. In particular, [SSK19b] trained a GAN to synthesize facial textures, [MPN+19] trained a GAN that models face geometry based on UV representations for neutral faces and [WZX+16] synthesizes high-quality 3D objects by volumetric convolutional GANs.

#### Variational Auto-Encoders

Another recently popularized deep generative approach is Variational Auto-Encoders (VAE), firstly proposed by [KW14] and improved the generation quality over time [VK20]. Essentially, a VAE

is a probabilistic version of an auto-encoder (AE) which compresses a set of data to a latent space by sequential encoder and decoder networks. In VAE, the encoder network outputs a probability distribution over the latent space and a separate sampler generates a latent code from this distribution. The code is then reconstructed by the decoder network and the reconstruction loss computed and backpropagated during training. Compared to AE, VAEs give smoother and more generalized latent space, therefore the quality of novel generations is diverse and more visually plausible.

Similar to GANs, VAEs can be used for 2D image generation [vVk17, VK20] as well as 3D face modeling [BWS+18, LSSS18, WSS+19, RBSB18]. [BWS+18] models 3D facial shape variation hierarchically by different layers of the network from global to local features. [LSSS18, WSS+19] learn a personalized appearance model by encoding shape and texture jointly in a VAE space conditioned by pose. [RBSB18] models identity and expression geometry by variational autoencoders with mesh convolutions and [ZDKZ19] extends this approach to model both texture and geometry. In Chapter 6, we present an approach to improve personalized appearance models [LSSS18, WSS+19] by VAEs.

### 2.1.3 Person-specific face models

There have been several studies that propose to model the appearance and geometry of only a single or a few identities with superior quality. [GGW+05] utilizes appearance models for the identities whose large number of images captured in a controlled environment. Similarly, [LSSS18] propose deep appearance models by variational autoencoders. Even though it produces very high rendering quality with various expressions, this method also requires to capture ∼20 million images of subjects in a controlled environment. [CWW+16] and [NSX+19] reduce this number to a few in-the-wild images by interpolating different captures, but the quality of identity geometry and appearance is limited with the number and quality of the provided images.

## 2.2 2D & 3D Face Synthesis

3DMMs can synthesize novel 3D faces by sampling from the latent space of both linear or non-linear models. However, sampling from PCA space is not trivial. Closer to the mean vector, generated faces become not sufficiently distinctive from each other and further from it, they become unrealistic with overly-exaggerated facial features. Usually, keeping the latent vectors within the standard deviation gives reasonable results with the sacrifice of potential realistic faces beyond that point. A similar concept is introduced in GANs as *truncation trick* [KLA19, BDS19] which limits latent vector by a threshold to improve individual sample quality while reducing overall variation. One way or another one can generate 3D face of a novel identity with an arbitrary expression by $\mathbf{S}(\mathbf{p}_s, \mathbf{p}_e), \mathbf{T}(\mathbf{p}_t)$ as explained in Sec. 2.1. And this generated 3D face can either be used in a 3D platform like AR/VR or be rendered to synthesize many 2D face images by manipulating camera and illumination parameters.

From a computer graphics point of view, a photorealistic face rendering requires some elements to be tailored, i.e. shape, normals, and albedo maps, some of which should be specific to a particular identity. Currently, 3D faces rendered in computer games and movies are captured by expensive camera systems or by professional technical artists. 3DMMs can offer automated 3D face generation at large-scale, yet photorealism and level of details of the capturing systems are quite challenging to obtain by a generic model. This dilemma is one of the main issues that this thesis will cover in the following chapters.

### 2.2.1 Image Formation

Image formation from a 3D face, i.e. *Rendering*, is not only important to synthesize 2D face images but also required by 3D face reconstruction methods that are optimized or trained with analysis-by-synthesis approaches. 3D face rendering is an extensively studied subject in computer graphics and we briefly explain its basics in the following sections.

### Geometry

In this section, we briefly describe *perspective projection* in order to estimate the projections of every 3D face vertices to the 2D image plane. This camera model consists of two sets of parameters: *intrinsic* which refers to camera-specific parameters and *extrinsic* which consist of pose parameters that transform the world to camera coordinates. *intrinsic* parameters include focal lengths of both axes ($f_x, f_y \in \mathbb{R}_{>0}$ ) and the principal points $(c_x, c_y)$ which often correspond to the center of the image. Assuming that the camera view vector is perpendicular to the camera sensor, the intrinsic transformation matrix can be written as the following:

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \tag{2.5}$$

The extrinsic pose parameters usually consist of rotation ($\mathbf{R} \in \mathbb{R}^{3\times3}$) and translation $\mathbf{t} \in \mathbb{R}^{1\times3}$ matrices. And the final projection can be computed by the following:

$$\mathbf{s} = \mathbf{K}[\mathbf{R} \;\; \mathbf{t}]\mathbf{S}$$

where $\mathbf{s}$ denotes the 2D projection of the vertices $\mathbf{S}$.

### Photometry

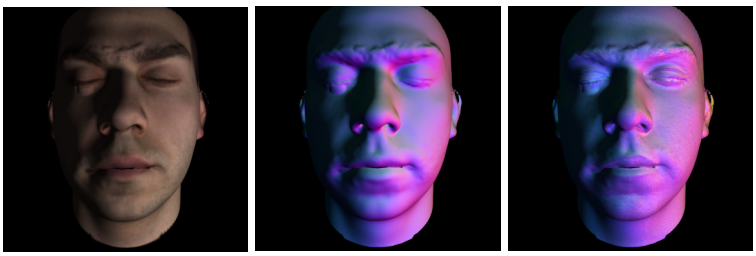
The final image is formed based on the 2D image projection of the geometry ($\mathbf{s} \in \mathbb{R}^{N\times2}$) as well as material properties (i.e. $[\mathbf{T}_{d,s}, \mathbf{N}_{d,s}]$), illumination (i.e. $[\mathbf{l}_a, \mathbf{l}_i, \hat{\mathbf{w}}_i]$), and the preferred reflectance model. Usually, a *Bidirectional Reflectance Distribution Function* (BRDF) is used to define reflectance of light on a facial surface which essentially depends on two direction vectors: one pointing to a light source ($\hat{\mathbf{w}}_d$), and the other to the camera ($\hat{\mathbf{w}}_c$). In a typical BRDF, there are four facial components to describe material properties: diffuse albedo ($\mathbf{T}_d$), specular albedo ($\mathbf{T}_s$), diffuse normals ($\hat{\mathbf{N}}_d$), and specular normals ($\hat{\mathbf{N}}_s$).



(a) Dif. Albedo ($\mathbf{T}_d$) (b) Spec. Albedo ($\mathbf{T}_s$) (c) Dif. Normals ($\hat{\mathbf{N}}_d$) (d) Spec. Normals($\hat{\mathbf{N}}_d$) (e) Shininess Coef.($\alpha$)



(f) $\mathbf{T}_d$ (g) $\mathbf{T}_d, \mathbf{T}_s, \alpha = 0.7$ (h) $\mathbf{T}_d, \mathbf{T}_s, \hat{\mathbf{N}}_d, \alpha = 0.7$ (i) $\mathbf{T}_d, \mathbf{T}_s, \hat{\mathbf{N}}_d, \hat{\mathbf{N}}_s, \alpha = 0.7$ (j) $\mathbf{T}_d, \mathbf{T}_s, \hat{\mathbf{N}}_d, \hat{\mathbf{N}}_s, \alpha =$(e)



(k) $\mathbf{T}_d$ rendered per-vertex (l) $\hat{\mathbf{N}}_d$ rendered per-vertex (m) $\hat{\mathbf{N}}_d$ rendered as UV

Figure 2.3: Progressive renderings by *Blinn-Phong Reflectance Model*. (a-e) BRDF components of a subject scanned in a light stage. (f-j) Full face and zoom-in rendering results by adding one component at a time. (k) Per-vertex diffuse albedo rendering. (l) Per-vertex diffuse normals rendering. (m) Diffuse normal rendering from UV map. All images rendered by a differentiable renderer implemented within the scope of this thesis.

Rendering starts with triangle rasterization to the image plane. This is achieved by barycentric interpolation of UV maps into the 2D projection of the mesh ($\mathbf{S}$) regarding the texture coordinates ($t_{coord}$). After that, we can render the image $\mathbf{I}^{\mathcal{R}} \in \mathbb{R}^{w \times h \times 3}$ by *Blinn-Phong Reflectance Model* as the following:

$$\mathbf{I}^{\mathcal{R}} = \mathcal{R}(\mathbf{S}, \mathbf{T}_{d,s}, \mathbf{N}_{d,s}, \mathbf{K}, [\mathbf{R} \;\; \mathbf{t}]) = k_a \mathbf{l}_a + \sum_{i=1}^{m} \left( k_d \mathbf{l}_i \mathbf{T}_d(|\hat{\mathbf{w}}_i \cdot \hat{\mathbf{N}}_d|^+) + k_s \mathbf{l}_i \mathbf{T}_s(|\frac{\hat{\mathbf{w}}_i + \hat{\mathbf{w}}_c}{||\hat{\mathbf{w}}_i + \hat{\mathbf{w}}_c||} \cdot \hat{\mathbf{N}}_s|^+)^\alpha \right) \tag{2.6}$$

where $\mathbf{l}_a \in \mathbb{R}^3$ denotes RGB ambient lighting intensity, $\mathbf{l}_i, \hat{\mathbf{w}}_i \in \mathbb{R}^3, i = \{1, \ldots, m\}$ denote RGB intensities and direction of the $m$ light sources respectively. Rectification and vector normalization functions are defined as $|*|^+ = \max(*, 0)$ and $\hat{*} = \frac{*}{||*||}$. Finally, $\alpha$ stands for the shininess coefficient of the face which can be defined as a constant value or as a UV map.

As can be seen in Fig. 2.3, *Blinn-Phong Reflectance Model* provides excellent shading for facial details and quite efficient to compute; nevertheless, many 3D face reconstruction approaches omit specular components, which becomes *Lambertian Reflectance Model*, and calculate diffuse normals ($\hat{\mathbf{N}}_d$) from the mesh ($\mathbf{S}$) to simplify optimization and avoid ambiguities. On the other side, BRDFs are short to describe the actual subsurface scattering, which can be better addressed by a *Bidirectional Subsurface Scattering Reflectance Distribution Function* (BSSRDF) [NRH+92, JMLH01]. That being said, BSSRDFs are considered to be very difficult to optimize in 3DMM fitting pipelines.

### 2.2.2 Differentiable Renderers

Every pixel in the rendered image is a weighted average based on barycentric coordinates as described in the previous section. Thus, the rendering function is continuous and differentiable. Even the original 3DMM fitting study [BV99] and others that follows it [RV05, BSBW17], optimized 3D reconstruction through object-space per-vertex differentiable renderer. For example, [BAP+17, BRV+18] only project vertices on image-plane and optimize based on per-vertex SIFT features. Recent studies employ differentiable renderers into deep learning frameworks to train 3DMM regression network [TZK+17, GCM+18] or to learn corrective space supplementary to 3DMM [TZG+18] or a complete non-linear 3DMM model [TL18, TL19, TBG+19] from 2D images.

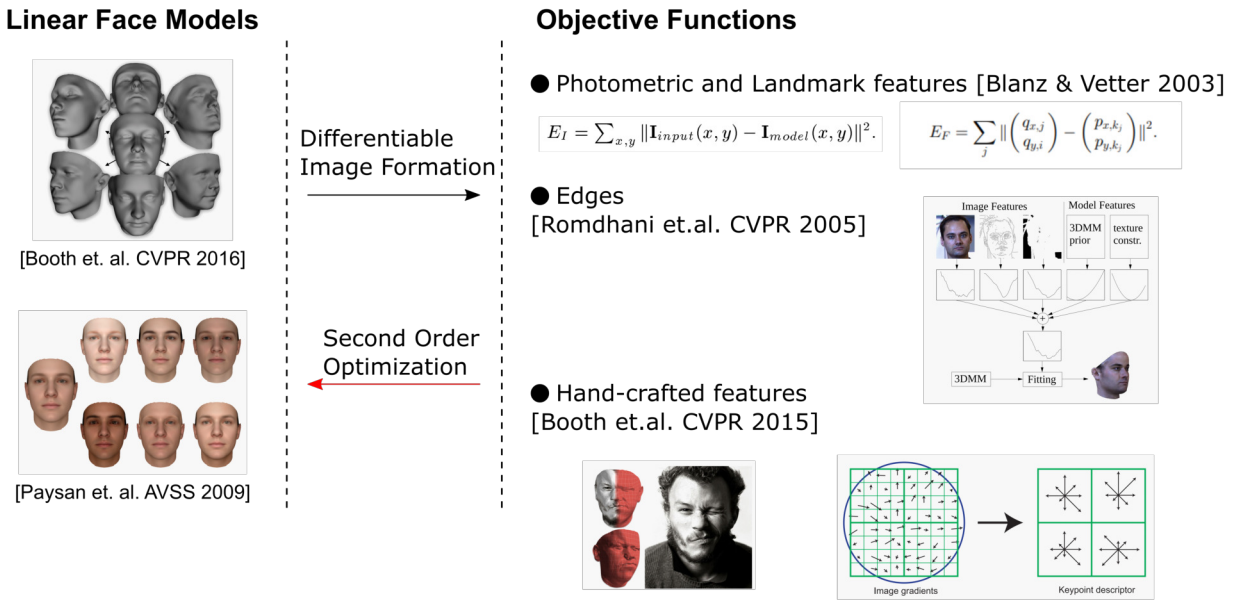



Figure 2.4: An overview of traditional linear 3D Morphable Model Fitting.

Object-space differentiable renderers optimize based on projecting the mesh to the target image and sampling image intensities or features for the visible vertices. Image-space renderers, however, directly render the reconstruction as an image, minimize its distance with the target image. The latter is more favorable as it supports high-resolution texture UV maps and feature extraction directly from the reconstruction image. To the best of our knowledge, we are the first to exploit image-based differentiable renderers for 3DMM fitting by an optimization-based approach. Unlike object-based renderer used in the previous fitting approaches [BV99, RV05, BSBW17, BAP+17, BRV+18], image-based renderer facilitates deep features as shown in Chapter 3.

Many of the renderers in the previous studies, omit specular components for simplicity, thus, they are unable to optimize based on the specularity of a target image. On the other hand, Fig. 2.3 illustrates renderings of a full BRDF differentiable renderer according to the *Blinn-Phong Reflectance Model* which is implemented as a part of this thesis.

## 2.3 3D Face Reconstruction from a Single Image

Estimation of the 3D facial geometry and appearance from single images is an interesting problem with many challenges and applications. The original study of Blanz and Vetter [BV99] has shown twenty years ago that it is possible to reconstruct 3D face by solving a non-linear optimization problem that is constrained by linear statistical models of facial texture and shape. Despite very promising results and advancements since then, there are still open challenges waiting to be fulfilled such as ambiguities such as camera-shape parameters and illumination-texture; robustness to occlusion, low-resolution, extreme pose, illumination, and expression; the disentanglement of expression and identity of 3D mesh and texture; missing correlation between shape, texture, and expression, etc.

The ambiguities can be addressed by limiting the capacities of models by regularization terms, which, in fact, means to keep the reconstruction closer to the mean face. This may or may not be a problem depending on the strength of the model. Booth et al. [BAP+17, BRV+18] addressed the ambiguities between illumination and texture by bypassing the texture model with a SIFT feature model which is robust to illumination. Kyle et al. [GCM+18] proposed to use deep identity features to improve robustness to occlusion, illumination, pose, and expression. Learning-based approaches are often more robust and generalized but they cannot iteratively refine the reconstruction. These are just a small subset of achievements in the literature and the challenges that are waiting to be explored more deeply.

Essentially, 3D face reconstruction can be formulated as optimization of model parameters or learning a regression network. We can summarize the mainstream methodology for both optimization-based and learning-based 3D face reconstruction approaches by the following: (1) a 3D face is synthesized by given latent parameters of 3D shape and texture models, (2) then it is rendered with the camera and illumination parameters by a differentiable renderer, (3) the reconstruction error of the target image is computed by the preferred metrics, and (4) the regression network or parameter fitting is updated based on the derivatives of the error. Finally, after some iterations, the latent parameters or the regression network can generate the 3D reconstruction of the/a target image. In the previous section, we have covered models and image formation, and the following sections will cover the reconstruction error and optimization/learning pipelines.

### 2.3.1 Energy Functions

In order to supervise the optimization or training, we can define one or more distance metrics $\mathcal{L}(\mathbf{P}, \mathbf{I}^0) : (\bullet, \mathbb{R}^{w\times h\times 3}) \rightarrow \mathbb{R}$ between the target image ($\mathbf{I}^0 \in \mathbb{R}^{w\times h\times 3}$) and the rendering of the model ($\mathcal{R}(\mathbf{P}) \in \mathbb{R}^{w\times h\times 3}$). The synthesized image $\mathcal{R}(\mathbf{P})$ is either depends on 3DMM parameters ($\mathbf{P} = [\mathbf{p}_{s,e}, \mathbf{p}_t, \mathbf{p}_c, \mathbf{p}_l]$ [1]) in optimization-based approaches or depends on regression network parameters ($\mathbf{P} = \mathbf{\Theta}$) in regression-based approaches. A good metric should be differentiable, robust to environmental factors, and sensitive to prominent facial features. Below we list common distance metrics in the literature:

#### Photometric distance

One of the most primitive and widely used metrics is *photometric distance*. It is mostly useful to reconstruct appearance as its receptive field is only 1 pixel and it heavily relies on the assumption that shape and camera parameters are more or less converged.

While it is relatively simple, it can optimize texture to some extent, lighting parameters such as ambient colors, direction, distance, and color of a light source. The photometric distance,which is also called *pixel loss*, is defined by pixel level $\ell_1$ or $\ell_2$ loss functions as respectively:

$$\mathcal{L}_{\text{pix}_1}(\mathbf{P}, \mathbf{I}^0) = \sum_{(w,h,3)} \frac{||\mathbf{I}^0 - \mathcal{R}(\mathbf{P})||_1}{w \times h \times 3} \tag{2.7}$$

$$\mathcal{L}_{\text{pix}_2}(\mathbf{P}, \mathbf{I}^0) = \sum_{(w,h,3)} \frac{||\mathbf{I}^0 - \mathcal{R}(\mathbf{P})||_2^2}{w \times h \times 3} \tag{2.8}$$

The main advantage of these functions is their simplicity, e.g. they can be applied at high resolutions very efficiently. Especially, $\ell_1$ loss is good at capturing fine, local patterns. In order to solve the locality problem of these functions, recently there have been some soft rasterization approaches [LLCL19, ZWC+20] proposed to increase the field of view of this metric. Nevertheless, it is still too sensitive to misalignments and illumination and fails to factorize global facial features effectively.

[1] $\mathbf{p}_c$ and $\mathbf{p}_l$ denotes camera and illumination parameters respectively

### Landmarks

The early 3DMM fitting studies had two essential supervision: photometric distance (which is dense and local) and landmarks distance (which is sparse and global). In a way, these two losses are complementary as the former provides bottom-up and the latter provides top-down information. The landmarks distance metric detects the landmark locations of the input image and aligns the rendered geometry onto it by updating the shape, expression, and camera parameters. That is, camera parameters are optimized to align with the pose of image $\mathbf{I}$ and geometry parameters are optimized for the rough shape estimation.

Let us say we can estimate landmarks of the target image ($\mathbf{I}_0$) by a landmark detection network such that $\mathcal{M}(\mathbf{I}) : \mathbb{R}^{w \times h \times 3} \rightarrow \mathbb{R}^{68 \times 2}$. The landmarks of a 3D face can be acquired by a predefined set of vertices in the mesh ($\mathbf{S}_{\text{lan}} \in \mathbb{R}^{68 \times 3}$) that corresponds to standardized landmarks (e.g. iBUG-68 landmarks), thus their projection ($\mathbf{s}_{\text{lan}} \in \mathbb{R}^{68 \times 2}$) can be used in the metric. The alignment error is achieved by point-to-point euclidean distances between the detected landmark locations of the input image and the projection of the landmark vertices as the following:

$$\mathcal{L}_{\text{lan}}(\mathbf{P}, \mathbf{I}^0) = \sum_{j}^{68} ||\mathcal{M}(\mathbf{I}^0)_j - \mathbf{s}_{\text{lan}_j}||_2^2 \tag{2.9}$$

### Hand-crafted Features

For robust reconstruction, a recent approach [BAP+17] proposed to use handcrafted features, such as SIFT [Low99] or HOG [DT05] and fit based on these features. Let us denote feature extraction function by $\mathcal{H}$, the proposed energy function would be as follows:

$$\mathcal{L}_{\text{SIFT}}(\mathbf{P}, \mathbf{I}^0) = ||\mathcal{H}(\mathbf{I}^0) - \mathcal{H}(\mathcal{R}(\mathbf{P}))||_2 \tag{2.10}$$

Despite the equation above, the original method [BAP+17] did not need to render the reconstruction in the image-space as it is based on a texture model built by SIFT features. Thus it bypassed the ambiguity between the true albedo and illumination. The main drawback of the original method is that the facial texture is not reconstructed, however, this can be resolved by extracting such features

differentiably on the image-space rendered reconstructions (i.e. $\mathcal{H}(\mathcal{R}(\mathbf{P}))$), as we explain in the next section.

### Deep Identity Features

With the availability of large scale datasets, CNNs have shown incredible performance on many face recognition benchmarks. Their strong identity features are robust to many variations including pose, expression, illumination, age, etc. A recent study [GCM+18] has shown that they are also effective at face reconstruction by learning the correlation between identity features and 3DMM parameters. This can be generalized by the following cosine distance between the target image and the rendering of the reconstruction:

$$\mathcal{L}_{\text{id}}(\mathbf{P}, \mathbf{I}^0) = 1 - \frac{\mathcal{F}^n(\mathbf{I}^0).\mathcal{F}^n(\mathcal{R}(\mathbf{P}))}{||\mathcal{F}^n(\mathbf{I}^0)||_2||\mathcal{F}^n(\mathcal{R}(\mathbf{P}))||_2} \tag{2.11}$$

where a pretrained face recognition network with $n$ layers denoted by $\mathcal{F}^n(\mathbf{I}) : \mathbb{R}^{w\times h\times 3} \rightarrow \mathbb{R}^{512}$. This way, one can take advantage of an off-the-shelf face recognition network in order to optimize the reconstruction based on the identity-related features of a target face image.

### Regularization terms

Linear and non-linear face models provide statistical prior in addition to the above-mentioned energy functions. This prior keep the reconstruction within the manifold of the training set used to build the model. Especially, it encourages avoiding overfitting to particular pose and disentanglement of texture-illumination and expression-shape ambiguities. Adjusting the weight of this term in the total energy function is not trivial. Higher weight leads to getting close to the mean face, lower weight might result in exaggerated facial patterns. Further discussions on this dilemma can be found in [LMAR14, PS16]. We also refer the reader to the other solutions proposed to address this issue in the literature [PS12, RSOK17, TZB+18]. For linear 3DMM, regularization can be achieved by minimizing latent parameters (e.g. $\mathbf{p_s}$) with respect to the associated variance of the principal components, as

follows:

$$\mathcal{L}_{\text{reg}}(\mathbf{p}) = ||\mathbf{p}||^2_{\mathbf{A}} = \mathbf{p}^T\mathbf{A}\mathbf{p} \tag{2.12}$$

where $\mathbf{A}$ stands for the orthogonal space to the statistical model of the texture.

### 2.3.2 Optimization-based Approaches

For many years optimization-based 3DMM fitting and its variants were the methods of choice for 3D face reconstruction [BV99, ZTG+18, EST+20]. Ever since the original work [BV99], there have been many studies improving the generalization and the precision of 3D face reconstruction by adding other sources of information such as edges [RV05, BSBW17, HYK+17], specular highlights [RV05], and so on. [TZS+19] follows a standard linear 3DMM fitting strategy followed by a number of steps for texture completion and refinement. Furthermore, with appropriate statistical texture models on image features such as Scale Invariant Feature Transform (SIFT) and Histogram Of Gradients (HOG), 3DMM-based methodologies can achieve great performance in 3D shape estimation on images captured under unconstrained conditions [BAP+17] or videos [BRV+18].

Optimization-based, or in other name *analysis-by-synthesis* approaches can reconstruct 3D facial geometry and texture by fitting a 3DMM on a given target image $\mathbf{I}^0$. This is often achieved by solving a non-linear energy-based cost optimization problem that recovers a set of parameters $\mathbf{P} = [\mathbf{p}_{s,e}, \mathbf{p}_t, \mathbf{p}_c, \mathbf{p}_l]$. The optimization can be formulated as:

$$\min_{\mathbf{P}} \mathcal{E}(\mathbf{P}, \mathbf{I}^0) = \sum_i \lambda_i \mathcal{L}_i(\mathbf{P}, \mathbf{I}^0) + \text{Reg}(\mathbf{P}) \tag{2.13}$$

where $i$ can be any of the metrics defined above or a novel energy function. $\lambda_i$ denotes the balancing factor to control the contribution of different metrics to the total energy. Finally, Reg is the regularization term to keep texture and shape parameters in a realistic range of PCA.

The analysis-by-synthesis approach iteratively updates 3DMM parameters based on the derivatives of Eq. 2.13. The optimization can be done by any gradient-based solvers such as the Gradient De-

scent [BV99], Gauss-Newton [BAP+17, PB16], Levenberg-Marquardt [RV05], or Iteratively Reweighted Least Squares (IRLS) [HW77]. Another optimization approach [SEMV17] introduces stochastic sampling method based on Bayesian inference. This method searches for the posterior distribution of the model parameters rather than solving for an optimal solution and it is based on proposing a set of parameters and verifying them with the model (i.e. Data-Driven Markov Chain Monte Carlo). In contrast to the gradient-based methods, it does not need gradients, thus less sensitive to local minima. Moreover, since it is not gradient-based, it makes use of facial features [Low99] without the need for differentiable image formation.

### 2.3.3 Regression-based Approaches

Recent advances in deep learning popularized regression-based 3D face reconstruction approaches. Similar to other deep learning tasks, the goal is to regress 3DMM parameters (or 3D face directly) by a deep neural network. Deep networks are very powerful to learn sophisticated tasks when proper supervision and data are provided. However, providing such is not trivial, because accurate estimation of 3D shape and texture can be done under controlled environments which has some domain gap with 'in-the-wild' conditions. Below we discuss and refer to some techniques used in the literature.

#### Supervised Learning

Neural networks can easily learn a well-defined task when trained with a well-annotated dataset. 3D face reconstruction task is no exception. However, the 3D face reconstruction task aims for face images taken in arbitrary conditions, thus the regression network should be robust to environmental changes such as lighting, occlusion, perspective, and even camera parameters.

The simplest solution is to collect a large scale annotated 3D face dataset to train a regression CNN. However, acquiring such a dataset under unconstrained settings is challenging. [OLSL16] manually annotated expression blendshapes of video clips by professional animators. [LKA+17] proposed to train based on images captured in a controlled environment. These two studies demonstrate the difficulty of direct supervision as the former is very costly and the latter contains a domain gap with

in-the-wild images. Some other datasets are semi-automatically annotated by landmarks [STZP13b, ZLL+16]. The early CNNs [ZLL+16, YYS+17a] trained by these datasets have shown promising results.

Moreover, it is safe to assume the facial shape and texture of an individual are consistent across videos and multiple images. 3D Menpo dataset [ZCR+17] takes this advantage to fine-tune 3D landmark annotations by aggregation of an individual in all frames of a video. Similarly, [THMM17] relies on the same assumption by first fitting 3DMM individually for multiple images of the same person and then pool across the parameters to train a deep neural network.

#### Self-supervised Learning

Another approach is to utilize random in-the-wild face images to train a network by reconstruction consistency. Typically, the 3D reconstruction is rendered by a differentiable renderer and the difference between the input image and the rendered image is penalized during the training. [TZK+17] trains by 2D landmark coordinates and pixel intensities as supervision, while [RSOK17] learns fine-details in texture in a similar fashion. [GCM+18] trains a CNN by matching deep identity features between the training image and the rendered reconstruction. These self-supervised approaches allow to exploit large-scale face datasets without annotation, however, energy functions should be carefully designed for robust and precise reconstruction.

#### Learning from Synthetic data

Without a doubt, the ground truth provided by synthetic data has absolute accuracy, yet the data itself has a domain gap with real face images. There have been many approaches for synthetic training such as masking [RSK16, RSOK17], self-supervised bootstrapping [KZT+18], pretraining the network with synthetic data and fine-tuning with ready data [GCM+18], or training image-to-image domain transfer networks [SRK17].

## 2.4 Conclusion

This chapter covered the basics of 3D face modeling, reconstruction, and synthesis in the conventional methods and the recent studies to prepare for the upcoming chapters. The main components necessary for these tasks are gathered under three main sections: '3D Face Modeling', '2D & 3D Face Synthesis', and '3D Face Reconstruction from a Single Image'. Each of these sections gives a sneak peek of a well-established problem domain with a large literature in the context of recent advances of deep learning. Nevertheless, the isolation and the formulation of these problems as presented here, help us to contemplate how each module can be improved or replaced by its more advanced alternative in the following chapters. The remainder of the thesis particularly focuses on incorporating deep generative and discriminative learning into 3D face modeling, reconstruction, and synthesis.

# Chapter 3

# Generative Adversarial Network Regression and Fitting for 3D Face Reconstruction

---


**3.1 Introduction . . . . . 56**

**3.2 History of 3DMM Fitting . . . . . 59**

3.2.1 3DMM representation . . . . . 59

3.2.2 Fitting . . . . . 61

**3.3 Approach . . . . . 62**

3.3.1 GAN Texture Model . . . . . 63

3.3.2 Differentiable Renderer . . . . . 64

3.3.3 Cost Functions . . . . . 65

3.3.4 Model Fitting . . . . . 67

**3.4 Drawbacks and Improvements on GANFit . . . . . 68**

3.4.1 Fitting with Multiple Images (i.e. Video): . . . . . 68

3.4.2 Regressing parameters by Encoder . . . . . 68

**3.5 Experiments . . . . . 71**

3.5.1 Implementation Details . . . . . 71

3.5.2 Qualitative Results . . . . . 71
3.5.3 Quantitative Experiments . . . . . 74
3.5.4 Ablation Study . . . . . 79
**3.6 Conclusion . . . . . 80**

In this chapter, we propose an approach that harnesses the power of Generative Adversarial Networks (GANs) and DCNNs in order to reconstruct the facial texture and shape from single images. That is, we utilize GANs to train a very powerful generator of facial texture in UV space. Then, we revisit the original 3D Morphable Models (3DMMs) fitting approaches making use of non-linear optimization to find the optimal latent parameters that best reconstruct the test image but under a new perspective. We optimize the parameters with the supervision of pretrained deep identity features through our end-to-end differentiable framework. We demonstrate excellent results in photorealistic and identity preserving 3D face reconstructions and achieve for the first time, to the best of our knowledge, facial texture reconstruction with high-frequency details.

## 3.1 Introduction

Estimation of the 3D facial surface and other intrinsic components of the face from single images (e.g. albedo, etc.) is a very important problem at the intersection of computer vision and machine learning with countless applications (e.g. face recognition, face editing, virtual reality). It is now twenty years from the seminal work of Blanz and Vetter [BV99] which showed that it is possible to reconstruct shape and albedo by solving a non-linear optimization problem that is constrained by linear statistical models of facial texture and shape. This statistical model of texture and shape is called a 3D Morphable Model (3DMM). Arguably the most popular publicly available 3DMM is the Basel model built from 200 people [PKA+09]. Recently, large scale statistical models of face and head shape have been made publicly available [BRZ+16, DPSD17].

For many years 3DMMs and its variants were the methods of choice for 3D face reconstruction [RV05, ZLL+16, JZD+18]. Furthermore, with appropriate statistical texture models on image fea-

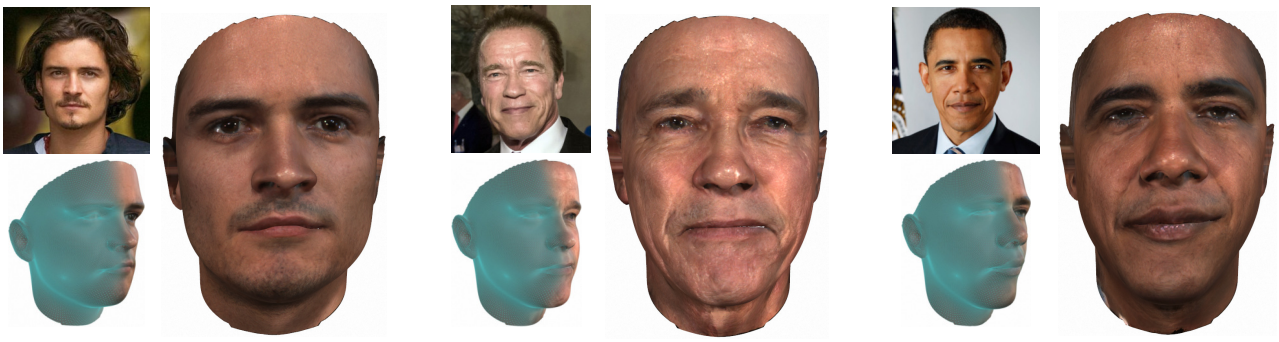


Figure 3.1: The proposed deep fitting approach can reconstruct high quality texture and geometry from a single image with precise identity recovery. The reconstructions in the figure and the rest of the chapter are represented by a vector of size 700 floating points and rendered without any special effects. We would like to highlight that the depicted texture is reconstructed by our model and none of the features taken directly from the image.

tures such as Scale Invariant Feature Transform (SIFT) and Histogram Of Gradients (HOG), 3DMM-based methodologies can still achieve state-of-the-art performance in 3D shape estimation on images captured under unconstrained conditions [BAP+17]. Nevertheless, those methods [BAP+17] can reconstruct only the shape and not the facial texture. Another line of research in [YSN+18, SWH+17] decouples texture and shape reconstruction. A standard linear 3DMM fitting strategy [TZS+19] is used for face reconstruction followed by a number of steps for texture completion and refinement. In these papers [SWH+17, YSN+18], the texture looks excellent when rendered under professional renderers (e.g. Arnold), nevertheless when the texture is overlaid on the images the quality significantly drops [1].

In the past two years, a lot of work has been conducted on how to harness Deep Convolutional Neural Networks (DCNNs) for 3D shape and texture reconstruction. The first such methods either trained regression DCNNs from image to the parameters of a 3DMM [THMM17] or used a 3DMM to synthesize images [RSK16, GZC+19] and formulate an image-to-image translation problem using DCNNs to estimate the depth[2] [SRK17]. The more recent unsupervised DCNN-based methods are trained to regress 3DMM parameters from identity features by making use of differentiable image formation architectures [CBK+17] and differentiable renderers [GCM+18, TZK+17, RSOK17].

The most recent methods such as [TZG+18, TL18, GZC+16] use both the 3DMM model, as well as additional network structures (called correctives) in order to extend the shape and texture representa-

[1] Please see the supplementary materials for a comparison with [SWH+17, YSN+18].
[2] The depth was afterwards refined by fitting a 3DMM and then changing the normals by using image features.

tion. Even though the paper [TZG+18] shows that the reconstructed facial texture has indeed more details than a texture estimated from a 3DMM [THMM17, TZK+17], it is still unable to capture high-frequency details in texture and subsequently many identity characteristics (please see the Fig. 5.5). Furthermore, because the method permits the reconstructions to be outside the 3DMM space, it is susceptible to outliers (e.g. glasses etc.) which are baked in shape and texture. Although rendering networks (i.e. trained by VAE [LSSS18]) generates outstanding quality textures, each network is capable of storing up to few individuals whom should be placed in a controlled environment to collect $\sim$20 millions of images.

In this chapter, we still propose to build upon the success of DCNNs but take a radically different approach for 3D shape and texture reconstruction from a single in-the-wild image. That is, instead of formulating regression methodologies or auto-encoder structures that make use of self-supervision [TZG+18, GCM+18, TL18], we revisit the optimization-based 3DMM fitting approach by the supervision of deep identity features and by using Generative Adversarial Networks (GANs) as our statistical parametric representation of the facial texture.

In particular, the novelties that this chapter brings are:

- We show for the first time, to the best of our knowledge, that a large-scale high-resolution statistical reconstruction of the complete facial surface on an unwrapped UV space can be successfully used for reconstruction of arbitrary facial textures even captured in unconstrained recording conditions[3].
- We formulate a novel 3DMM fitting strategy which is based on GANs and a differentiable renderer.
- We devise a novel cost function which combines various content losses on deep identity features from a face recognition network.
- We demonstrate excellent facial shape and texture reconstructions in arbitrary recording condi-

[3]In the very recent works, it was shown that it is feasible to reconstruct the non-visible parts a UV space for facial texture completion[DCX+18] and that GANs can be used to generate novel high-resolution faces[SSK19b]. Nevertheless, our work is the first one that demonstrates that a GAN can be used as powerful statistical texture prior and reconstruct the complete texture of arbitrary facial images.

tions that are shown to be both photorealistic and identity preserving in qualitative and quantitative experiments.

## 3.2 History of 3DMM Fitting

Our methodology naturally extends and generalizes the ideas of texture and shape 3DMM using modern methods for representing texture using GANs, as well as defines loss functions using differentiable renderers and very powerful publicly available face recognition networks [DGXZ19]. Before we define our cost function, we will briefly outline the history of 3DMM representation and fitting.

### 3.2.1 3DMM representation

The first step is to establish dense correspondences between the training 3D facial meshes and a chosen template with fixed topology in terms of vertices and triangulation.

#### Texture

Traditionally 3DMMs use a UV map for representing texture. UV maps help us to assign 3D texture data into 2D planes with universal per-pixel alignment for all textures. A commonly used UV map is built by cylindrical unwrapping the mean shape into a 2D flat space formulation, which we use to create an RGB image $\mathbf{I}_{UV}$. Each vertex in the 3D space has a texture coordinate $t_{coord}$ in the UV image plane in which the texture information is stored. A universal function exists, where for each vertex we can sample the texture information from the UV space as $\mathbf{T} = \mathcal{P}(\mathbf{I}_{UV}, t_{coord})$.

In order to define a statistical texture representation, all the training texture UV maps are vectorized and Principal Component Analysis (PCA) is applied. Under this model any test texture $\mathbf{T}^0$ is approximated as a linear combination of the mean texture $\mathbf{m}_t$ and a set of bases $\mathbf{U}_t$ as follows:

$$\mathbf{T}(\mathbf{p}_t) \approx \mathbf{m}_t + \mathbf{U}_t\mathbf{p}_t \tag{3.1}$$

where $\mathbf{p}_t$ is the texture parameters for the text sample $\mathbf{T}^0$. In the early 3DMM studies, the statistical model of the texture was built with few faces captured in strictly controlled conditions and was used to reconstruct the test albedo of the face. Since, such texture models can hardly represent faces captured in uncontrolled recording conditions (in-the-wild). Recently it was proposed to use statistical models of hand-crafted features such as SIFT or HoG [BAP+17] directly from in-the-wild faces. The interested reader is referred to [BV03, RBV02] for more details on texture models used in 3DMM fitting algorithms.

The recent 3D face fitting methods [TZG+18, TL18, GZC+16] still make use of similar statistical models for the texture. Hence, they can naturally represent only the low-frequency components of the facial texture (please see Fig. 5.5).

**Shape**

The method of choice for building statistical models of facial or head 3D shapes is still PCA [WEG87]. Assuming that the 3D shapes in correspondence comprise of $N$ vertexes, i.e. $\mathbf{s} = \left[\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_N^\mathsf{T}\right]^\mathsf{T} = [x_1, y_1, z_1, \ldots, x_N, y_N, z_N]^\mathsf{T}$. In order to represent both variations in terms of identity and expression, generally two linear models are used. The first is learned from facial scans displaying the neutral expression (i.e. representing identity variations) and the second is learned from displacement vectors (i.e. representing expression variations). Then a test facial shape $\mathbf{S}(\mathbf{p}_{s,e})$ can be written as

$$\mathbf{S}(\mathbf{p}_{s,e}) \approx \mathbf{m}_{s,e} + \mathbf{U}_{s,e}\mathbf{p}_{s,e} \tag{3.2}$$

where $\mathbf{m}_{s,e}$ in the mean shape vector, $\mathbf{U}_{s,e} \in \mathbb{R}^{3N \times n_{s,e}}$ is $\mathbf{U}_{s,e} = [\mathbf{U}_s, \mathbf{U}_e]$ where the $\mathbf{U}_s$ are the bases that correspond to identity variations, and $\mathbf{U}_e$ the bases that correspond to expression. Finally, $\mathbf{p}_{s,e}$ are the $n_{s,e}$ shape parameters which can be split accordingly to the identity and expression bases: $\mathbf{p}_{s,e}$ = $[\mathbf{p_s}, \mathbf{p_e}]$.

### 3.2.2 Fitting

3D face and texture reconstruction by fitting a 3DMM is performed by solving a non-linear energy based cost optimization problem that recovers a set of parameters $\mathbf{p} = [\mathbf{p}_{s,e}, \mathbf{p}_t, \mathbf{p}_c, \mathbf{p}_l]$ where $\mathbf{p}_c$ are the parameters related to a camera model and $\mathbf{p}_l$ are the parameters related to an illumination model. The optimization can be formulated as:

$$\min_{\mathbf{p}} \mathcal{E}(\mathbf{p}) = ||\mathbf{I}^0(\mathbf{p}) - \mathbf{W}(\mathbf{p})||_2^2 + \text{Reg}(\{\mathbf{p}_{s,e}, \mathbf{p}_t\}) \tag{3.3}$$

where $\mathbf{I}^0$ is the test image to be fitted and $\mathbf{W}$ is a vector produced by a physical image formation process (i.e. rendering) controlled by $\mathbf{p}$. Finally, Reg is the regularization term that is mainly related to texture and shape parameters.

Various methods have been proposed for numerical optimization of the above cost functions [HYK+17, BSBW17]. A notable recent approach is [BAP+17] which uses handcrafted features (i.e. $\mathbf{H}$) for texture representation simplified the cost function as:

$$\min_{\mathbf{p}^r} \mathcal{E}(\mathbf{p}^r) = ||\mathbf{H}(\mathbf{I}^0(\mathbf{p}^r)) - \mathbf{H}(\mathbf{W}(\mathbf{p}^r))||_{\mathbf{A}}^2 + \text{Reg}(\mathbf{p}_{s,e}) \tag{3.4}$$

where $||\mathbf{a}||_{\mathbf{A}}^2 = \mathbf{a}^T\mathbf{A}\mathbf{a}$, $\mathbf{A}$ is the orthogonal space to the statistical model of the texture and $\mathbf{p}^r$ is the set of reduced parameters $\mathbf{p}^r = \{\mathbf{p}_{s,e}, \mathbf{p}_c\}$. The optimization problem in Eq. 3.4 is solved by Gauss-Newton method. The main drawback of this method is that the facial texture in not reconstructed.

In this chapter, we generalize the 3DMM fittings and introduce the following novelties:

- We use a GAN on high-resolution UV maps as our statistical representation of the facial texture. That way we can reconstruct textures with high-frequency details.
- Instead of other cost functions used in the literature such as low-level $\ell_1$ or $\ell_2$ loss (e.g. RGB values [PB16], edges [RV05]) or hand-crafted features (e.g. SIFT [BAP+17]), we propose a novel cost function that is based on feature loss from the various layers of publicly available face recognition embedding network [DGXZ19]. Unlike others, deep identity features are very powerful at preserving identity characteristics of the input image.

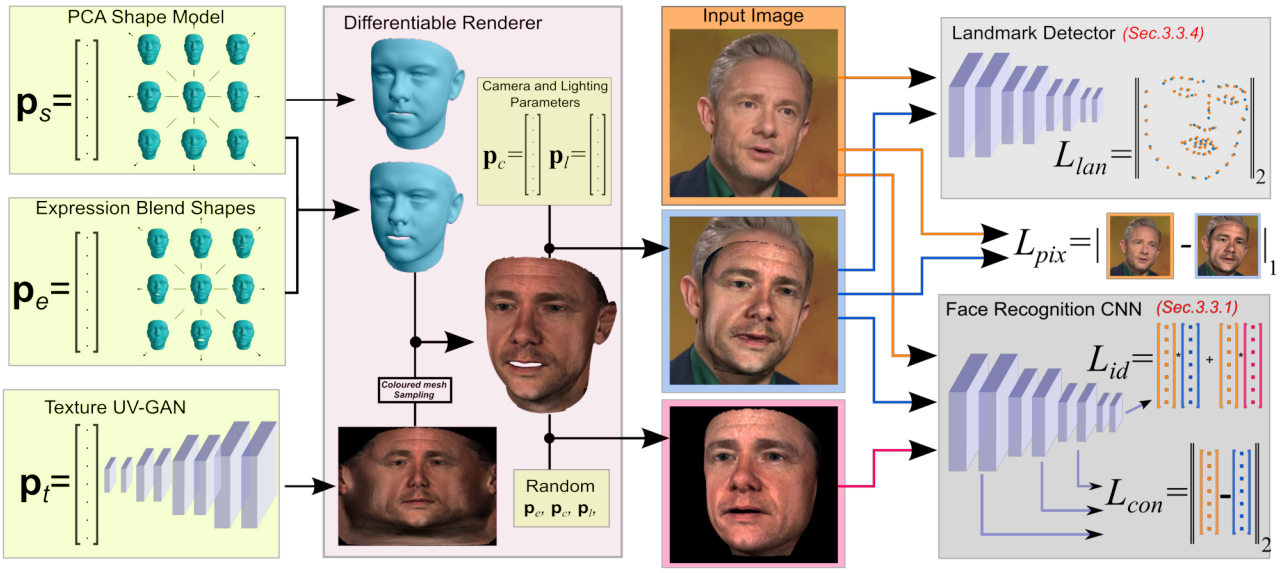



Figure 3.2: Detailed overview of the proposed approach. A 3D face reconstruction is rendered by a differentiable renderer (shown in purple). Cost functions are mainly formulated by means of identity features on a pretrained face recognition network (shown in gray) and they are optimized by flowing the error all the way back to the latent parameters ($p_s, p_e, p_t, c, i$, shown in green) with gradient descent optimization. End-to-end differentiable architecture enables us to use computationally cheap and reliable first order derivatives for optimization thus making it possible to employ deep networks as a generator (i.e. statistical model) or as a cost function.

- We replace physical image formation stage with a differentiable renderer to make use of first order derivatives (i.e. gradient descent). Unlike its alternatives, gradient descent provides computationally cheaper and more reliable derivatives through such deep architectures (i.e. above-mentioned texture GAN and identity DCNN).

## 3.3 Approach

We propose an optimization-based 3D face reconstruction approach from a single image that employs a high fidelity texture generation network as statistical prior as illustrated in Fig. 5.1. To this end, the reconstruction mesh is formed by 3D morphable shape model; textured by the generator network's output UV map; and projected into 2D image by a differentiable renderer. The distance between the rendered image and the input image is minimized in terms of a number of cost functions by updating the latent parameters of 3DMM and the texture network with gradient descent. We mainly formulate these functions based on rich features of face recognition network [DGXZ19, SKP15, PVZ15] for

smoother convergence and landmark detection network [DZCZ18] for alignment and rough shape estimation.

The following sections introduce firstly our novel texture model that employs a generator network trained by progressive growing GAN framework. After describing the procedure for image formation with differentiable renderer, we formulate our cost functions and the procedure for fitting our shape and texture models onto a test image.

### 3.3.1 GAN Texture Model

Although conventional PCA is powerful enough to build a decent shape and texture model, it is often unable to capture high frequency details and ends up having blurry textures due to its Gaussian nature. This becomes more apparent in texture modelling which is a key component in 3D reconstruction to preserve identity as well as photo-realism.

GANs are shown to be very effective at capturing such details. However, they suffer from preserving 3D coherency [Goo16] of the target distribution when the training images are semi-aligned. We found that a GAN trained with UV representation of real textures with per pixel alignment avoids this problem and is able to generate realistic and coherent UVs from $99.9\%$ of its latent space while at the same time generalizing well to unseen data.

In order to take advantage of this perfect harmony, we train a progressive growing GAN [KALL18] to model distribution of UV representations of 10,000 high resolution textures and use the trained generator network

$$\mathcal{G}(\mathbf{p}_t) : \mathbb{R}^{512} \rightarrow \mathbb{R}^{H \times W \times C} \tag{3.5}$$

as texture model that replaces 3DMM texture model in Eq. 3.1.

While fitting with linear models, i.e. 3DMM, is as simple as linear transformation, fitting with a generator network can be formulated as an optimization that minimizes per-pixel Manhattan distance between target texture in UV space $\mathbf{I}_{uv}$ and the network output $\mathcal{G}(\mathbf{p}_t)$ with respect to the latent parameter $\mathbf{p}_t$, i.e. $\min_{\mathbf{p}_t} |\mathcal{G}(\mathbf{p}_t) - \mathbf{I}_{uv}|$.

### 3.3.2 Differentiable Renderer

Following [GCM+18], we employ a differentiable renderer to project 3D reconstruction into a 2D image plane based on deferred shading model with given camera and illumination parameters. Since color and normal attributes at each vertex are interpolated at the corresponding pixels with barycentric coordinates, gradients can be easily backpropagated through the renderer to the latent parameters.

A 3D textured mesh at the center of Cartesian origin $[0, 0, 0]$ is projected onto 2D image plane by a pinhole camera model with the camera standing at $[x_c, y_c, z_c]$, directed towards $[x'_c, y'_c, z'_c]$, with world's up direction $[\hat{x}_c, \hat{y}_c, \hat{z}_c]$, and with the focal length $f_c$. The illumination is modelled by phong shading given 1) direct light source at 3D coordinates $[x_l, y_l, z_l]$ with color values $[r_l, g_l, b_l]$, and 2) color of ambient lighting $[r_a, g_a, b_a]$.

Finally, we denote the rendered image given camera ( $\mathbf{p}_c = [x_c, y_c, z_c, x'_c, y'_c, z'_c, \hat{x}_c, \hat{y}_c, \hat{z}_c, f_c]$ ), geometry ($\mathbf{p}_{s,e}$), texture ($\mathbf{p}_t$), and lighting parameters ($\mathbf{p}_l = [x_l, y_l, z_l, r_l, g_l, b_l, r_a, g_a, b_a]$) by the following:

$$\mathbf{I}^{\mathcal{R}} = \mathcal{R}(\mathbf{S}(\mathbf{p}_s, \mathbf{p}_e), \mathcal{P}(\mathcal{G}(\mathbf{p}_t)), \mathbf{p}_c, \mathbf{p}_l) \tag{3.6}$$

where we construct shape mesh by 3DMM as given in Eq. 3.2 and texture by GAN generator network as in Eq. 3.5. Since our differentiable renderer supports only color vectors, we sample from our generated UV map to get vectorized color representation as explained in Sec. 3.2.1.

Additionally, we render a secondary image with random expression, pose and illumination in order to generalize identity related parameters well with those variations. We sample expression parameters from a normal distribution as $\hat{\mathbf{p}}_e \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$ and sample camera and illumination parameters from the Gaussian distribution of 300W-3D dataset as $\hat{\mathbf{p}}_c \sim \mathcal{N}(\hat{\mu}_c, \hat{\sigma}_c)$ and $\hat{\mathbf{p}}_l \sim \mathcal{N}(\hat{\mu}_l, \hat{\sigma}_l)$. This rendered image of the same identity as $\mathbf{I}^{\mathcal{R}}$ (i.e. with same $\mathbf{p}_s$ and $\mathbf{p}_t$ parameters) is expressed by the following:

$$\hat{\mathbf{I}}^{\mathcal{R}} = \mathcal{R}(\mathbf{S}(\mathbf{p}_s, \hat{\mathbf{p}}_e), \mathcal{P}(\mathcal{G}(\mathbf{p}_t)), \hat{\mathbf{p}}_c, \hat{\mathbf{p}}_l) \tag{3.7}$$

### 3.3.3 Cost Functions

Given an input image $\mathbf{I}^0$, we optimize all of the aforementioned parameters simultaneously with gradient descent updates. In each iteration, we simply calculate the forthcoming cost terms for the current state of the 3D reconstruction, and take the derivative of the weighted error with respect to the parameters using backpropagation.

#### Identity Loss

With the availability of large scale datasets, CNNs have shown incredible performance on many face recognition benchmarks. Their strong identity features are robust to many variations including pose, expression, illumination, age etc. These features are shown to be quite effective at many other tasks including novel identity synthesizing (as in Chapter 5), face normalization [CBK+17] and 3D face reconstruction [GCM+18]. In our approach, we take advantage of an off-the-shelf state-of-the-art face recognition network [DGXZ19][4] in order to capture identity related features of an input face image and optimize the latent parameters accordingly. More specifically, given a pretrained face recognition network $\mathcal{F}^n(\mathbf{I}) : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{512}$ consisting of $n$ convolutional filters, we calculate the cosine distance between the identity features (i.e. embeddings) of the real target image and our rendered images as following:

$$\mathcal{L}_{\text{id}} = 1 - \frac{\mathcal{F}^n(\mathbf{I}^0).\mathcal{F}^n(\mathbf{I}^{\mathcal{R}})}{||\mathcal{F}^n(\mathbf{I}^0)||_2||\mathcal{F}^n(\mathbf{I}^{\mathcal{R}})||_2} \tag{3.8}$$

We formulate an additional identity loss on the rendered image $\hat{\mathbf{I}}^{\mathcal{R}}$ that is rendered with random pose, expression and lighting. This loss ensures that our reconstruction resembles the target identity under different conditions. We formulate it by replacing $\mathbf{I}^{\mathcal{R}}$ by $\hat{\mathbf{I}}^{\mathcal{R}}$ in Eq. 3.8 and it is denoted as $\hat{\mathcal{L}}_{\text{id}}$.

#### Content Loss

Face recognition networks are trained to remove all kinds of attributes (e.g. expression, illumination, age, pose) other than abstract identity information throughout the convolutional layers. Despite their

[4]We empirically deduced that other face recognition networks work almost equally well and this choice is orthogonal to the proposed approach.

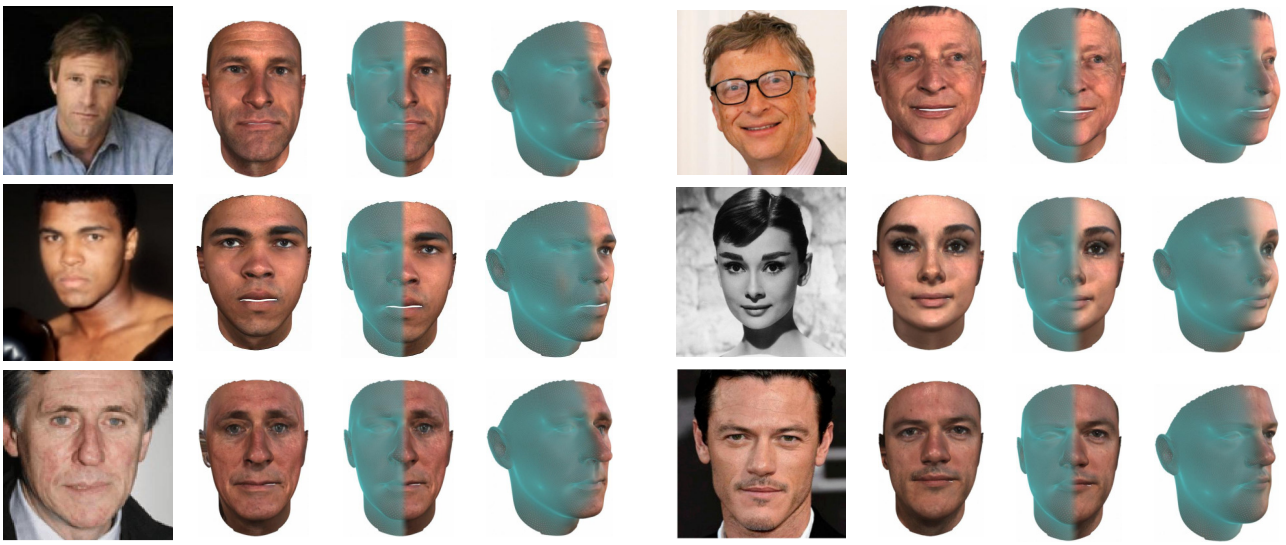


Figure 3.3: Example fits of our approach for the images from various datasets. Please note that our fitting approach is robust to occlusion (e.g. glasses), low resolution and black-white in the photos and generalizes well with ethnicity, gender and age. The reconstructed textures are very well at capturing high frequency details of the identities; likewise, the reconstructed geometries from 3DMM are surprisingly good at identity preservation thanks to the identity features used, e.g. crooked nose at bottom-left, dull eyes at bottom-right and chin dimple at top-left

strength, the activations in the very last layer discard some of the mid-level features that are useful for 3D reconstruction, e.g. variations that depend on age. Therefore we found it effective to accompany identity loss by leveraging intermediate representations in the face recognition network that are still robust to pixel-level deformations and not too abstract to miss some details. To this end, normalized euclidean distance of intermediate activations, namely content loss, is minimized between input and rendered image with the following loss term:

$$\mathcal{L}_{\text{con}} = \sum_{j}^{n} \frac{||\mathcal{F}^j(\mathbf{I}^0) - \mathcal{F}^j(\mathbf{I}^{\mathcal{R}})||_2}{H_{\mathcal{F}^j} \times W_{\mathcal{F}^j} \times C_{\mathcal{F}^j}} \tag{3.9}$$

**Pixel Loss**

While identity and content loss terms optimize albedo of the visible texture, lighting conditions are optimized based on pixel value difference directly. While this cost function is relatively primitive, it is sufficient to optimize lighting parameters such as ambient colors, direction, distance and color of a light source. We found that optimizing illumination parameters jointly with others helped to improve

albedo of the recovered texture. Furthermore, pixel loss support identity and content loss with fine-grained texture as it supports highest available resolution while images needs to be downscaled to $112 \times 112$ before identity and content loss. The pixel loss is defined by pixel level $\ell_1$ loss function as:

$$\mathcal{L}_{\text{pix}} = ||\mathbf{I}^0 - \mathbf{I}^{\mathcal{R}}||_1 \tag{3.10}$$

**Landmark Loss**

The face recognition network $\mathcal{F}$ is pre-trained by the images that are aligned by similarity transformation to a fixed landmark template. To be compatible with the network, we align the input and rendered images under the same settings. However, this process disregards the aspect ratio and scale of the reconstruction. Therefore, we employ a deep face alignment network [DZCZ18] $\mathcal{M}(\mathbf{I}) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{68 \times 2}$ to detect landmark locations of the input image and align the rendered geometry onto it by updating the shape, expression and camera parameters. That is, camera parameters are optimized to align with the pose of image $\mathbf{I}$ and geometry parameters are optimized for the rough shape estimation. As a natural consequence, this alignment drastically improves the effectiveness of the pixel and content loss, which are sensitive to misalignment between the two images.

The alignment error is achieved by point-to-point euclidean distances between detected landmark locations of the input image and 2D projection of the 3D reconstruction landmark locations that is available as meta-data of the shape model. Since landmark locations of the reconstruction heavily depend on camera parameters, this loss is great a source of information the alignment of the reconstruction onto input image and is formulated as following:

$$\mathcal{L}_{\text{lan}} = \sum_{j}^{68} ||\mathcal{M}(\mathbf{I}^0)_j - \mathcal{M}(\mathbf{I}^{\mathcal{R}})_j||_2 \tag{3.11}$$

### 3.3.4 Model Fitting

We first roughly align our reconstruction to the input image by optimizing shape, expression and camera parameters by: $\min_{\mathbf{p}^r} \mathcal{E}(\mathbf{p}^r) = \lambda_{\text{lan}}\mathcal{L}_{\text{lan}} + \lambda_{\text{reg}}\text{Reg}(\{\mathbf{p}_{s,e}\})$. We then simultaneously optimize all of

our parameters with gradient descent and backpropagation so as to minimize weighted combination of above loss terms in the following:

$$\begin{aligned}\min_{\mathbf{p}} \mathcal{E}(\mathbf{p}) = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \hat{\lambda}_{id}\hat{\mathcal{L}}_{\text{id}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} \\ + \lambda_{\text{lan}}\mathcal{L}_{\text{lan}} + \lambda_{\text{reg}}\text{Reg}(\{\mathbf{p}_{s,e}, \mathbf{p}_l\})\end{aligned} \tag{3.12}$$

where we weight each of our loss terms with $\lambda$ parameters. In order to prevent our shape and expression models and lighting parameters from exaggeration to arbitrarily bias our loss terms, we regularize those parameters by $\text{Reg}(\{\mathbf{p}_{s,e}, \mathbf{p}_l\})$.

## 3.4 Drawbacks and Improvements on GANFit

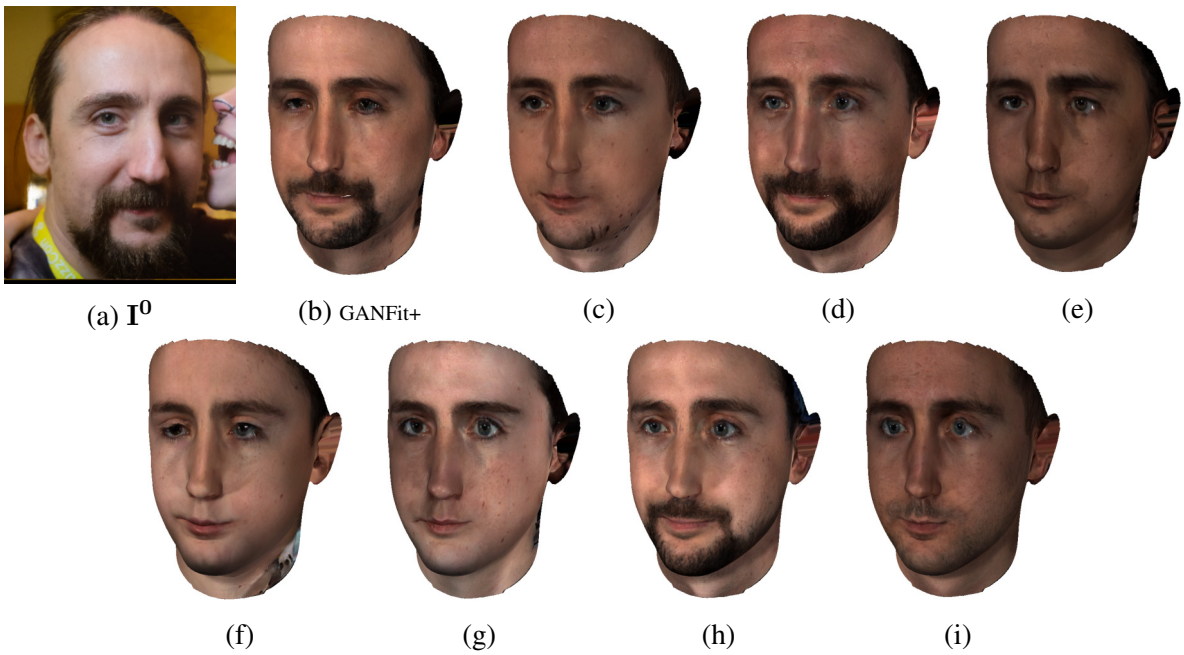
### 3.4.1 Fitting with Multiple Images (i.e. Video):

Although the proposed approach can fit a 3D reconstruction from a single image, one can take advantage of more images effectively when available, e.g. from a video recording. This often helps to improve reconstruction quality under challenging conditions, e.g. outdoor, low resolution. While state-of-the-art methods follow naive approaches by averaging either the reconstruction [THMM17] or features-to-be-regressed [GCM+18] before making a reconstruction, we utilize the power of iterative optimization by averaging identity reconstruction parameters $(\mathbf{p}_s, \mathbf{p}_t)$ after every iteration. For an image set $\mathbf{I} = \{\mathbf{I}^0, \mathbf{I}^1, \ldots, \mathbf{I}^i, \ldots, \mathbf{I}^{n_i}\}$, we reformulate our parameters as $\mathbf{p} = [\mathbf{p}_s, \mathbf{p}_e^i, \mathbf{p}_t, \mathbf{p}_c^i, \mathbf{p}_l^i]$ in which we average shape and texture parameters by the following:

$$\mathbf{p}_s = \sum_i^n \mathbf{p}_s^i, \mathbf{p}_t = \sum_i^n \mathbf{p}_t^i \tag{3.13}$$

### 3.4.2 Regressing parameters by Encoder

One weakness of the proposed GANFit approach is its sensitivity to the initialization of the generated texture as shown in Fig. 3.4. Due to the non-linear generator network and first order derivatives,



(a) $\mathbf{I}^0$ (b) GANFit+ (c) (d) (e)

(f) (g) (h) (i)

Figure 3.4: The fitting results of the proposed approach for the input image (a) initialized by FastGANFit (b) or randomly resulting reconstructions given random initial textures (c-i)

optimization is not guaranteed to find the global minima and sometimes result in sub-optimal reconstructions. In order to initialize GANFit optimization parameters closer to global/good minima, we propose to train an encoder network to by the same image formation and loss functions that regress latent parameters (i.e. $p_s, p_t, p_e, p_c, p_l$) from the input image.

In order to train this network $N$, we modify GANFit optimization by simply replacing trainable latent parameters $\mathbf{p}$ by the activations of the encoder network ($\mathbf{p}^N = N(\mathbf{I}^0)$) as shown in Fig. 3.5. The architecture of the regression network particularly benefit from different levels of identity features (i.e. content features) of a pretrained face recognition network [DGXZ19] by passing and concatenating same-resolution activations in the regressor network. Moreover, we flatten and concatenate normalized landmark locations by a pretrained landmark detection network [DZCZ18] before the final fully connected layers. This design leverage the information of the state of the art facial features and, unlike its alternatives [GCM+18], it allows the regressor network to benefit all levels of features (i.e. [GCM+18] restrict the features from face recognition network to its final layer which is fully abstract and ignore state-relevant features, e.g. age, facial hair etc.). We call this regression network *FastGANFit* in the rest of the chapter.

The activations of FastGANFit ($\mathbf{p}^N = N(\mathbf{I}^0)$) is connected with GANFit's image formation and loss

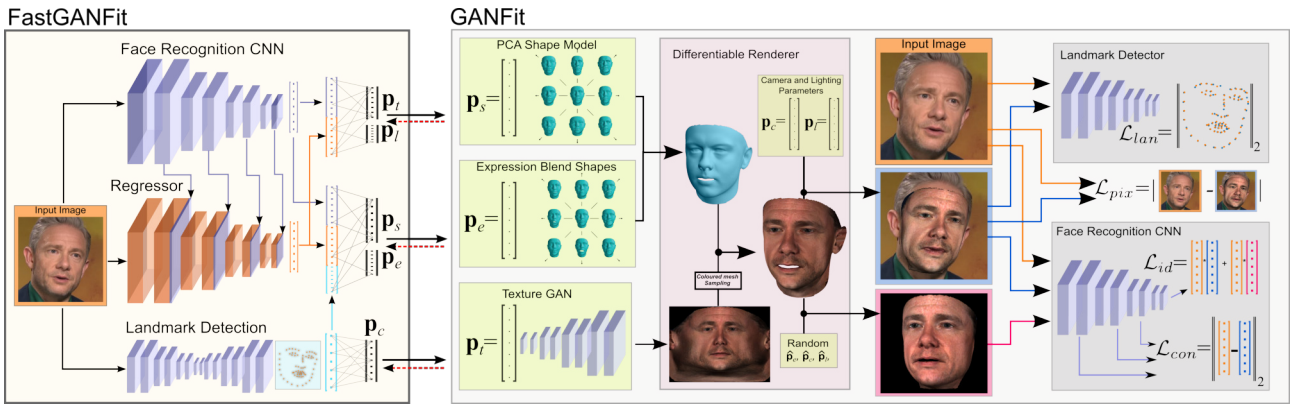



Figure 3.5: Overview of the approach with regression network. The network is end-to-end connected with the differentiable renderer and the lost functions of GANFit. It benefits from the activations of all layers of a pretrained face recognition network and detection of a hourglass landmark detector. The network is trained similar to GANFit optimization: 1) alignment 2) full objective. The only difference is that now the regression network is being optimized (i.e. trained) instead of the trainable latent parameters of GANFit.

functions in an end-to-end manner. The network is then pre-trained for camera, shape and expression parameters by optimizing only $\min_{\mathbf{p}^{rN}} \mathcal{E}(\mathbf{p}^{rN}) = \lambda_{\text{lan}}\mathcal{L}_{\text{lan}} + \lambda_{\text{reg}}\text{Reg}(\{\mathbf{p}_{s,e}\}$. After having a good alignment of the reconstructions of a given training data, we train the regressor network again with our full objective function as in the following:

$$\begin{aligned}\min_{\mathbf{p}^N} \mathcal{E}(\mathbf{p}^N) = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \hat{\lambda}_{id}\hat{\mathcal{L}}_{\text{id}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} \\ + \lambda_{\text{lan}}\mathcal{L}_{\text{lan}} + \lambda_{\text{reg}}\text{Reg}(\{\mathbf{p}_{s,e}, \mathbf{p}_l\})\end{aligned} \tag{3.14}$$

The resulting network can either regress 3D reconstruction parameters directly (called FastGANFit) or initialize GANFit's latent parameters $\mathbf{p}$ by the regressed parameters $N(\mathbf{I}^0)$ before running GANFit optimization as explained in Section 5.3. We call this initialization trick GANFit+ in the rest of the chapter. This combination of optimization- and regression-based 3D reconstruction leverage best of both worlds: stability of inference and high-fidelity of iterative optimization. Furthermore, one can also enjoy flexibility of trade-off between speed and quality depending on the application.

## 3.5 Experiments

This section demonstrates the excellent performance of the proposed approach for 3D face reconstruction and shape recovery. We verify this by qualitative results in Figures 3.1, 3.3, qualitative comparisons with the state-of-the-art in Sec. 3.5.2 and quantitative shape reconstruction experiment on a database with ground truth in Sec. 3.5.3. In our experiments, we evaluate the proposed approach under three different settings: 1) GANFit, i.e. Fitting with random initialization, 2) Fast-GANFit, i.e. Regressing the parameters by the regressor network, 3) GANFit++, i.e. Fitting after initializing with Fast-GANFit.

### 3.5.1 Implementation Details

For all of our experiments, a given face image is aligned to our fixed template using 68 landmark locations detected by an hourglass 2D landmark detection [DZCZ18]. For the identity features, we employ ArcFace [DGXZ19] network's pretrained models. For the generator network $\mathcal{G}$, we train a progressive growing GAN [KALL18] with around 10,000 UV maps from [BRZ+16] at the resolution of $512 \times 512$. We use the Large Scale Face Model [BRZ+16] for 3DMM shape model with $n_s = 158$ and the expression model learned from 4DFAB database [CKPZ18] with $n_e = 29$. During fitting process, we optimize parameters using Adam Solver [KB15] with 0.01 learning rate. And we set our balancing factors as the following: $\lambda_{\text{id}} : 2.0, \hat{\lambda}_{id} : 2.0, \lambda_{\text{con}} : 50.0, \lambda_{\text{pix}} : 1.0, \lambda_{\text{lan}} : 0.001, \lambda_{\text{reg}} : \{0.05, 0.01\}$. The Fitting converges in around 30 seconds on an Nvidia GTX 1080 TI GPU for a single image.

### 3.5.2 Qualitative Results

#### Comparison on MOFA test set

Fig. 5.5 compares our results with the most recent face reconstruction studies [TZK+17, TZG+18, GCM+18, THMM17, TL18] on a subset of MoFA test-set. The first four rows after input images

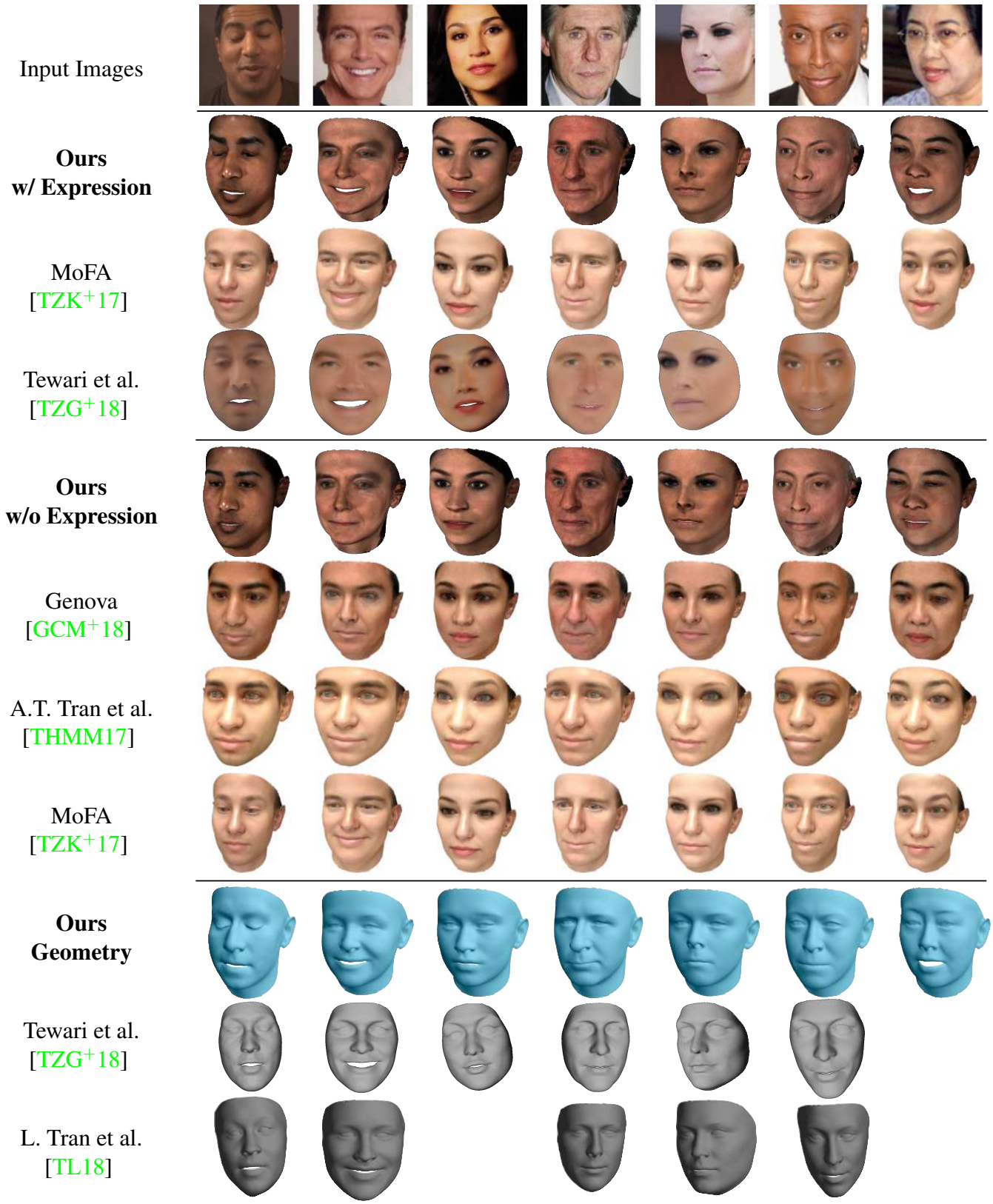



Figure 3.6: Comparison of our qualitative results with other state-of-the-art methods in MoFA-Test dataset. Rows 2-4 show comparison of textured geometry with original expression, rows 5-8 show comparison of textured geometry with neutral expression and rows 9-11 compare only shapes.

show a comparison of our shape and texture reconstructions to [GCM+18, THMM17, TZG+18] and the last three rows show our reconstructed geometries without texture compared to [TZG+18, TL18]. All in all, our method outshines all others with its high fidelity photorealistic texture reconstructions. Both of our texture and shape reconstructions manifest strong identity characteristics of the corresponding input images from the thickness and shape of the eyebrows to wrinkles around the mouth and forehead.

Rows 2 and 5 show results of our method with and without expression fitting (i.e. $\mathbf{p}_e$) visualized. It is visible that our method even captured that eyelids are closed in the geometry (i.e. first column) and that the expression is well disentangled from shape compared to MoFA [TZK+17].

**Comparison on high-quality texture generation**

In order to support our claim to generate high-quality textures, in Fig. 3.8 we provide comparative results with [SWH+17] which renders 3D reconstruction using commercial renderer tweaks. Our method provides excellent textures even without any improvements by renderers (which is orthogonal our method). This is more visible when the reconstruction is overlaid on the input images as in Fig. 3.7.

One general drawback of our texture model is that it sometimes contains highlights and shadow patterns, i.e. around the nose. We believe that these flaws are inherited from the training data that is collected in a controlled environment with multiple point light sources. The intensities of the shadows and the highlights are affected by the intensity of the daylight at the time of collection, thus, these artifacts may not be always visible.

**Results under challenging conditions**

Fig. 3.9 illustrates results of GANFit under more challenging conditions such as strong illuminations, self-occlusions and facial hair. Please note that our method succesfully reconstruct these challenging face images in 3D thanks to strong face recognition features. Particularly, Fig. 3.9(d) shows that the direction and the intensity of the illumination in the scene are successfully estimated by our lighting

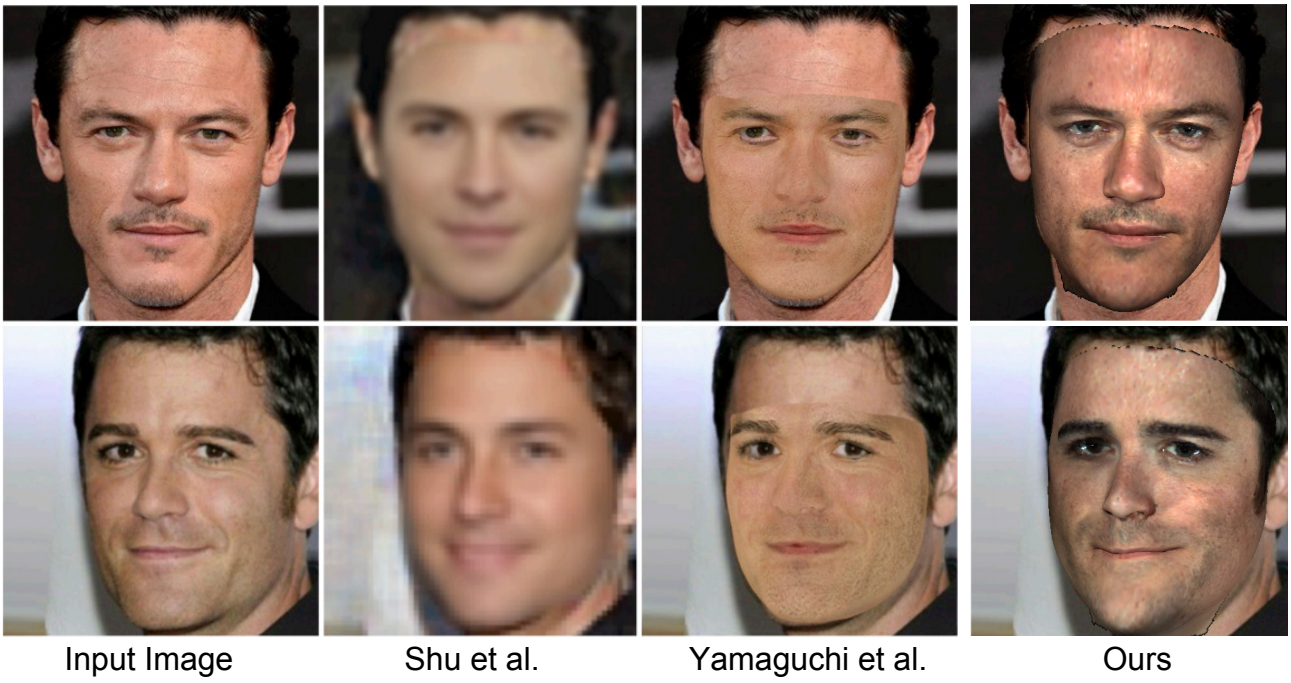



Figure 3.7: Qualitative comparison with [YSN+18, SYH+17] by overlaying the reconstructions on the input images. Our method can generate high fidelity texture with accurate shape, camera and illumination fitting.

model which consists of one RGB point light source and RGB ambient lighting. This result indicates that our combined loss function can effectively optimize the position of the point light source and its colour intensities.

Fig. 3.10 shows 3D reconstruction results given paintings from BAM dataset [WFJ+17].

### 3.5.3 Quantitative Experiments

#### 3D shape recovery on MICC dataset

We evaluate the shape reconstruction performance of our method on MICC Florence 3D Faces dataset (MICC) [BDBM11] in Table 3.1. The dataset provides 3D scans of 53 subjects as well as their short video footages under three difficulty settings: 'cooperative', 'indoor' and 'outdoor'. Unlike [GCM+18, THMM17] which processes all the frames in a video, we uniformly sample only 5 frames from each video regardless of their zoom level. And, we run our method with multi-image support for these 5 frames for each video separately as shown in Eq. 3.13. Each test mesh is cropped at a

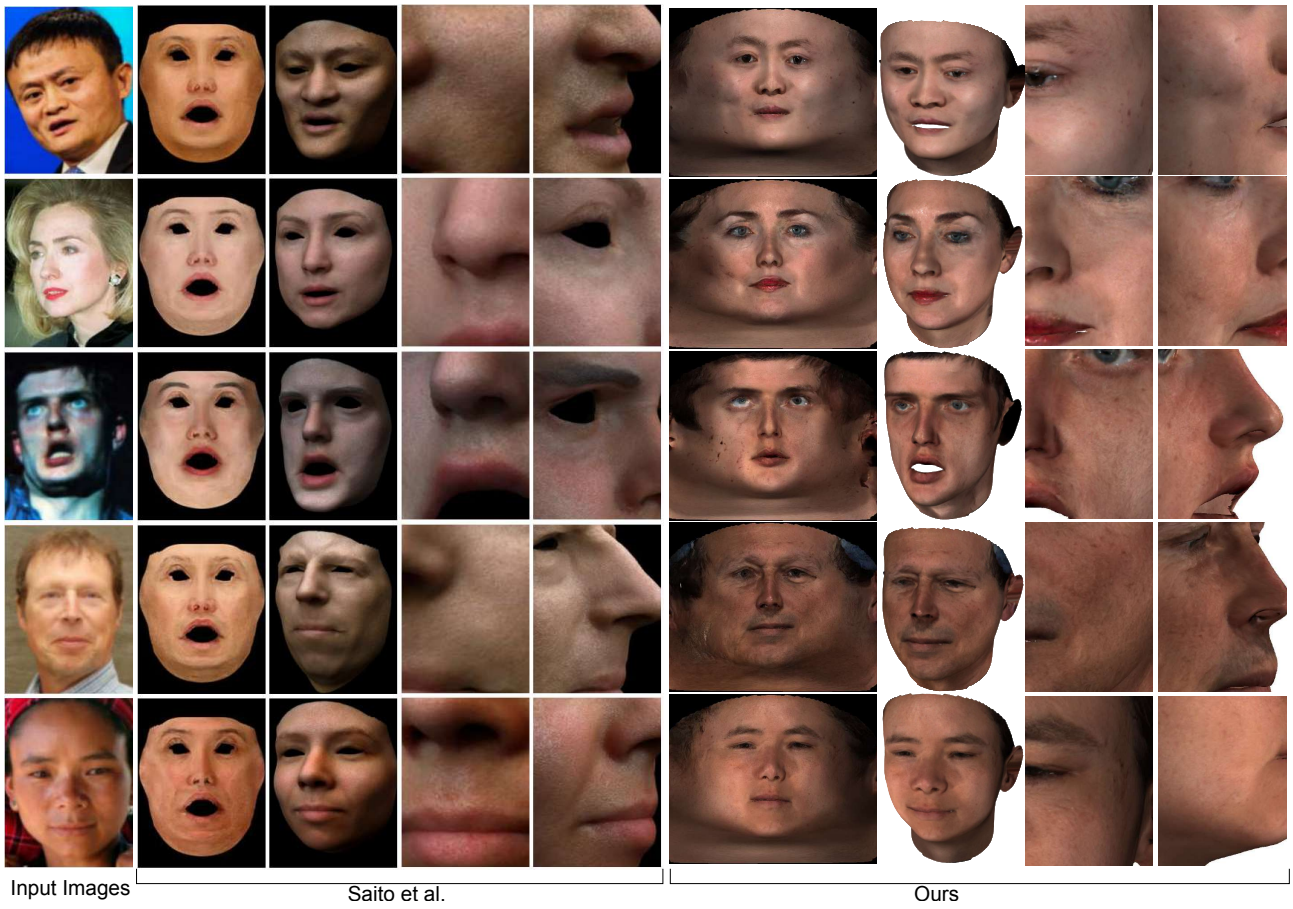
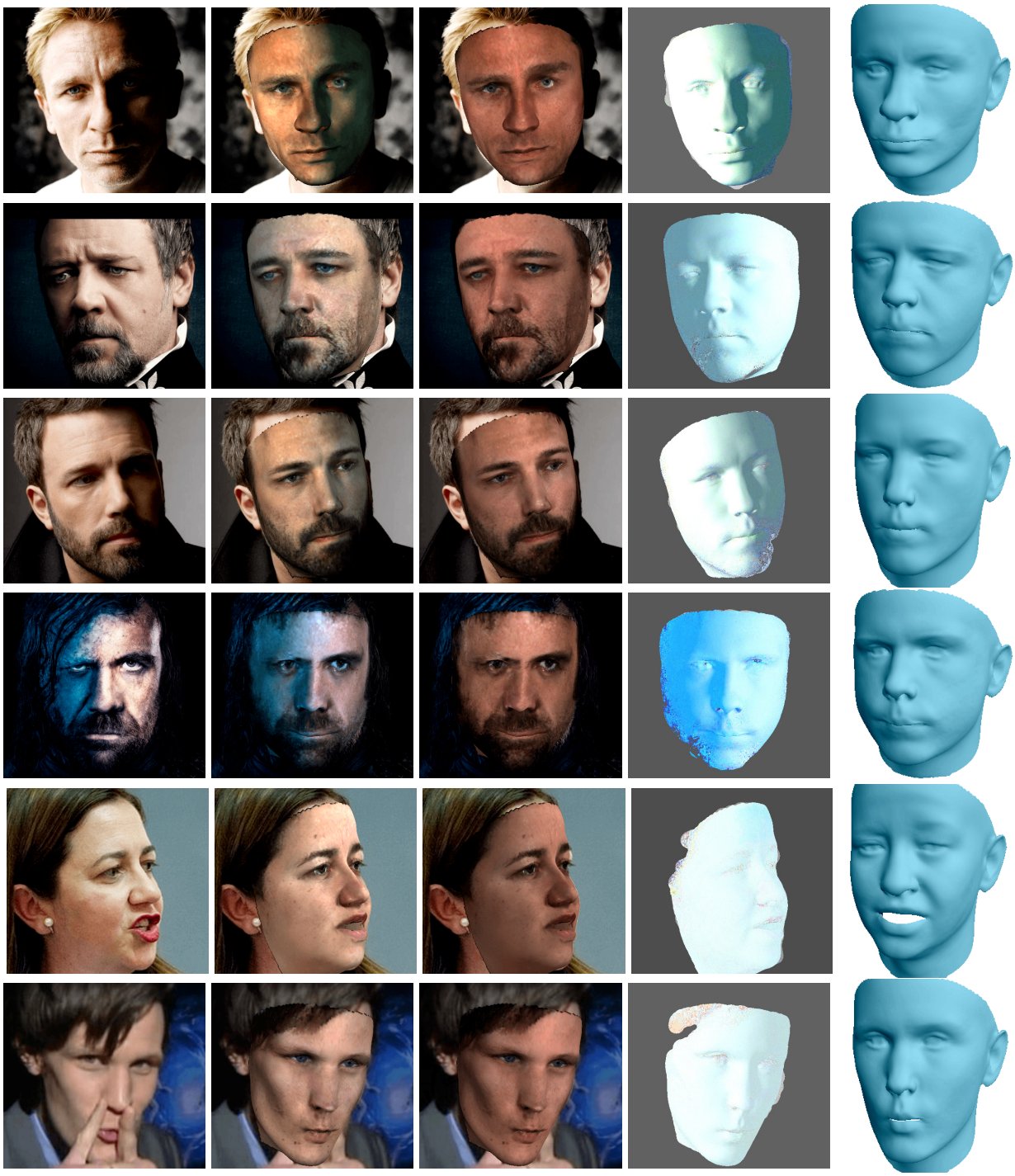



Figure 3.8: Qualitative comparison with [SWH+17] by means of texture maps, whole and partial face renderings. Please note that while our method does not require any particular renderer for special effects, e.g. lighting, [SWH+17] produce these renderings with a commercial renderer called Arnold.



(a) $\mathbf{I^0}$ (b) $\mathbf{I}^{\mathcal{R}}$ (c) $\mathbf{I}^{\mathcal{R}}_{\mathbf{alb.}}$ (d) $\mathbf{I}^{\mathcal{R}}/\mathbf{I}^{\mathcal{R}}_{\mathbf{alb.}}$ (e) $\mathbf{S}$

Figure 3.9: Results under more challenging conditions, i.e. strong illuminations, self-occlusions and facial hair. (a) Input image. (b) Estimated fitting overlayyed including illumination estimation. (c) Overlayyed fitting without illumination. (d) Normalized pixel-wise intensity ratio of (b) to (c). (e) Estimated shape mesh

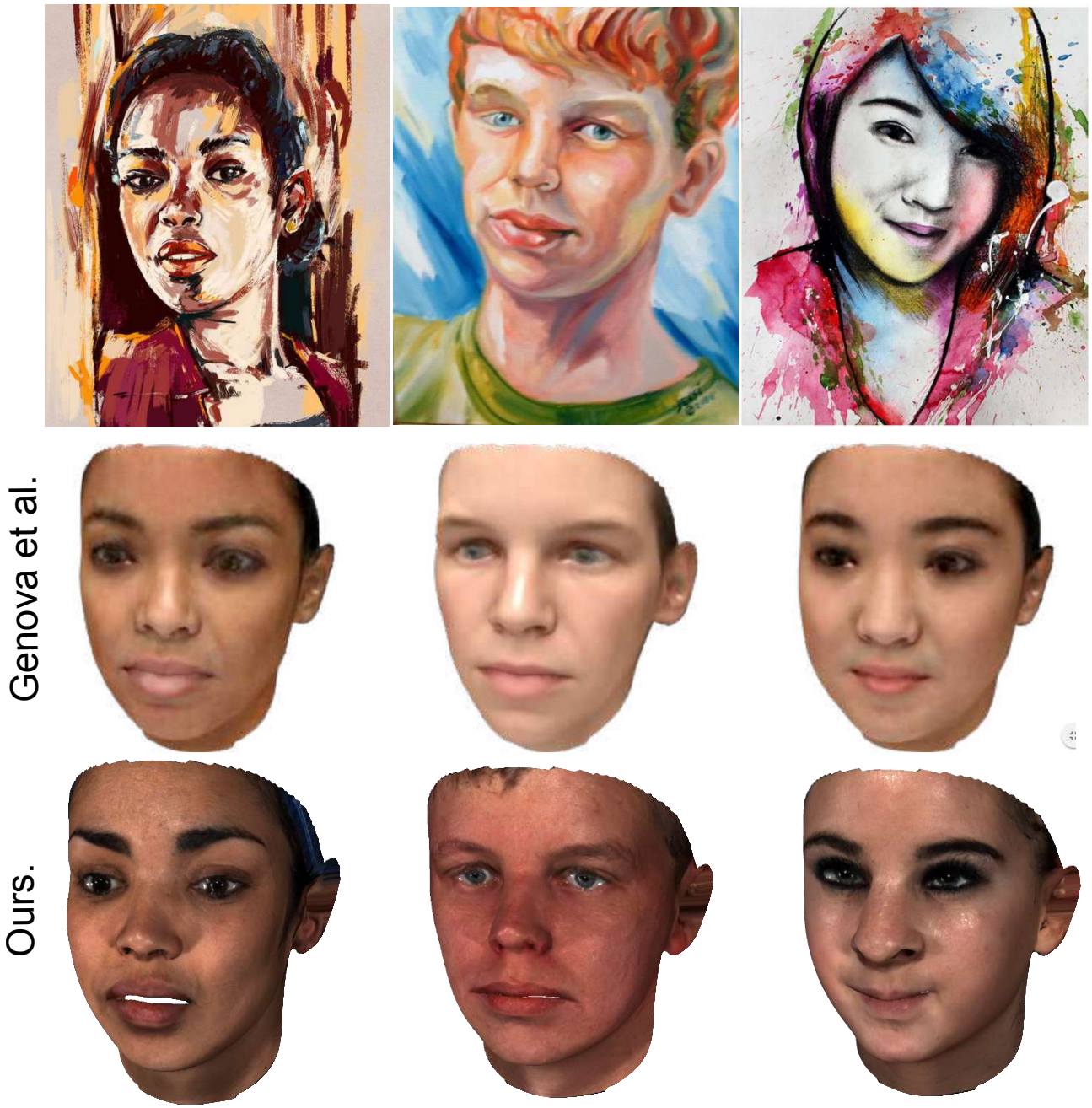



Figure 3.10: Our results on BAM dataset[WFJ+17] compared to [GCM+18]. Our method is robust to many image deformations and even capable of recovering identities from paintings thanks to strong identity features.

| *Method* | *Cooperative* Mean | Std. | *Indoor* Mean | Std. | *Outdoor* Mean | Std. |
|---|---|---|---|---|---|---|
| Tran et al. [THMM17] | 1.93 | 0.27 | 2.02 | 0.25 | 1.86 | 0.23 |
| Booth et al. [BAP+17] | 1.82 | 0.29 | 1.85 | 0.22 | 1.63 | 0.16 |
| Genova et al. [GCM+18] | 1.50 | 0.13 | 1.50 | 0.11 | 1.48 | 0.11 |
| GANFit (*Ours*) | 0.95 | 0.107 | 0.94 | 0.106 | 0.94 | 0.106 |
| Fast-GANFit (*Ours*) | 0.94 | 0.24 | 0.94 | 0.173 | 0.93 | **0.095** |
| GANFit++ (*Ours*) | **0.93** | **0.091** | **0.92** | **0.101** | **0.92** | 0.099 |

Table 3.1: Accuracy results for the meshes on the MICC Dataset using point-to-plane distance. The table reports the mean error (Mean), the standard deviation (Std.).

radius of $95mm$ around the tip of the nose according to [THMM17] in order to evaluate the shape recovery of the inner facial mesh. We perform dense alignment between each predicted mesh and its corresponding ground truth mesh, by implementing an iterative closest point (ICP) method [BM92]. As evaluation metric, we follow [GCM+18] to measure the error by average symetric point-to-plane distance.

Table 3.1 reports the normalized point-to-plain errors in millimeters. It is evident that we have improved the absolute error compared to the other two state-of-the-art methods by $36\%$. Our results are shown to be consistent across all different settings with minimal standard deviation from the mean error.

**Experiments on LFW**

In order to evaluate identity preservation capacity of the proposed method, we run two face recognition experiments on Labelled Faces in the Wild (LFW) dataset [HMLL12]. Following [GCM+18], we feed real LFW images and rendered images of their 3D reconstruction by our method to a pretrained face recognition network, namely VGG-Face[PVZ15]. We then compute the activations at the embedding layer and measure cosine similarity between 1) real and rendered images and 2) renderings of same/different pairs.

In Fig. 3.11 and 3.12, we have quantitatively showed that our method is better at identity preservation and photorealism (i.e. as the pretrained network is trained by real images) than other state-of-the-art deep 3D face reconstruction approaches [GCM+18, THMM17].

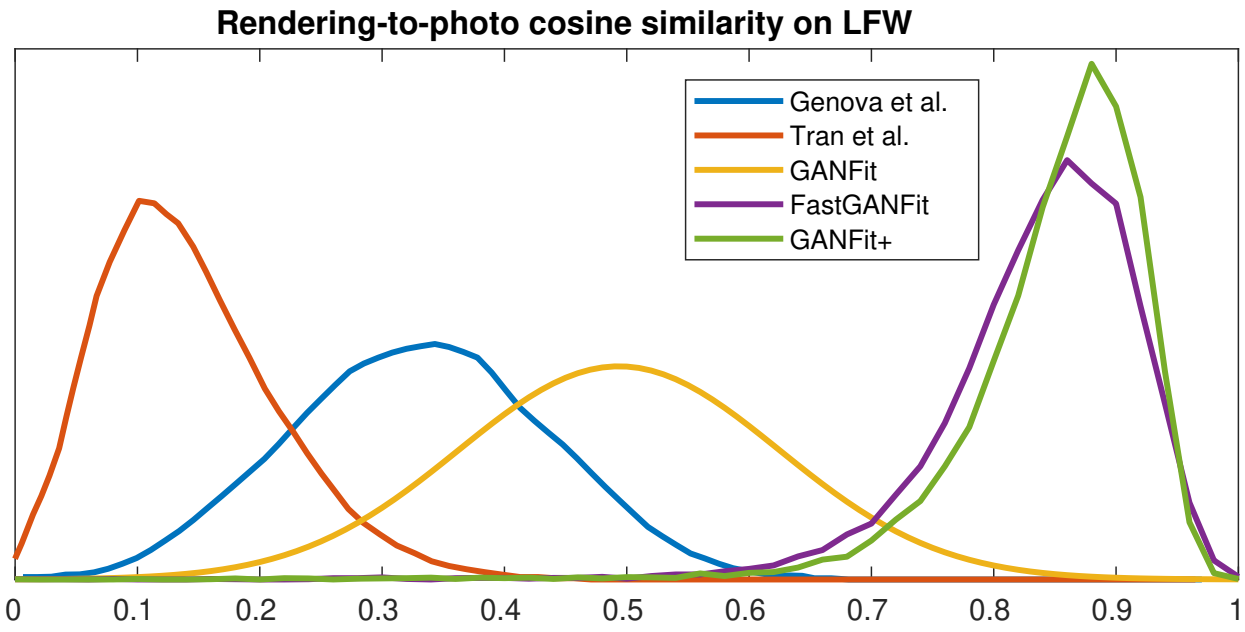



Figure 3.11: Cosine similarity distributions of rendered and real images LFW based on activations at the embedding layer of VGG-Face network[PVZ15]. Our method achieves more than 0.5 similarity on average which [GCM+18] has 0.35 average similarity and [THMM17] 0.16 average similarity. Camera and lighting parameters are fixed for all renderings.

### 3.5.4 Ablation Study

Fig. 3.13 shows an ablation study on our method where the full model reconstructs the input face better than its variants, something that suggests that each of our components significantly contributes towards a good reconstruction. Fig. 3.13(c) indicates albedo is well disentangled from illumination and our model capture the light direction accurately.

While Fig. 3.13(d-f) shows each of the identity terms contributes to preserve identity, Fig. 3.13(h) demonstrates the significance identity features altogether. Still, overall reconstruction utilizes pixel intensities to capture better albedo and illumination as shown in Fig. 3.13(g). Finally, Fig. 3.13(i) shows the superiority of our textures over PCA-based ones.

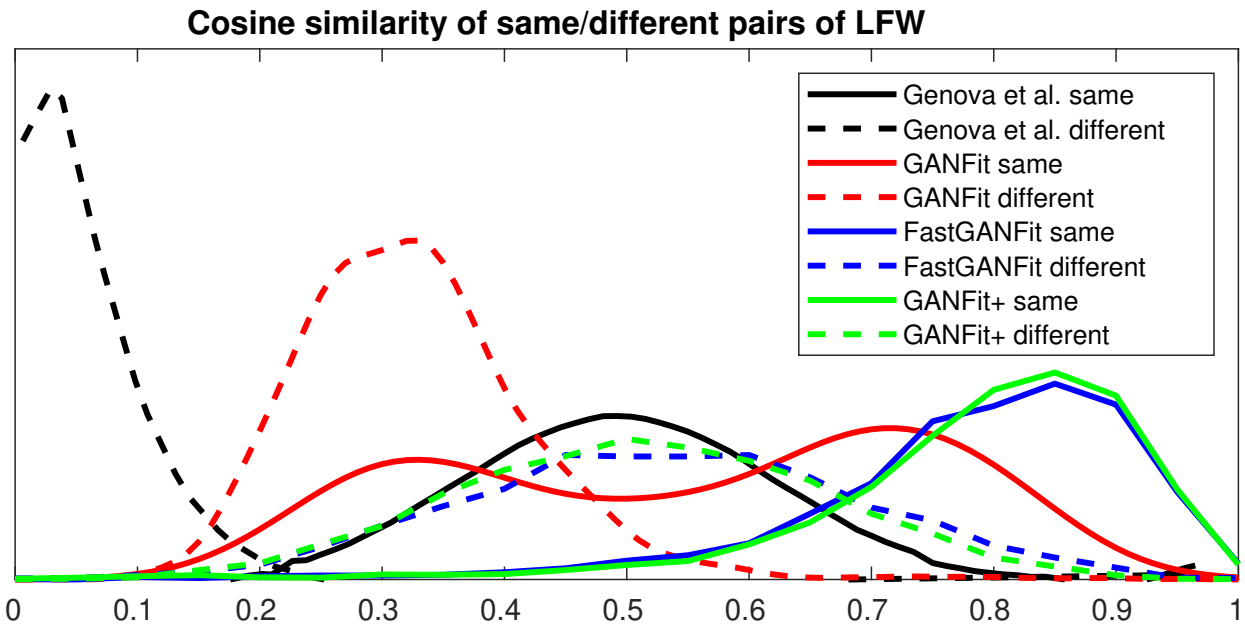



Figure 3.12: Our method successfully preserve identity so that distribution of cosine similarity of same/different pairs is separable by thresholding. Camera and lighting parameters are fixed for all renderings.

## 3.6 Conclusion

In this chapter, we revisit optimization-based 3D face reconstruction under a new perspective, that is, we utilize the power of recent machine learning techniques such as GANs and face recognition network as statistical texture model and as energy function respectively.

To the best of our knowledge, this is the first time that GANs are used for model fitting and they have shown excellent results for high quality texture reconstruction. The proposed approach shows identity preserving high fidelity 3D reconstructions in qualitative and quantitative experiments.

The following chapters focus on improving non-linear 3D face modeling in other domains such as shape, normals and expression, as well as 2D novel face synthesization.



(a) $\mathbf{I^0}$ (b) $\mathbf{I}^{\mathcal{R}}$ (c) $\mathbf{I}^{\mathcal{R}}$ albedo

(d) $\mathbf{I}^{\mathcal{R}} \setminus \mathcal{L}_{\text{id}}$ (e) $\mathbf{I}^{\mathcal{R}} \setminus \hat{\mathcal{L}}_{\text{id}}$ (f) $\mathbf{I}^{\mathcal{R}} \setminus \mathcal{L}_{\text{con}}$

(g) $\mathbf{I}^{\mathcal{R}} \setminus \mathcal{L}_{\text{pix}}$ (h) $\mathbf{I}^{\mathcal{R}} \setminus \{\mathcal{L}_{\text{id}}, \hat{\mathcal{L}}_{\text{id}}, \mathcal{L}_{\text{con}}\}$ (i) $\mathbf{I}^{\mathcal{R}}$ with $\mathbf{T}(\mathbf{p_t})$

Figure 3.13: Contributions of the components or loss terms of the proposed approach with an leave-one-out ablation study.

# Chapter 4

# Synthesizing Coupled 3D Face Modalities by Trunk-Branch GAN


**4.1 Introduction** . . . . . **83**

**4.2 Related Work** . . . . . **86**

4.2.1 3D face modeling . . . . . 86

4.2.2 Photorealistic face synthesis . . . . . 87

4.2.3 Boosting face recognition by synthetic training data . . . . . 88

**4.3 Approach** . . . . . **89**

4.3.1 UV Maps for Shape, Texture and Normals . . . . . 89

4.3.2 Trunk-Branch GAN to Generate Coupled Texture, Shape and Normals . . . 90

4.3.3 Expression Augmentation by Conditional GAN . . . . . 91

4.3.4 Photorealistic Rendering with Generated UV maps . . . . . 92

**4.4 Results** . . . . . **93**

4.4.1 Qualitative Results . . . . . 94

4.4.2 Pose-invariant Face Recognition . . . . . 98

**4.5 Conclusion** . . . . . **101**

Generating realistic 3D faces is of high importance for computer graphics and computer vision applications. Generally, research on 3D face generation revolves around linear statistical models of the facial surface. Nevertheless, these models cannot represent faithfully either the facial texture or the normals of the face, which are very crucial for photo-realistic face synthesis. Chapter 3 demonstrated that Generative Adversarial Networks (GANs) can be used for modeling and generating high-quality textures of faces. Nevertheless, the generation process either omits the geometry and normals, or independent processes are used to produce 3D shape information. In this chapter, we present the first methodology that generates high-quality texture, shape, and normals jointly, which can be used for photo-realistic synthesis. To do so, we propose a novel GAN that can generate data from different modalities while exploiting their correlations. Furthermore, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions.

## 4.1 Introduction

Generating 3D faces with high-quality texture, shape, and normals is of paramount importance in computer graphics, movie post-production, computer games, etc. Other applications of such approaches include generating synthetic training data for face recognition (covered in Chapter 5) and modeling the face manifold for 3D face reconstruction (covered in Chapter 3). Currently, 3D face generation in computer games and movies is performed by expensive capturing systems or by professional technical artists. The current state-of-the-art methods generate faces, which can be suitable for applications such as caricature avatar creation in mobile devices [HSW+17] but do not generate high-quality shape and normals that can be used for photo-realistic face synthesis. In this chapter, we propose the first methodology for high-quality face generation that can be used for photo-realistic face synthesis (i.e. joint generation of texture, shape, and normals) by capitalizing on the recent developments on Generative Adversarial Networks (GANs).

The early face models, such as [BV99], represent 3D face by disentangled PCA models of geometry, expression [CWZ+14], and colored texture, called 3D morphable models (3DMM). 3DMMs and its

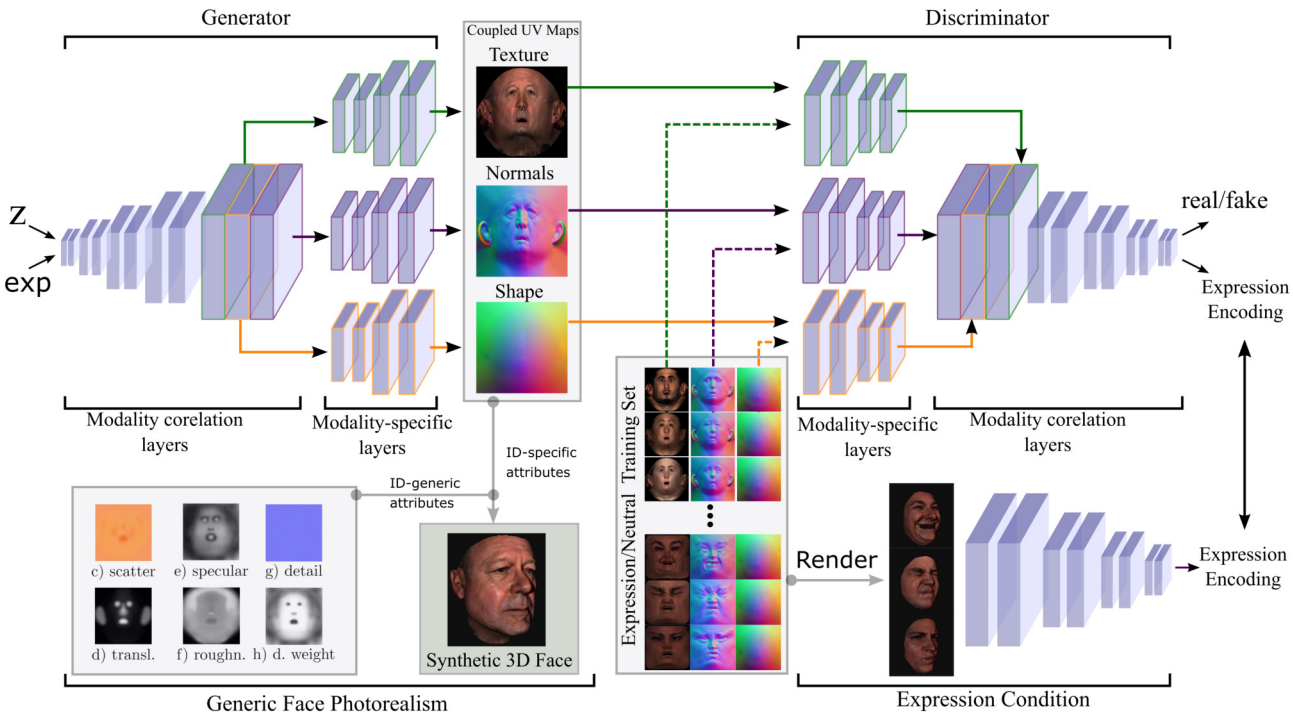



Figure 4.1: We propose a novel GAN that can synthesize high-quality texture, shape, and normals jointly for realistic and coherent 3D faces of novel identities. The separation of branch networks allows the specialization of the characteristic of each one of the modalities while the trunk network maintains the local correspondences among them. Moreover, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions. We annotate the training dataset automatically by an expression recognition network to couple those expression encodings to the texture, shape, and normals UV maps.

variants were the most popular method for modeling shape and texture separately. However, the linear nature of PCA is often unable to capture high-frequency signals properly, thus the quality of generation and reconstruction by PCA is sub-optimal.

GANs is a recently introduced family of techniques that train samplers of high-dimensional distributions [GPM+14]. It has been demonstrated that when a GAN is trained on facial images, it can generate images that have realistic characteristics. In particular, the recently introduced GANs [KALL18, KLA19, BDS19] can generate photo-realistic high-resolution faces. Nevertheless, because they are trained on partially-aligned 2D images, they cannot properly model the manifold of faces and thus (a) inevitably create many unrealistic instances and (b) it is not clear how they can be used to generate photo-realistic 3D faces.

Recently, GANs have been applied for generating facial texture for various applications. In particular, [SRK17] and our method in Chapter 5 utilize style transfer GANs to generate photorealistic

images of 3DMM-sampled novel identities. [SSK19b] directly generates high-quality 3D facial textures by GANs and our method presented in Chapter 3 replaces 3D Morphable Models (3DMMs) with GAN models for 3D texture reconstruction while the shape is still maintained by statistical models. [LMG+20] propose to generate 4K diffuse and specular albedo and normals from a texture map by an image-to-image GAN. On the other hand, [MPN+19] model 3D shape by GANs in a parametric UV map and [RBSB18] utilize mesh convolutions with variational autoencoders to model shape in its original structure. Although one can model 3D faces with such shape and texture GAN approaches, these studies omit the correlation between shape, normals, and texture which is very important for photorealism in identity space. The significance of such correlation is most visible with inconsistent facial attributes such as age, gender, and ethnicity (i.e. old-aged texture on a baby-face geometry).

In order to address these gaps, we propose a novel multi-branch GAN architecture that preserves the correlation between different 3D modalities (such as texture, shape, normals, and expression). After converting all modalities into UV space and concatenate over channels, we train a GAN that generates all modalities in a meaningful local and global correspondence. In order to prevent incompatibility issues due to the intensity distribution of different modalities, we propose a trunk-branch architecture that can synthesize photorealistic 3D faces with coupled texture and geometry. Further, we condition this GAN by expression labels to generate faces in any desired expression.

From a computer graphics point of view, a photorealistic face rendering requires a number of elements to be tailored, i.e. shape, normals and albedo maps, some of which should or can be specific to a particular identity. However, the cost of hand-crafting novel identities limits their usage on large-scale applications. The proposed approach tackles this down with reasonable photorealism with a massively generalized identity space. Although the results in this chapter are limited to aforementioned modalities by the dataset at hand, the proposed method allows adding more identity-specific modalities (i.e. cavity, gloss, scatter) once such a dataset becomes available.

The contributions of this chapter can be summarized as follows:

- We propose to model and synthesize coherent 3D faces by jointly training a novel Trunk-branch based GAN (TBGAN) architecture for shape, texture, and normals modalities. TBGAN is

designed to maintain correlation while tolerating domain-specific differences of these three modalities and can be easily extended to other modalities and domains.

- In the domain of identity-generic face modeling, we believe this is the first study that utilizes normals as an additional source of information.
- We propose the first methodology for face generation that correlates expression and identity geometries (i.e. modeling personalized expression) and also the first attempt to model expression in texture and normals space.

## 4.2 Related Work

### 4.2.1 3D face modeling

There is an underlying assumption that human faces lie on a manifold with respect to the appearance and geometry. As a result, one can model the geometry and appearance of the human face analytically based upon the identity and expression space of all individuals. Two of the first attempts in the history of face modeling were [ASW93], which proposes part-based 3D face reconstruction from frontal and profile images, and [PB81], which represents expression action units by a set of muscle fibers.

Twenty years ago methods that generated 3D faces revolved around parametric generative models that are driven by a small number of anthropometric statistics (e.g. sparse face measurements in a population) which act as constraints [DMS98]. The seminal work of 3D morphable models (3DMMs) [BV99] demonstrated for the first time that is possible to learn a linear statistical model from a population of 3D faces [PS09, BSBW14]. 3DMMs are often constructed by using a Principal Component Analysis (PCA) based on a dataset of registered 3D scans of hundreds [PKA+09] or thousands [BRP+18] subjects. Similarly, facial expressions are also modeled by applying PCA [YMW+11, LBB+17, BBC11, AKV08], or are manually defined using linear blendshapes [LWP10, TZN+15, BWP13]. 3DMMs, despite their advantages, are bounded by the capacity of linear space that underrepresents the high-frequency information and often result in overly-smoothed geometry and texture models. [CCZ+19] and [THM+18] attempt to address this issue by using local displacement maps.

Furthermore, the 3DMM line of research assumes that texture and shape are uncorrelated, hence they can only be produced by separate models (i.e. separate PCA models for texture and shape). Early attempts in correlated shape and texture have been made in Active Appearance Models (AAMs) by computing joint PCA models of sparse shape and texture [CET98]. Nevertheless, due to the inherent limitations of PCA to model high-frequency texture, it is rarely used to correlate shape and texture for 3D face generation.

Recent progress in generative models [KW14, GPM+14] is being utilized in 3D face modeling to tackle this issue. [MPN+19] trained a GAN that models face geometry based on UV representations for neutral faces, and likewise, [RBSB18] modeled identity and expression geometry by variational autoencoders with mesh convolutions. In Chapter 3, we presented a GAN-based texture modeling for 3D face reconstruction while modeling geometry by PCA and [SSK19b] trained a GAN to synthesize facial textures. To the best of our knowledge, these methodologies totally omit the correlation between geometry and texture and moreover, they ignore identity-specific expression modeling by decoupling them into separate models. In order to address this issue, we propose a trunk-branch GAN that is trained jointly for texture, shape, normals, and expression in order to leverage non-linear generative networks for capturing the correlation between these modalities.

### 4.2.2 Photorealistic face synthesis

Although most of the aforementioned 3D face models can synthesize 2D face images, there are also some dedicated 2D face generation studies. [MPK09] combines non-parametric local and parametric global models to generate various set of face images. Recent family of GAN approaches [RMC16, KALL18, KLA19, BDS19] offers the state-of-the-art high quality random face generation without constraints.

Some other GAN-based studies allow to condition synthetic faces by rendered 3DMM images (i.e. Chapter 5), by landmarks [BJC18] or by another face image [BCW+18] (i.e. by disentangling identity and certain facial attributes). Similarly, facial expression is also conditionally synthesized by an audio input [JCZ19], by action unit codes [PAM+18], by predefined 3D geometry [ZLGS03] or by expression

of an another face image [LDW+14].

In this work, we jointly synthesize the aforementioned modalities for coherent photorealistic face synthesis by leveraging high-frequency generation by GANs. Unlike many of its 2D and 3D alternatives, the resulting generator models provide absolute control over disentangled identity, pose, expression and illumination spaces. Unlike many other GAN works that are struggling due to misalignments among the training data, our entire latent space correspond to realistic 3D faces as the data representation is naturally aligned on UV space.

### 4.2.3 Boosting face recognition by synthetic training data

There have been also some works to synthesize face images to be used as synthetic training data for face recognition methods either by directly using GAN-generated images [TMH18] or by controlling pose-space with a conditional-GAN [TYL19, HWY+18, SLY+18]. [MTH+16] propose many augmentation techniques, such as rotation, expression, and shape, based on 3DMMs. Other GAN-based approaches that capitalize 3D facial priors include [ZXJ+17], which rotates faces by fitting 3DMM and preserves photorealism by translation GANs and [YYS+17a], which frontalize face images by a GAN and 3DMM regression network. [DCX+18] complete missing parts of UV texture representations of 2D images after 3DMM fitting by a translation GAN. Our method in Chapter 5 first synthesizes face images of novel identities by sampling from 3DMM and then removes the photorealistic domain gap by an image-to-image translation GAN.

All of these studies show the significance of photorealistic and identity-generic face synthesization for the next generation of facial recognition algorithms. Although this study focuses more on the graphical aspect of face synthesization, we show that synthetic images can also improve face recognition performance.

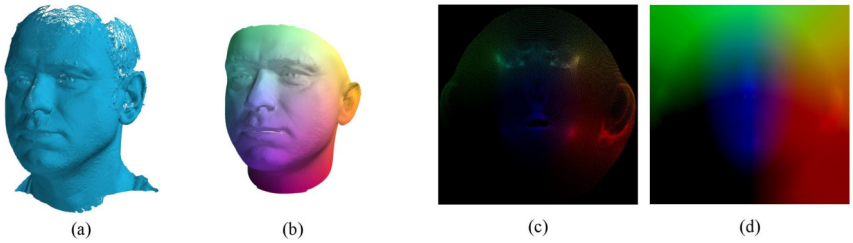



Figure 4.2: UV extraction process. In (a) we present a raw mesh, in (b) the registered mesh using the Large Scale Face Model (LSFM) template [BRZ+16], in (c) the unwrapped 3D mesh in the 2D UV space, and (d) the interpolated 2D UV map. Interpolation is carried out using the barycentric coordinates of each pixel in the registered 3D mesh.

## 4.3 Approach

### 4.3.1 UV Maps for Shape, Texture and Normals

In order to feed the shape, the texture, and the normals of the facial meshes into a deep network we need to reparameterize them into an image-like tensor format to apply 2D-convolutions [1]. We begin by describing all the raw 3D facial scans with the same topology and number of vertices (dense correspondence). This is achieved by morphing non-rigidly a template mesh to each one of the raw scans. We employ a standard non-rigid iterative closest point algorithm as described in [ARV07, DSVG11] and we deform our chosen template so that it captures correctly the facial surface of the raw scans. As a template mesh, we choose the mean face of the LSFM model proposed in [BRZ+16], which consists approximately of $54K$ vertices that are sufficient enough to depict non-linear, high facial details.

After reparameterizing all the meshes into the LSFM [BRZ+16] topology, we cylindrically unwrap the mean face of the LSFM [BRZ+16] to create a UV representation for that specific mesh topology. In the literature, a UV map is commonly utilized for storing only the RGB texture values. Apart from storing the texture values of the 3D meshes, we utilize the UV space to store the 3D coordinates of each vertex $(x, y, z)$ and the normal orientation $(n_x, n_y, n_z)$. Before storing the 3D coordinates into the UV space, all meshes are aligned in the 3D spaces by performing General Procrustes Analysis

[1] Another line of research is mesh convolutional networks [CBZ+19, LBBM18, RBSB18] which cannot preserve high-frequency details of the texture and normals at the current state-of-the-art.

(GPA) [Gow75] and are normalized to be in the scale of $[1, -1]$. Moreover, we store each 3D coordinate and normals in the UV space given the respective UV pixel coordinate. Finally, we perform a barycentric interpolation based on the barycentric coordinates of each pixel on the registered mesh to fill out the missing areas in order to produce a dense illustration of the UV map. In Fig. 4.2, we illustrate a raw 3D scan, the registered 3D scan on the LSFM [BRZ+16] template, the sparse UV map of 3D coordinates and finally the interpolated one.

### 4.3.2 Trunk-Branch GAN to Generate Coupled Texture, Shape and Normals

In order to train a model that handles multiple modalities, we propose a novel trunk-branch GAN architecture to generate entangled modalities of the 3D face such as texture, shape, and normals as UV maps. For this task, we exploit the MeIn3D dataset [BRZ+16] which consists of approximately 10,000 neutral 3D facial scans with wide diversity in age, gender, and ethnicity.

Given a generator network $\mathcal{G}^L$ with a total of $L$ convolutional upsampling layers and gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as input, the activation at the end of layer $d$ (i.e. $\mathcal{G}^d(\mathbf{z})$) is split into three branch networks $\mathcal{G}_T^{L-d}$, $\mathcal{G}_N^{L-d}$, $\mathcal{G}_S^{L-d}$ each of which consists of $L-d$ upsampling convolutional layers that generate texture, normals and shape UV maps respectively. The discriminator $\mathcal{D}^L$ starts with the branch networks $\mathcal{D}_T^{L-d}$, $\mathcal{D}_N^{L-d}$, $\mathcal{D}_S^{L-d}$ whose activations are concatenated before fed into trunk network $\mathcal{D}^d$. The output of $\mathcal{D}^L$ is regression of real/fake score.

Although the proposed approach is compatible with most of the GAN architectures and loss functions, in our experiments, we base TBGAN on progressive growing GAN architecture [KALL18] train it by WGAN-GP loss [GAA+17] as following:

$$\mathcal{L}_{\mathcal{G}^L} = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[-\mathcal{D}^L\left(\mathcal{G}^L(\mathbf{z})\right)\right] \tag{4.1}$$

$$\mathcal{L}_{\mathcal{D}^L} = \mathbb{E}_{x\sim p_{\text{data}},\, \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathcal{D}^L\left(\mathcal{G}^L(\mathbf{z})\right) - \mathcal{D}^L(x) + \lambda * GP(x, \mathcal{G}^L(\mathbf{z}))\right] \tag{4.2}$$

where gradient penalty calculated by $GP(x, \hat{x}) = \left(\|\nabla\mathcal{D}^L\left(\alpha\hat{x} + (1-\alpha)x\right)\|_2 - 1\right)^2$ and $\alpha$ denotes uniform random numbers between 0 and 1. $\lambda$ is a balancing factor which is typically $\lambda = 10$. An overview of this trunk-branch architecture is illustrated in Fig. 4.1

### 4.3.3 Expression Augmentation by Conditional GAN

Further, we modify our GAN in order to generate 3D faces with expression by conditioning it with expression annotations ($\mathbf{p_e}$). Similar to the MeIn3D dataset, we have captured approximately $35,000$ facial scans of around $5,000$ distinct identities during a special exhibition in the Science Museum, London. All subjects were recorded in various guided expressions with a 3dMD face capturing apparatus. All of the subjects were asked to provide meta-data regarding their age, gender, and ethnicity. The database consists of $46\%$ male, $54\%$ female, $85\%$ White, $7\%$ Asian, $4\%$ Mixed Heritage, $3\%$ Black, and $1\%$ other.

In order to avoid the cost and potential inconsistency of manual annotation, we render those scans and automatically annotate them by an expression recognition network. The resulting expression encodings ($(*, \mathbf{p_e}) \sim p_{\text{data}}$) are used as label vector during the training of our trunk-branch conditional GAN. This training scheme is illustrated in Fig. 4.1. $\mathbf{p_e}$ is basically a vector of 7 for universal expressions (neutral, happy, angry etc.), randomly drawn from our dataset. During the training, Eq. 4.1 and 4.2 are updated to condition expression encodings by AC-GAN [OOS17] as following:

$$\mathcal{L}_{\mathcal{G}^L} \mathrel{+}= \mathbb{E}_{(*,\mathbf{p_e})\sim p_{\text{data}},\ \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\sum_e \mathbf{p_e}\log(\mathcal{D}_e^L(\mathcal{G}^L(\mathbf{z},\mathbf{p_e})))\right] \tag{4.3}$$

$$\mathcal{L}_{\mathcal{D}^L} \mathrel{+}= \mathbb{E}_{(x,\mathbf{p_e})\sim p_{\text{data}},\ \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\sum_e \mathbf{p_e}\log(\mathcal{D}_e^L(x)) + \mathbf{p_e}\log(\mathcal{D}_e^L(\mathcal{G}^L(\mathbf{z},\mathbf{p_e})))\right] \tag{4.4}$$

which performs softmax cross entropy between expression prediction of the discriminator ($\mathcal{D}_e^L(x)$) and the random expression vector input ($\mathbf{p_e}$) for real ($x$) and generated samples ($\mathcal{G}^L(\mathbf{z},\mathbf{p_e})$).

Unlike previous expression models that omit the effect of the expression on textures, the resulting generator is capable of generating coupled texture, shape, and normals map of a face with controlled expression. Similarly, our generator respects the identity-expression correlation thanks to correlated supervision provided by the training data. This is in contrast to the traditional statistical expression models which decouples expression and identity models into two separate entities.

### 4.3.4 Photorealistic Rendering with Generated UV maps

For the renderings to appear photorealistic, we use the generated identity-specific mesh, texture, and normals, in combination with the generic reflectance properties, and employ a commercial rendering application: *Marmoset Toolbag* [Mar19].

In order to extract the 3D representation from the UV domain we employ the inverse procedure explained in section 4.3.1 based on the UV pixel coordinates of each vertex of the 3D mesh. Fig. 4.3 shows the rendering results, under a single light source, when using the generated geometry (Fig. 4.3(a)) and the generated texture (Fig. 4.3(b)). Here the specular reflection is calculated on the per-face normals of the mesh and exhibits steep changes between on the face's edges. By interpolating the generated normals on each face (Fig. 4.3(c)), we are able to smooth the specular highlights and correct any high-frequency noise on the geometry of the mesh. However, these results do not correctly model the human skin and resemble a metallic surface. In reality, the human skin is rough and as a body tissue, it both reflects and absorbs light, thus exhibiting specular reflection, diffuse reflection, and subsurface scattering.

Although we can add such modalities as additional branches with the availability of such data, we find that rendering can be still improved by adding some identity-generic maps. Using our training data, we create maps that define certain reflectance properties per-pixel, which will match the features of the average generated identity, as shown in bottom-left of Fig. 4.1. *Scattering* (c) defines the intensity of subsurface scattering of the skin. *Translucency* (d) defines the amount of light, that travels inside the skin and gets emitted in different directions. *Specular albedo* (e) gives the intensity of the specular highlights, which differ between hair-covered areas, the eyes, and the teeth. *Roughness* (f) describes the scattering of specular highlights and controls the glossiness of the skin. A *detail normal map* (g) is also tilled and added on the generated normal maps, to mimic the skin pores and a *detail weight map* (h) controls the appearance of the detail normals, so that they do not appear on the eyes, lips, and hair. The final result (Fig. 4.3(d)) properly models the skin surface and reflection, by adding plausible high-frequency specularity and subsurface scattering, both weighted by the area of the face where they appear.



(a) shape (b) Shp+tex (c) Shp+tex+nor (d) Final

Figure 4.3: Zoom-in on rendering results with (a) only the shape, (b) adding the albedo texture, (c) adding the generated normals, and (d) using identity-generic detail normal, specular albedo, roughness, scatter and translucency maps.

## 4.4 Results

In this section, we give qualitative and quantitative results of our method for generating 3D faces with novel identities and various expressions. In our experiments, there are total $L = 8$ up- and down-sampling layers where $d = 6$ of them in the trunk and 2 layers in each branch. These choices are empirically validated to ensure sufficient correlation among modalities without incompatibility artifacts. Running time is a few milliseconds to generate UV images from a latent code on a high-end GPU. Transforming from UV image to mesh is just sampling with UV coordinates and can be considered free of cost. Renderings in this chapter take a few seconds due to high resolution but this cost depends on the application. The memory needed for the generator network is 1.25GB compared to the 6GB PCA model of the same resolution and $95\%$ of the total variance.

In the following sections, we first visualize generated UV maps and their contributions to the final renderings on several generated faces. Next, we show the generalization ability of the identity and expression generators on some facial characteristics. We also demonstrate its well-generalization latent space by interpolating between different identities. Additionally, we perform full-head completion to the interpolated faces. Finally, we perform face recognition experiments by using the generated face images as additional training data.

### 4.4.1 Qualitative Results

**Combining coupled modalities:**

Fig. 4.4 presents the generated shape, normals, and texture maps by the proposed GAN and their additive contributions to the final renderings. As can be seen from local and global correspondences, the generated UV maps are highly correlated and coherent. Attributes like age, gender, race, etc. can be easily grasped from all of the UV maps and rendered images. Please also note that some of the minor artifacts of the generated geometry in Fig. 4.4(d) are compensated by the normals in Fig. 4.4(e).

**Diversity:**

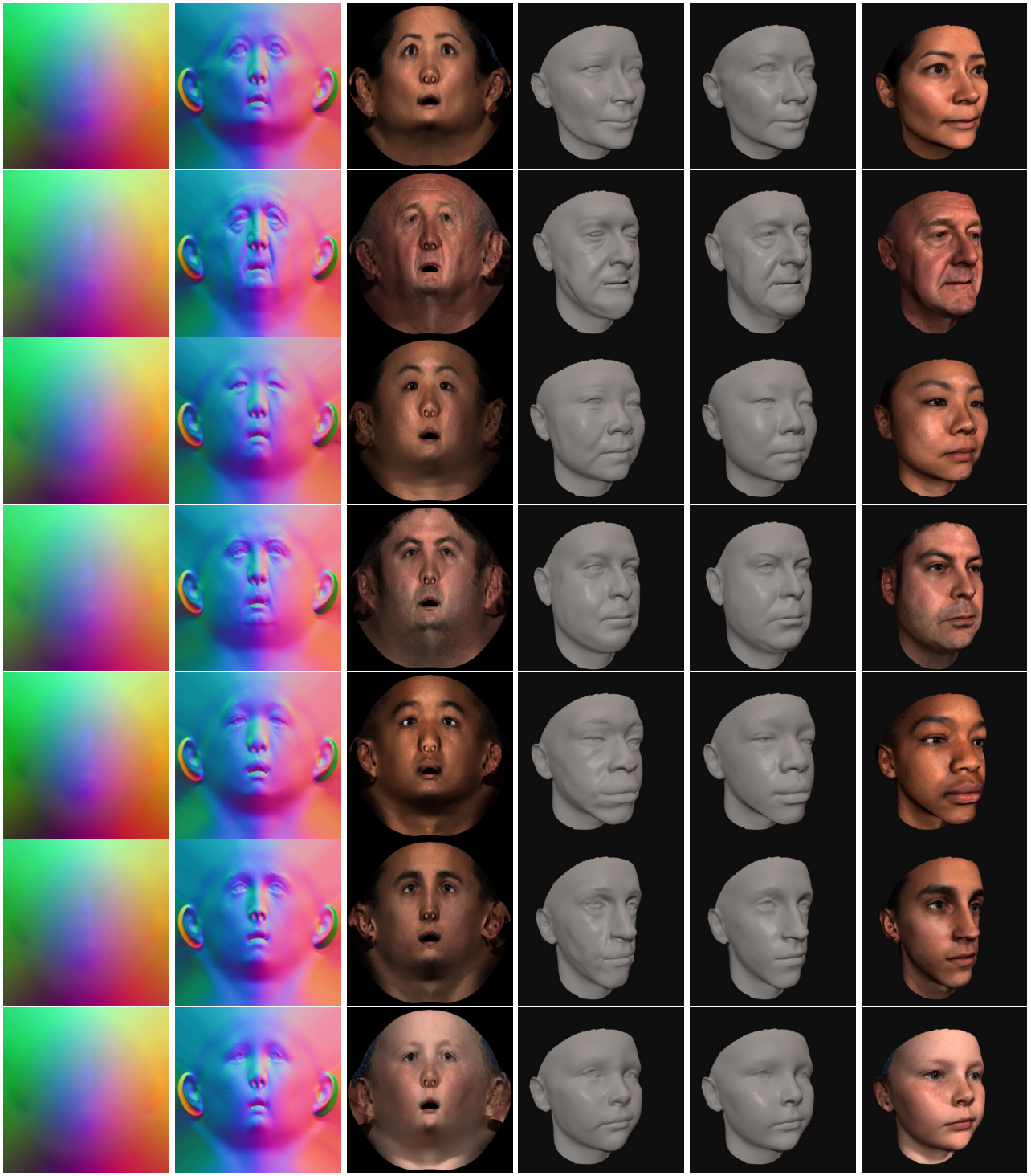
Our model is well-generalized with different age, gender, ethnicity groups and many facial attributes. Although Fig. 4.5 shows diversity in some of those categories, the reader is encouraged to see identity variation throughout the chapter and the supplementary video.

**Expression:**

We also show that our expression generator is capable of synthesizing quite a diverse set of expressions. Moreover, the expressions can be controlled by the input label as can be seen in Fig. 4.6. The reader is encouraged to see more expression generations in the supplementary video.

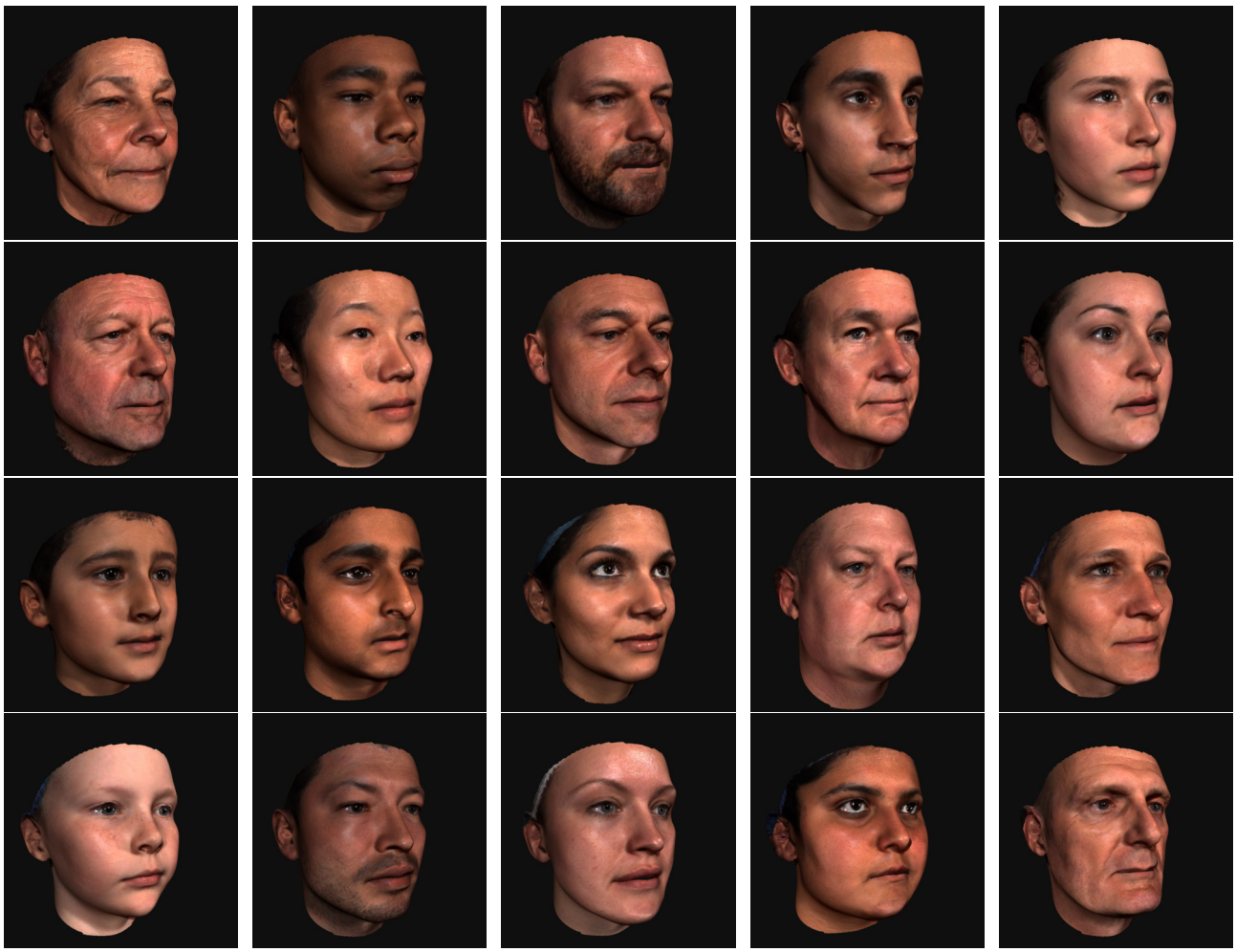
**Interpolation between identities:**

As shown in the supplementary video and in Fig. 4.7, our model can easily interpolate between any generation in a visually continuous set of identities which is another indication that the model is free from mode collapse. Interpolation is done by randomly generating two identities and generates faces by evenly spaced samples in latent space between the two.



(a) Shape (b) Normals (c) Texture (d) Shape (e) Shp+Nor (f) All

Figure 4.4: Generated UV representations and their corresponding additive renderings. Please note the strong correlation between UV maps, high fidelity and photorealistic renderings. The figure is best viewed in zoom.



(a) Age (b) Ethnicity (c) Gender (d) Weight (e) Roundness

Figure 4.5: Variation of generated 3D faces by our model. Each block shows diversity in a different aspect. Readers are encouraged to zoom in on a digital version.

**Full head completion:**

We also extend our facial 3D meshes to full head representations by employing the framework proposed in [PWP+19]. We achieve this by regressing from a latent space that represents only the 3D face to the PCA latent space of the Universal Head Model (UHM) [PWP+19, PVOS+20]. We begin by building a PCA model of the inner face based on the $10,000$ neutral scans of the MeIn3D dataset. Similarly, we exploit the extended full head meshes of the same identities utilized by UHM model and project them to the UHM subspace to acquire the latent shape parameters of the entire head topology. Finally, we learn a regression matrix by solving a linear least-square optimization problem as proposed in [PWP+19], which maps the latent space of the face shape to the full head representation. Fig. 4.7 demonstrates the extended head representations of our approach in conjunction with the



(a) Happy (b) Sadness (c) Anger (d) Fear (e) Disgust (f) Surprise

Figure 4.6: First and forth rows shows generations of six universal expressions. The other rows show texture and normals maps are used to generate the corresponding 3D faces. Please note how expressions are represented and correlated in the texture and normals space.

synthesized crop faces.

**Comparison to decoupled modalities and PCA:**

Results in Fig. 4.9 reveal a set of advantages of such unified 3D face modeling over separate GAN and statistical models. Clearly, the figure shows that the correlation among texture, shape, and normals is an important component for realistic face synthesis. Also, generations by PCA models are missing

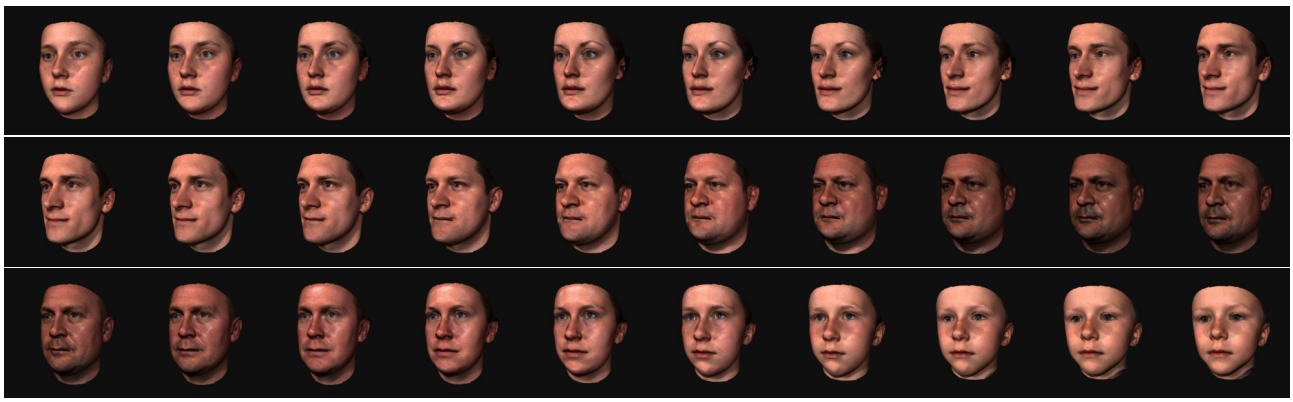


Figure 4.7: The figure shows the interpolation between pair of identities. Smooth transition indicates generalization of our GAN model.

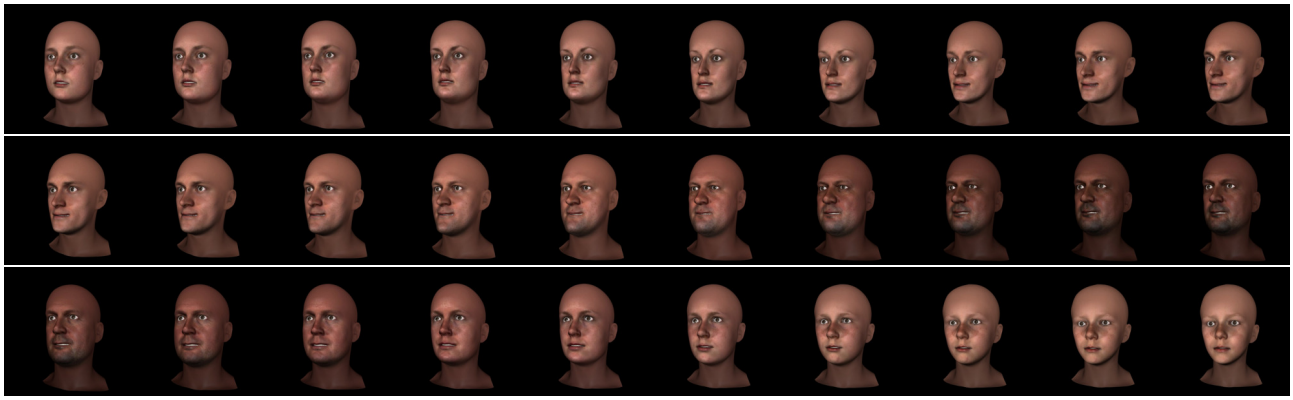


Figure 4.8: Complete full head representations in association with the facial topology corresponding in Fig.4.7. Even in the full head topology our generation methodology ensures a smooth transition during interpolation.

photorealism and details significantly.

### 4.4.2 Pose-invariant Face Recognition

In this section, we present an experiment that demonstrates that the proposed methodology can generate faces of different and diverse identities. That is, we use the generated faces to train one of the most recent state-of-the-art face recognition method, ArcFace [DGXZ19], and show that the proposed shape and texture generation model can boost the performance of pose-invariant face recognition.

**Training Data:** We randomly synthesize 10K new identities from the proposed model and render 50 images per identity with a random camera and illumination parameters from the Gaussian distribution of the 300W-LP dataset [ZLL+16, GBK+18]. For clarity, we call this dataset 'Gen' in the rest of




Figure 4.9: Comparison with seperate GAN models and PCA model. (a) Generation by our model. (b) Same texture with random shape and normals. (c) Same texture and shape with random normals (i.e. beard). (d) Generation by a PCA model constructed by the same training data and the same identity-generic rendering tools as explained in Sec. 3.4.

the text. Fig. 4.10 illustrates some examples of 'Gen' dataset which show larger pose variations than the real-world collected data. We augment 'Gen' with an in-the-wild training data, CASIA dataset [YLLL14], which consists of 10,575 identities with 494,414 images.

**Test Data:** For evaluation, we employ Celebrities in Frontal Profile (CFP) [SCC+16] and Age Database (AgeDB) [MPS+17]. **CFP** [SCC+16] consists of 500 subjects, each with 10 frontal and 4 profile images. The evaluation protocol includes frontal-frontal (FF) and frontal-profile (FP) face verification. In this chapter, we focus on the most challenging subset, CFP-FP, to investigate the performance of pose-invariant face recognition. There are 3,500 same-person pairs and 3,500 different-person pairs in CFP-FP for the verification test. **AgeDB** [MPS+17, DZZ17] contains $12,240$ images of $440$ distinct subjects. The minimum and maximum ages are $3$ and $101$, respectively. The average age range for each subject is $49$ years. There are four groups of test data with different year gaps ($5$ years, $10$ years, $20$ years and $30$ years, respectively) [DZZ17]. In this chapter, we only use the most challenging subset, AgeDB-30, to report the performance. There are 3,000 positive pairs and 3,000 negative pairs in AgeDB-30 for the verification test.

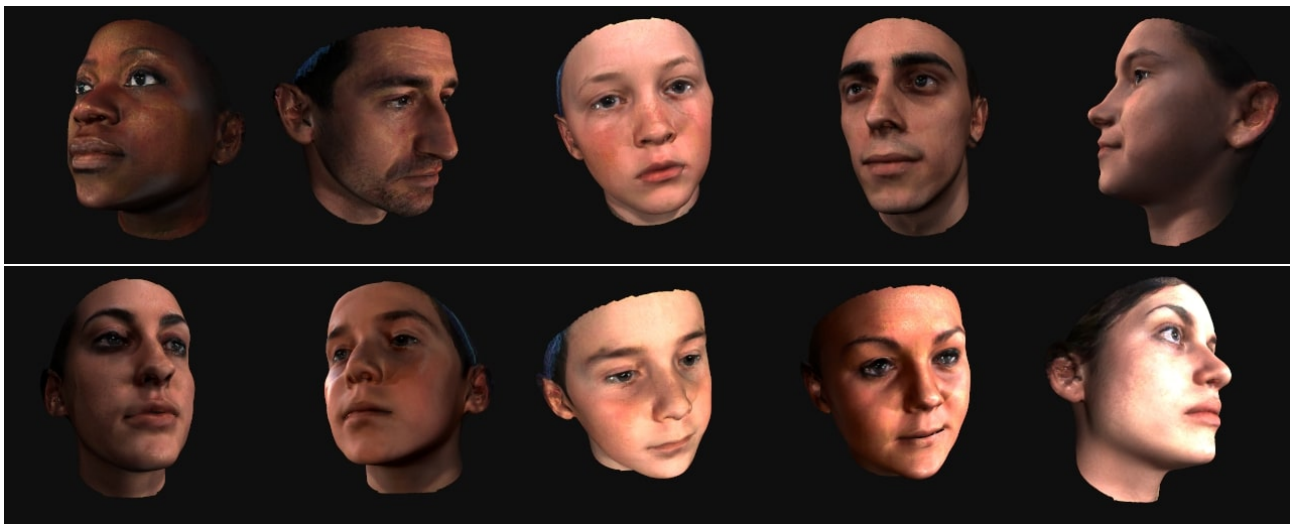


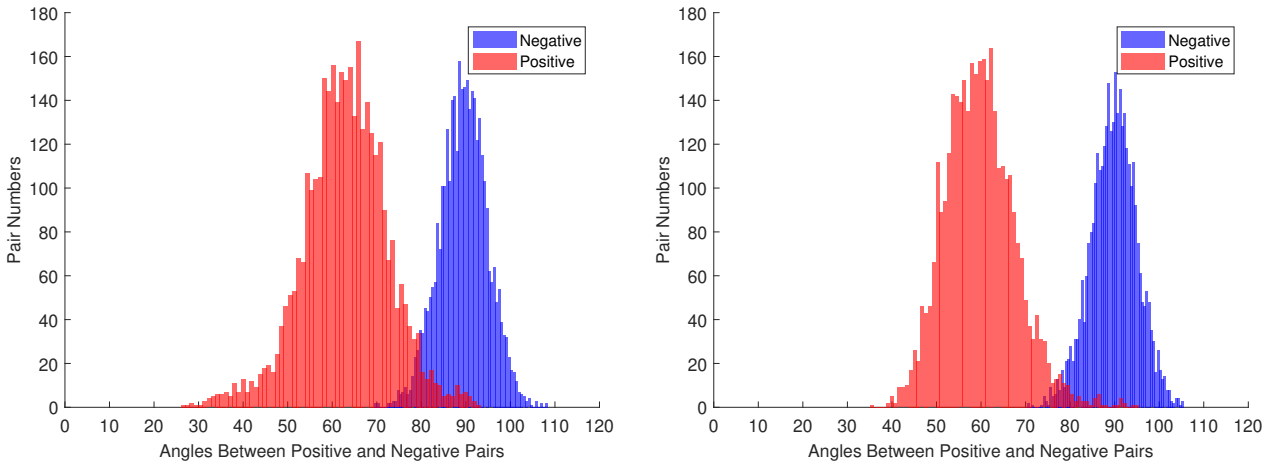
Figure 4.10: Examples of generated data ('Gen') by the proposed method.

**Data Preprocessing:** We follow the baseline [DGXZ19] to generate the normalized face crops ($112 \times 112$) by utilizing five facial points.

**Training and testing Details:** For the embedding networks, we employ the widely used ResNet50 architecture [HZRS16]. After the last convolutional layer, we also use the BN-Dropout-FC-BN [DGXZ19] structure to get the final $512$-$D$ embedding feature. For the hyper-parameter setting and loss functions, we follow [DGXZ19, GBK17, DZZ17]. The overlapping identities between the CASIA data set and the test set are removed for strict evaluations, and we only use a single crop for all testing.

**Result Analysis:** In Table 4.1, we show the contribution of the generated data on pose-invariant face recognition. We take UV-GAN [DCX+18] as the baseline method, which attaches the completed UV texture map onto the fitted mesh and generates instances of arbitrary poses to increase pose variation during training and minimize pose discrepancy during testing. As we can see from Table 4.1, generated data significantly boost the verification performance on CFP-FP from $95.56\%$ to $97.12\%$, decreasing the verification error by $51.2\%$ compared to the result of UV-GAN [DCX+18]. On AgeDB-30, combining CASIA and generated data achieves similar performance compared to using single CASIA because we only include intra-variance from pose instead of age.

In Figure 4.11, we show the angle distributions of all positive pairs and negative pairs from CFP-FP. By incorporating generation data, the overlap indistinguishable area between the positive histogram and the negative histogram is obviously decreased, which confirms that ArcFace can learn pose-




(a) CASIA (b) CASIA+Gen

Figure 4.11: Angle distributions of CFP-FP positive (red) and negative (blue) pairs in the 512-$D$ feature space. Red area indicates positive pairs while blue indicates negative pairs.

invariant feature embedding from the generated data. In Table 4.2, we select some verification pairs from CFP-FP and calculate the cosine distance (*angle*) between these pairs predicted by different models trained from the CASIA and combined data. Intuitively, the angles between these challenging pairs are significantly reduced when generated data are used for the model training.

## 4.5 Conclusion

In this chapter, we presented the first 3D face model for joint texture, shape, and normal generation based on Generative Adversarial Networks (GANs). The proposed GAN model implements a new architecture for exploiting the correlation between different modalities and can synthesize different facial expressions in accordance with the embeddings of an expression recognition network. We demonstrate that randomly synthesized images of our unified generator show strong relations between texture, shape, and normals and that rendering with normals provides excellent shading and overall

Table 4.1: Verification performance (%) of different models on CFP-FP and AgeDB-30.

| Methods | CFP-FP | AgeDB-30 |
|---|---|---|
| UVGAN [DCX+18] | 94.05 | 94.18 |
| Ours (CASIA) | 95.56 | 95.15 |
| Ours (CASIA+Gen) | **97.12** | **95.18** |

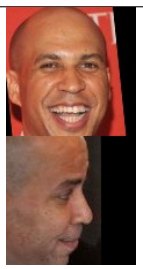
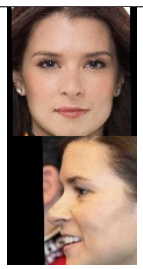
| Training Data |  |  |  |  |  |
|---|---|---|---|---|---|
| CASIA | 84.06° | 82.39° | 84.72° | 88.06° | 84.37° |
| CASIA+Gen | 57.60° | 63.12° | 66.10° | 59.72° | 60.25° |

Table 4.2: The angles between face pairs from CFP-FP predicted by different models trained from the CASIA and combined data. The generated data can obviously enhance the pose-invariant feature embedding.

visual quality. Finally, in order to demonstrate the generalization of our model, we have used a set of generated images to train a deep face recognition network.

Imposing 3D information into face generation can be in two ways, either we can synthesize 3D components as we presented in this chapter, or we can first render a 3D face into an image and formulate the problem as an image-to-image translation problem. In the next chapter, we tackle the problem of face synthesization with novel identities in this way.

# Chapter 5

# Generating Photorealistic Face Images from 3D Morphable Model


**5.1 Introduction** . . . . . **104**

**5.2 Related Works** . . . . . **107**

**5.3 Adversarial Identity Generation** . . . . . **110**

5.3.1 Unsupervised Domain Adaptation . . . . . 110

5.3.2 Adversarial Pair Matching . . . . . 111

5.3.3 Identity Preservation . . . . . 113

**5.4 Implementation Details** . . . . . **115**

**5.5 Results and Discussions** . . . . . **116**

5.5.1 Qualitative Evaluation . . . . . 117

5.5.2 The Added Realism and Identity Preservation . . . . . 118

5.5.3 Face Recognition with GANFaces dataset . . . . . 119

**5.6 Conclusions** . . . . . **121**


In this chapter, we propose a novel end-to-end semi-supervised adversarial framework to generate photorealistic face images of new identities with a wide range of expressions, poses, and illumina-

tions conditioned by synthetic images sampled from a 3D morphable model. Similar to the previous chapter, the proposed approach exploits 3D face modeling for photorealistic face synthesis. However, the method in this chapter, translate 3D faces into 2D image-space to simplify the problem.

Previous adversarial style-transfer methods either supervise their networks with a large volume of paired data or train highly under-constrained two-way generative networks in an unsupervised fashion. We propose a semi-supervised adversarial learning framework to constrain the two-way networks by a small number of paired real and synthetic images, along with a large volume of unpaired data. A set-based loss is also proposed to preserve identity coherence of generated images. Qualitative results show that generated face images of new identities contain pose, lighting and expression diversity. They are also highly constrained by the synthetic input images while adding photorealism and retaining identity information. We combine face images generated by the proposed method with a real data set to train face recognition algorithms and evaluate the model quantitatively on two challenging data sets: LFW and IJB-A.

## 5.1 Introduction

Deep learning has shown a great improvement in performance of several computer vision tasks [RHGS15, HGDG17, Gec16, GAM+18, DLHT16, DFI+15, YYS+17b] including face recognition [PVZ15, SKP15, XZT+15, LWY+17, XLZ+16] in the recent years. This was mainly thanks to the availability of large-scale datasets. Yet the performance is often limited by the volume and the variations of training examples. Larger and wider datasets improve the generalization and overall performance of the model [SKP15, BCRC17].

The process of collecting and annotating training examples for every specific computer vision task is laborious and non-trivial. To overcome this challenge, additional synthetic training examples along with limited real training examples can be utilised to train the model. Some of the recent works such as 3D face reconstruction [RSK16], gaze estimation [ZSFB15, WBM+16], human pose, shape and motion estimation [VRM+17] etc. use additional synthetic images generated from 3D models to train deep networks. One can generate synthetic face images using a 3D morphable model

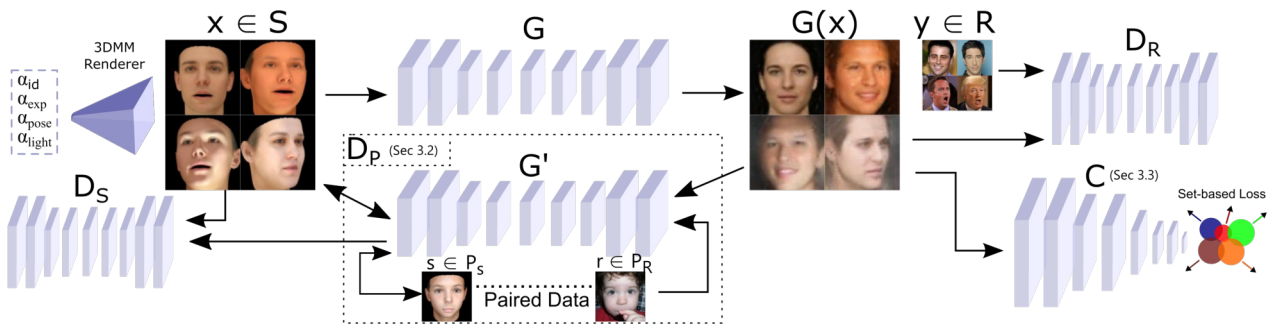



Figure 5.1: Our approach aims to synthesize photorealistic images conditioned by a given synthetic image by 3DMM. It regularizes cycle consistency [ZPIE17] by introducing an additional adversarial game between the two generator networks in an unsupervised fashion. Thus the under-constrained cycle loss is supervised to have correct matching between the two domains by the help of a limited number of paired data. We also encourage the generator to preserve face identity by a set-based supervision through a pretrained classification network.

(3DMM) [BV99] by manipulating identity, expression, illumination, and pose parameters. However, the resulting images are not photorealistic enough to be suitable for in-the-wild face recognition tasks. It is because the information of real face scans is compressed by the 3DMM and the graphical engine that models illumination and surface is not perfectly accurate. Thus, the main challenge of using synthetic data obtained from 3DMM model is the discrepancy in the nature and quality of synthetic and real images which poses the problem of domain adaptation [PGLC15]. Recently, adversarial training methods [SPT+17, SWL16, CGM+18] have become popular to mitigate such challenges.

Generative Adversarial Network (GAN), introduced by Goodfellow et al. [Goo16], and its variants [RMC16, KALL18, BSM17, DBP+19] are quite successful in generating realistic images. However, in practice, GANs are likely to stuck in mode collapse for large scale image generation. They are also unable to produce images that are 3D coherent and globally consistent [Goo16]. To overcome these drawbacks, we propose a semi-supervised adversarial learning framework to synthesize photorealistic face images of new identities exhibiting extensive data variation supplied by a 3DMM. We address these shortcomings by exciting a generator network with synthetic images sampled from 3DMM and transforming them into photorealistic domain using adversarial training as a bridge. Unlike most of the existing works that excite their generators with a noise vector [RMC16, BSM17], we feed our generator network by synthetic face images. Such a strong constraint naturally helps in avoiding the mode collapse problem, one of the main challenges faced by the current GAN methods. Fig. 5.1 shows a general overview of the proposed method. We discuss the proposed method in more

details in Sec. 5.3.

In this chapter, we address the challenge of generating photorealistic face images from 3DMM rendered faces of different identities with arbitrary poses, expressions, and illuminations. We formulate this problem as a domain adaptation problem i.e. aligning the 3DMM rendered face domain into realistic face domain. One of the previous works closest to ours [IZZE17] addresses the style transfer problem between a pair of domains with classical conditional GAN. The major bottleneck of this method is that it requires a large number of paired examples from both domains which are hard to collect. CycleGAN [ZPIE17], another recent method and closest to our work, proposes a two-way GAN framework for unsupervised image-to-image translation. However, the cycle consistency loss proposed in their method is satisfied as long as the transitivity of the two mapping networks is maintained. Thus, the resulting mapping is not guaranteed to produce the intended transformation. To overcome the drawbacks of these methods [IZZE17, ZPIE17], we propose to use a small amount of paired data to train an inverse mapping network as a matching aware discriminator. In the proposed method, the inverse mapping network plays the role of both the generator and the discriminator. To the best of our knowledge, this is the first attempt for adversarial semi-supervised style translation for an application with such limited paired data.

Adding realism to the synthetic face images and preserving their identity information is a challenging problem. Although synthetic input images, 3DMM rendered faces, contain distinct face identities, the distinction between them vanishes as a result of the inherent non-linear transformations induced by the discriminator to encourage realism. To tackle such a problem, prior works either employ a separate pre-trained network [YYS+17a] or embed Identity labels (id) [TYL17] into the discriminator. Unlike existing works, which are focused on generating new images of existing identities, we are interested in generating multiple images of new identities. Therefore, such techniques are not directly applicable to our problem. To address this challenge, we propose to use set-based center [WZLQ16] and pushing loss functions [GBK17] on top of a pre-trained face embedding network. This will keep track of the changing average of embeddings of generated images belonging to the same identity (i.e. centroids). In this way identity preservation becomes adaptive to the changing feature space during the training of the generator network unlike softmax layer that converges very quickly at the beginning of the training before meaningful images are generated.

Our contributions can be summarized as follows:

- We propose a novel end-to-end adversarial training framework to generate photorealistic face images of new identities constrained by synthetic 3DMM images with identity, pose, illumination and expression diversity. The resulting synthetic face images are visually plausible and can be used to boost face recognition as additional training data or any other graphical purposes.
- We propose a novel semi-supervised adversarial style transfer approach that trains an inverse mapping network as a discriminator with paired synthetic-real images.
- We employ a novel set-based loss function to preserve consistency among unknown identities during GAN training.

## 5.2 Related Works

In this Section we discuss the prior art that are closely related to the proposed method.

**Domain Adaptation.** As stated in the introduction, our problem of generating photorealistic face images from 3DMM rendered faces can be seen as a domain adaptation problem. A straightforward adaptation approach is to align the distributions at the feature level by simply adding a loss to measure the mismatch either through second-order moments [SS15] or with adversarial losses [THDS15, THSD17, GUA+16].

Recently, pixel level domain adaptation became popular due to practical breakthroughs on Kullback-Leibler divergence [GPM+14, Goo16, RMC16], namely GANs which optimize a generative and discriminative network through a mini-max game. It has been applied to a wide range of problems including fashion clothing [LPG17], person specific avatar creation [WTP17], text-to-image synthesis [ZXL+17], face frontalization [YYS+17a], and retinal image synthesis [CGM+18].

Pixel domain adaptation can be done in a supervised manner simply by conditioning the discriminator network [IZZE17] or directly the output of the generator [CK17] with the expected output when there

is enough paired data from both domains. Please note collecting a large number of paired training examples is expensive, and often requires expert knowledge. [RAY+16] proposes a text-to-image synthesis GAN with a matching aware discriminator. They optimize their discriminator for image-text matching beside requiring realism with the information provided by additional mismatched text-image pairs.

For the cases where paired data is not available, many approaches adapt unsupervised learning such as imposing pixel-level consistency between input and output of the generator network [BSD+17, SPT+17], an encoder architecture that is shared by both domains[BTS+16] and adaptive instance normalization [HB19]. An interesting approach is to have two way translation between domains with two distinct generator and discriminator networks. They constrain the two mappings to be inverses of each other with either ResNet [ZPIE17] or encoder-decoder network [LBK17] as the generator.

**Synthetic Training Data Generation.** The usage of synthetic data as additional training data is shown to be helpful even if they are graphically rendered images in many applications such as 3D face reconstruction [RSK16], gaze estimation [ZSFB15, WBM+16], human pose, shape and motion estimation [VRM+17]. Despite the availability of almost infinite number of synthetic images, those approaches are limited due to the domain difference from that of in-the-wild images.

Many existing works utilize adversarial domain adaptation to translate images into photorealistic domain so that they are more useful as training data. [ZZY17] generates many unlabeled samples to improve person re-identification in a semi-supervised fashion. RenderGAN [SWL16] proposes a sophisticated approach to refine graphically rendered synthetic images of tagged bees to be used as training data for a bee tag decoding application. WaterGAN [LSEJ18] synthesizes realistic underwater images by modeling camera parameters and environment effects explicitly to be used as training data for a color correction task. Some studies deform existing images by a 3D model to augment diverse datasets [MTH+16] without adversarial learning.

One of the recent works, simGAN [SPT+17], generates realistic synthetic data to improve eye gaze and hand pose estimation. It optimizes the pixel level correspondence between input and output of the generator network to preserve the content of the synthetic image. This is in fact a limited solution

since the pixel-consistency loss encourages the generated images to be similar to synthetic input images and it partially contradicts adversarial realism loss. Instead, we employ an inverse translation network similar to cycleGAN [ZPIE17] with an additional pair-wise supervision to preserve the initial condition without hurting realism. This network also behaves as a discriminator to a straight mapping network trained with real paired data to avoid possible biased translation.

**Identity Preservation.** To preserve the identity/category of the synthesized images, some of the recent works such as [CDH+16, TYL17] keep categorical/identity information in discriminator network as an additional task. Some of the others propose to employ a separate classification network which is usually pre-trained [LTT17, YYS+17a]. In both these cases, the categories/identities are known beforehand and are fixed in number. Thus it is trivial to include such supervision in a GAN framework by training the classifier with real data. However such setup is not feasible in our case as images of new identities to-be-generated are not available to pre-train a classification network.

To address the limitation of existing methods of retaining identity/category information of synthesized images, we employ a combination of different set-based supervision approaches for unknown identities to be distinct in the pre-trained embedding space. We keep track of moving averages of same-id features by the momentum-like centroid update rule of center loss [WZLQ16] and penalize distant same-id samples and close different-id samples by a simplified variant of the magnet loss[RPDB16] without its sophisticated sampling process and with only a single cluster per identity (see Section 5.3.3 for further discussions).

**Recent Related Works.** Very recently, Generative Adversarial Networks have shown great success in generating high resolution 2D face images [KALL18, KLA19, KLA+20] from a semantically meaningful latent space. Tewari et al. [TEB+20] disentangles these latent spaces in terms of 3DMM parameters to control face generation by identity, expression, pose, lighting etc. Their method trains a multi-layer perceptron in a self-supervised fashion to estimate StyleGAN [KLA19] latent parameters in correspondence to a given 3DMM sample. So, the translation happens in the latent space. In contrast, our method performs domain translation in the image-space which can focus on local changes (i.e. photorealism) and keep the global information more accurate (i.e. identity, pose). An-

other difference is that their method heavily relies on a good quality pretrained GAN network such as StyleGAN [KLA19], whereas we train the generator GAN within our approach.

## 5.3 Adversarial Identity Generation

In this Section, we describe in details the proposed method. Fig. 5.1 shows a schematic diagram of our method. Specifically, the synthetic image set $x \in \mathcal{S}$ is formed by a graphical engine for the randomly sampled of 3DMM with its identity, pose and lighting parameters $\alpha$. The generated images they are translated into a more photorealistic domain $G(x)$ through the network, $G$, and mapped back to its synthetic domain ($G'(G(x))$) through the network, $G'$, to retain $x$. The adversarial synthetic and real domain translation of $G$ and $G'$ networks are supervised by the discriminator networks $D_R$ and $D_S$, with an additional adversarial game between $G$ and $G'$ as a generator and a discriminator respectively. During training, the identities generated by 3DMM are preserved with a set-based loss on a pre-trained embedding network $C$. In the following sub-sections, we further describe these components i.e. domain adaptation, real-synthetic pair discriminator, and identity preservation.

### 5.3.1 Unsupervised Domain Adaptation

Given a 3D morphable model (3DMM) [BV99], we synthesize face images of new identities sampled from its Principal Components Analysis (PCA) coefficients' space with random variation of expression, lighting and pose. Similar to [ZPIE17], a synthetic input image ($x \in \mathcal{S}$) is mapped to a photorealistic domain by a residual network ($G : S \rightarrow \hat{R}$) and mapped back to the synthetic domain by a 3DMM fitting network ($G' : \hat{R} \rightarrow \hat{S}$) to complete the forward cycle only[1]. To preserve cycle consistency, the resulting image $G'(G(x))$ is encouraged to be the same as input $x$ by a pixel level $L_1$ loss:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x \in \mathcal{S}} \|G'(G(x)) - x\|_1 \tag{5.1}$$

[1]We empirically found that removing the backward cycle-loss improves performance when the task is to map from artificial images to real as also shown in Tab.4 of [ZPIE17]

In order to encourage the resulting images $G(x)$ and $G'(G(x))$ to have a similar distribution as real and synthetic domains respectively, those refiner networks are supervised by discriminator networks $D_R$ and $D_S$ with images of the respective domains. The discriminator networks are formed as auto-encoders as in the boundary equilibrium GAN (BEGAN) architecture [BSM17] in which the generator and discriminator networks are trained by the following adversarial training formulation:

$$\mathcal{L}_G = \mathbb{E}_{x \in \mathcal{S}} \|G(x) - D_R(G(x))\|_1 \tag{5.2}$$

$$\mathcal{L}_{G'} = \mathbb{E}_{x \in \mathcal{S}} \|G'(G(x)) - D_S(G'(G(x)))\|_1 \tag{5.3}$$

$$\mathcal{L}_{D_R} = \mathbb{E}_{x \in \mathcal{S}, y \in \mathcal{R}} \|y - D_R(y)\|_1 - k_t^{D_R} \mathcal{L}_G \tag{5.4}$$

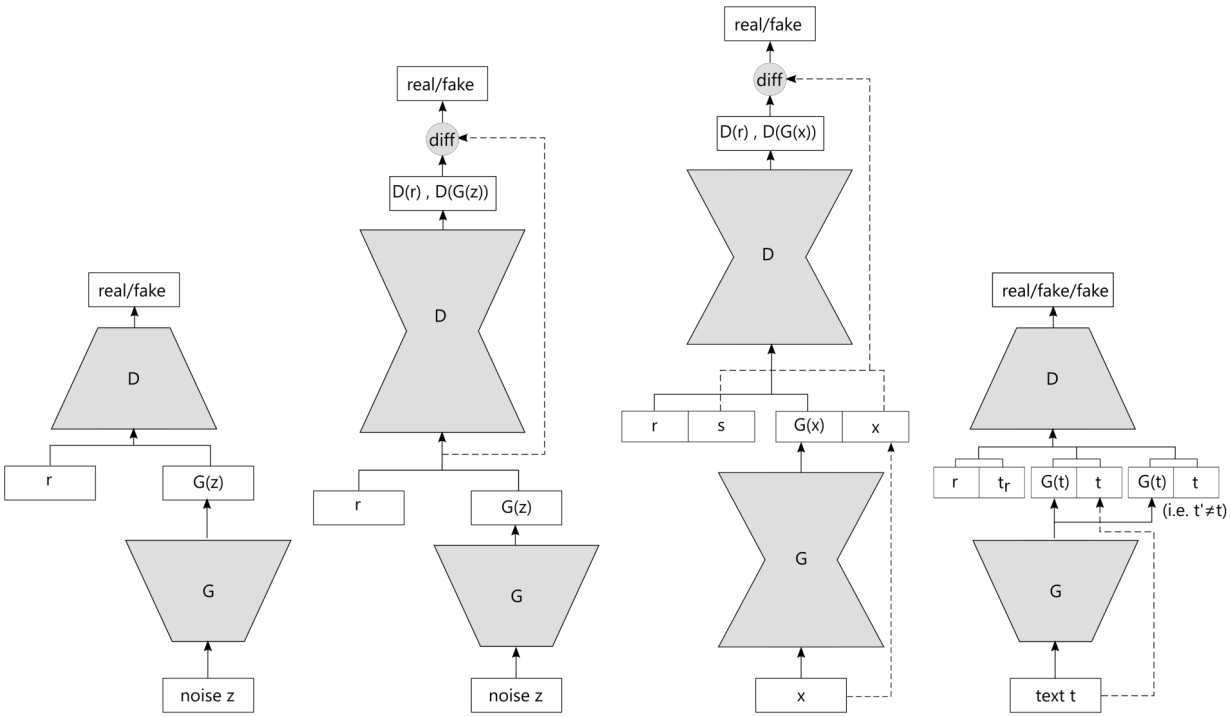
$$\mathcal{L}_{D_S} = \mathbb{E}_{x \in \mathcal{S}} \|x - D_S(x)\|_1 - k_t^{D_S} \mathcal{L}_{G'} \tag{5.5}$$

where for each training step $t$ and the generator network ($G$ for $k_t^{D_R}$, $G'$ for $k_t^{D_S}$) we update the balancing term with $k_t^D = k_{t-1}^D + 0.001(0.5\mathcal{L}_D - \mathcal{L}_G)$. As suggested by [BSM17], this term helps to maintain a balance between the interests of the generator and discriminator and stabilize the training.

### 5.3.2 Adversarial Pair Matching

The cycle consistency loss ensures the bijective transitivity of functions $G$ and $G'$ which means generated image $G(x) \in \hat{R}$ should be transformed back to $x \in \hat{S}$. Convolutional networks are highly under-constrained and they are free to make any unintended changes as long as the cycle consistency is satisfied. Therefore, without an additional supervision, it is not guaranteed to achieve the correct mapping that preserves shape, texture, expression, pose and lighting attributes of the face image from domains $S$ to $\hat{R}$ and $\hat{R}$ to $\hat{S}$. This problem is often addressed by introducing pixel-level penalization between input and output of the networks [ZPIE17, SPT+17] which is sub-optimal for domain adaptation as it encourages to stay in the same domain.

To overcome this issue, we propose an additional pair-wise adversarial loss that assigns the $G'$ network an additional role as a pair-wise discriminator to supervise the $G$ network. Given a set of paired




(a) DC-GAN[RMC16] (b) BEGAN [BSM17] (c) Ours (d) GAN-CLS [RAY+16]

Figure 5.2: Comparison of our pair matching method to the related work. (a) In the traditional GAN approach, the discriminator module aligns the distribution of real and synthetic images by means of a classification network. (b) BEGAN[BSM17] and many others showed that the alignment of reconstruction error distributions offers a more stable training. (c) We propose to utilize this autoencoder approach to align the distribution of pairs to encourage generated images to be transformed to the realistic domain with a game between real and synthetic pairs. (d) An alternative to our method is to introduce wrongly labeled images to the discriminator to teach pair-wise matching as proposed by [RAY+16] for text to images synthesis.

synthetic and real images $(\mathcal{P}_\mathcal{S}, \mathcal{P}_\mathcal{R})$, the discriminator loss is computed by BEGAN as follows:

$$\mathcal{L}_{D_P} = \mathbb{E}_{s\in\mathcal{P}_\mathcal{S}, r\in\mathcal{P}_\mathcal{R}} \|s - G'(r)\|_1 - k_t^{D_P}\mathcal{L}_{\text{cyc}} \tag{5.6}$$

While the $G'$ network is itself a generator network ($G' : \hat{R} \rightarrow \hat{S}$) with a separate discriminator ($D_S$), we use it as a third pair-matching discriminator to supervise $G$ by means of a distribution of paired correspondence of real and synthetic images. Thus while cycle-loss optimizes for the biject correspondences, we expect the resulting pairs of $(x \in S, G(x) \in \hat{R})$ to have the similar correlation distribution as paired training data $(s \in \mathcal{P}_\mathcal{S}, r \in \mathcal{P}_\mathcal{R})$. Fig 5.2 shows its relation to the previous related art and comparison to an alternative which is a matching aware discriminator with paired inputs for text to image synthesis, as suggested by [RAY+16]. Please note that how BEGAN autoencoder



Figure 5.3: Quality of 9 images of 3 identities (per row) during the training. Red plot shows the proposed identity preservation loss ($x$ axis) over the iterations ($y$ axis). Note the changes of fine-details on the faces which is the main motivation of set-based identity preservation.

architecture is utilized to align the distribution of pairs of synthetic and real images with synthetic and generated images.

Alternatively, one could pretrain the $G'$ network as a 3DMM fitting network as in [THMM17, SYH+17, TZK+17, CBK+17] . However, we trained it from scratch to balance the adversarial zero-sum game between the generator ($G$) and the pair-wise discriminator ($G'$). Otherwise the gradient would vanish as there would be no success in fooling the discriminator. Moreover, those networks provide only fitted 3DMM parameters which then would need to be rendered into 3DMM images by a differentiable tensor operation.

### 5.3.3 Identity Preservation

Although identity information is provided by the 3DMM in shape and texture parameters, it may be lost to some extent by virtue of a non-linear transformation. Some studies [YYS+17a, TYL17] address this issue by employing identity labels of known subjects as additional supervision either with a pre-trained classification network or within the discriminator network. However, we intend to generate images of new identities sampled from the 3DMM parameter space and their photorealistic images simply do not exist yet. Furthermore, training a new softmax layer and the rest of the framework simultaneously becomes a chicken-egg problem and results in failed training.

In order to preserve identity in the changing image space, we propose to adapt a set-based approach over a pre-trained face embedding network. We import the idea of pulling same-id samples as well as pushing close samples from different identities in the embedding space such that same-id images are

gathered and distinct from other identities regardless of the quality of the images during the training. At the embedding layer of a pre-trained network $C$, the generator network ($G$) is supervised by a combination of the center [WZLQ16] and pushing loss [GBK17] (which is also a simplified version of the Magnet loss [RPDB16] formulation) defined for a given mini-batch (M) as:

$$\mathcal{L}_C = \mathbb{E}_{x \in \mathcal{S}, i_x \in \mathbb{N}^+} \sum_{x}^{M} -\log \frac{\exp(\frac{1}{2\sigma^2}\|C(G(x)) - c_{i_x}\|_2^2 - \eta)}{\sum_{j \neq i_x} \exp(\frac{1}{2\sigma^2}\|C(G(x)) - c_j\|_2^2)} \tag{5.7}$$

where $i_x$ stands for the identity label of $x$ provided by 3DMM sampling and $c_j$ stands for the mean embedding of identity $j$. The margin term, $\eta$, is set to 1 and the variance is computed by $\sigma = \frac{\sum_x^M \|C(G(x)) - c_{i_x}\|_2^2}{M-1}$.

While the quality of images is improved during the training, i.e. better photo-realism, their projection on the embedding space is shifting. In order to adapt to those changes, we update identity centroids ($c_j$) with a momentum of $\beta = 0.95$ when new images of id $j$ become available. Following [WZLQ16], for a given $x$, the moving average of an identity centroid is calculated by $c_j^{t+1} = c_j^t - \beta\delta(i_x = j)(c_j^t - C(G(x)))$ where $\delta(condition) = 1$, if the condition is satisfied and $\delta(condition) = 0$ if not. Centroids ($c_j$) are initialized with zero and after few iterations, they converge to embedding centers and then continue updating to adapt to the changes caused by the simultaneous training of $G$. Fig. 5.3 shows the quality of 9 images of 3 identities over training iterations. Please note the difference of the images after convergence with the images at the beginning of the training, produced by the Softmax layer which fails to supervise the forthcoming images in later iterations.

## Full Objective

Overall, the framework is optimized by the following updates simultaneously:

$$\theta_G = \arg\min_{\theta_G} \mathcal{L}_G + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_C\mathcal{L}_C \tag{5.8}$$

$$\theta_{G'} = \arg\min_{\theta_{G'}} \mathcal{L}_{G'} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{D_P}\mathcal{L}_{D_P} \tag{5.9}$$

$$\theta_{D_R}, \theta_{D_S} = \arg\min_{\theta_{D_R}, \theta_{D_S}} \mathcal{L}_{D_R} + \mathcal{L}_{D_S} \tag{5.10}$$

where $\lambda$ parameters balance the contribution of different modules. The selection of those parameters is discussed in the next section.

## 5.4 Implementation Details

**Network Architecture:** For the generator networks ($G$ and $G'$), we use a shallow ResNet architecture as in [JAF16] which supplies smooth transition without changing the global structure because of its limited capacity, having only 3 residual blocks. In order to benefit from 3DMM images fully, we also add skip connections to the network $G$. We also add dropout layers after each block in the forward pass with a 0.9 keep rate to introduce some noise that could be caused by uncontrolled environmental changes.

We construct the discriminator networks ($D_R$ and $D_S$) as autoencoders trained by boundary equilibrium adversarial learning with Wasserstein distance as proposed by [BSM17]. The classification network $C$, is a shallow FaceNet architecture [SKP15], more specifically we use NN4 network with an input size of $96 \times 96$ where we randomly crop, rotate and flip generated images $G(x)$ which are of size $108 \times 108$.

**Data:** Our framework needs a large amount of real and synthetic face images. For real face images, we use CASIA-Web Face Dataset [YLLL14] that consists of ∼500K face images of ∼10K individuals.

The proposed method trains the $G'$ network as a discriminator ($D_P$) with a small number of paired real and synthetic images. For that, we use a combination of 300W-3D [STZP13b, STZP13a, BAP+17] and AFLW2000-3D datasets as our paired training set [ZLL+16] which consist of 5K real images with their corresponding 3DMM parameter annotations. We render synthetic images by those parameters and pair them with the matching real images. This dataset is relatively small, compared to the ones used by fully supervised transformation GANs (i.e. Amazon Handbag dataset used by [IZZE17] contains 137K bag images).

We randomly sample face images of new identities as our synthetic data set using Large Scale Face Model (LSFM) [BRZ+16] for shape, Basel Face Model [PKA+09] for texture and Face Warehouse model [CWZ+14] for expression. While the shape and texture parameters of new identities are sampled from the Gaussian distribution of the original model, expression, lighting and pose parameters are sampled with the same Gaussian distribution as that of synthetic samples of 300-3D [STZP13b, STZP13a, BAP+17] and AFLW2000-3D [ZLL+16]. All images are aligned by MTCNN [ZZLQ16] and centre cropped to the size of $108 \times 108$ pixels.

**Training Details:** We train all the components of our framework together from scratch except for the classification network $C$ which is pre-trained by using a subset of Oxford VGG Face Dataset [PVZ15]. The whole framework takes about 70 hours to converge on a Nvidia GTX 1080TI GPU in 248K iterations with a batch size of 16. We start with a learning rate of $8 \times 10^{-5}$ with ADAM solver [KB15] and halve it after 128Kth, 192Kth, 224Kth, 240Kth, 244Kth, 246Kth and 247Kth iterations.

As shown in Eqn. 5.8, 5.9, $\lambda$ is a balancing factor which controls the contribution of each optimization. We set $\lambda_{\text{cyc}} = 0.5$, $\lambda_{D_P} = 0.5$, $\lambda_C = 0.001$ to achieve a balance between realism, cycle-consistency, identity preservation and the supervision by the paired data. We also add identity loss ($\mathcal{L}_{\text{id}} = \|x - G(x)\|$) as suggested by [ZPIE17] to regularize the training with a balancing term $\lambda_{\text{id}} = 0.1$. During the training, we keep track of moving averages of the network parameters to generate images.

## 5.5 Results and Discussions

In this section, we show the qualitative and quantitative results obtained with the proposed framework. We also discuss and show the contribution of each module (i.e. $\mathcal{L}_{\text{cyc}}$, $D_P$, $C$) with an ablation study in the supplementary materials. For the experiments, we generate 500,000 and 5,000,000 images of 10,000 different identities with variations on expression, lighting and poses. We name this synthetic dataset **GANFaces**[2] (i.e. GANFaces-500K, GANFaces-5M).

[2]The dataset, training code, pre-trained models and face recognition experiments can be viewed at https://github.com/barisgecer/facegan

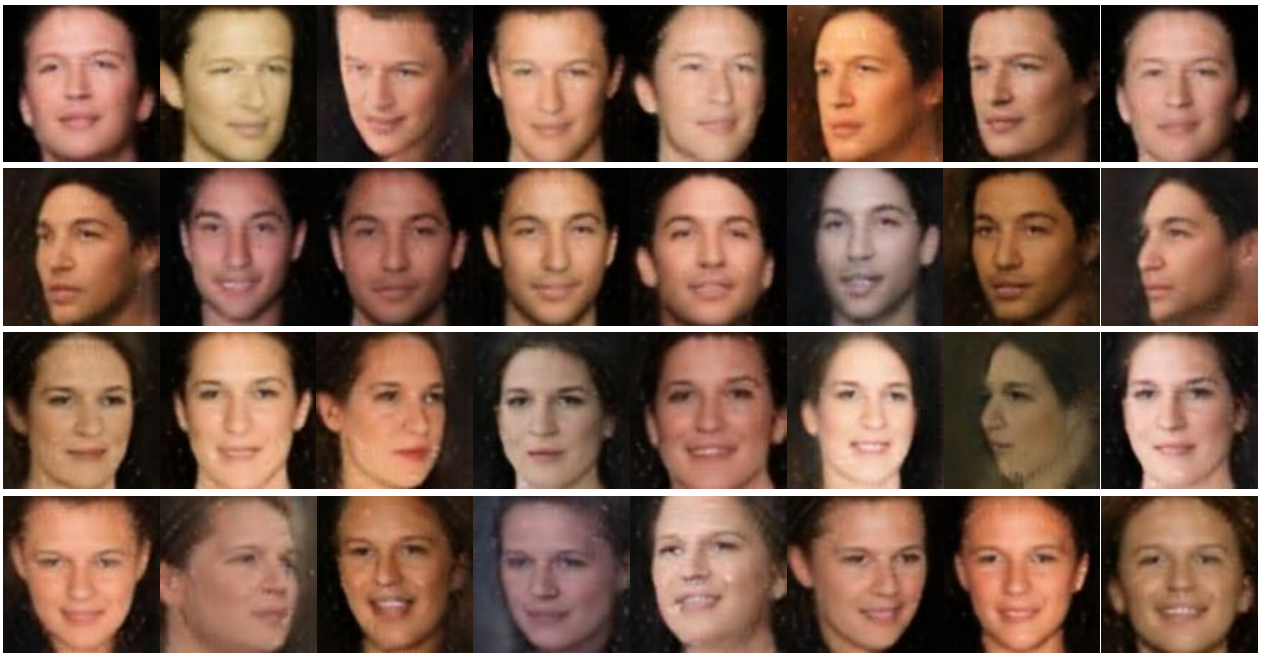


Figure 5.4: Random samples from GANFaces dataset. Each row belongs to the same identity. Notice the variation in pose, expression and lighting.

### 5.5.1 Qualitative Evaluation

Please see Fig. 5.4 for random samples from the dataset. Fig. 5.5 compares our results (left half of the Fig. 5.4) with the 3DMM inputs, the results with simGAN [BSD+17] and cycleGAN [ZPIE17] settings, and our setup with the addition of the reconstruction loss of the paired data within the $G$ network. We observe good correspondence when we compare first 4 columns of Fig.5.4 to Fig. 5.5(a) in terms of identity, pose, expression and lighting. Compared to ours (Fig. 5.4), [BSD+17] suffers from the loss of identity-specific facial features (Fig. 5.5(b)) while [ZPIE17] generates images visually less pleasant (Fig. 5.5(c)). An additional reconstruction loss used in our framework to train the $G$ network with the paired data produces the results in Fig. 5.5(d). We achieved less clear images by this step probably because of the severity of the influence of the direct reconstruction loss on the adversarial balance. The superiority of the proposed framework is also confirmed by the quantitative experiments shown in Table. 5.1.

One of the main goals of this work is to generate face images guided by the attributes of synthetic input images i.e. shape, expression, lighting, and poses. We can see from Fig. 5.6 that our model is capable of generating photorealistic images preserving the attributes conditioned by the synthetic



(a) 3DMM inputs (b) simGAN [[BSD+17] (c) CycleGAN [[ZPIE17] (d) Reconstruction Err.

Figure 5.5: Comparison to (a) input 3DMM images, (b) results with simGAN settings [BSD+17],(c) results with cycleGAN settings [ZPIE17] and (d) results with additional reconstruction loss. Figures correspond to left half of the Fig. 5.4 and each row belongs to the same identity.

input images. In the figure, top row shows the variations of pose and expression on input synthetic faces and the left column shows the input synthetic faces of different identities. The rest are the images generated by our model, conditioned on the corresponding attributes from the top row and the left column. We can clearly see that the conditioning attributes are preserved on the images generated by our model. We can also observe that fine-grained attributes such as shapes of chin, nose and eyes are also retained in the images generated by our model. In the case of extreme poses, the quality of the image generated by our model becomes less sharp as the CASIA-WebFace dataset, which we used to learn the parameters of discriminator network $D_R$, lacks a sufficient number of examples with extreme poses.

### 5.5.2 The Added Realism and Identity Preservation

In order to show that synthetic images are effectively transformed to the realistic domain with preserving identities, we perform a face verification experiments on GANFaces dataset. We took pre-trained face-recognition CNN network, namely FaceNet NN4 architecture [SKP15] trained on CASIA-WebFace [YLLL14] to compute the features of the face images. The verification performance of the network on LFW is $95.6\%$ accuracy and $95.5\%$ 1-EER which shows that the model is well optimized for in-the-wild face verification. We created 1000 similar (belonging to same identity) and 1000 dis-similar (belonging to different identities) face image pairs from GANFaces. Similarly, we also generated the same number of similar and dis-similar face image pairs from the VGG face dataset [PVZ15] and the synthetic 3DMM rendered faces dataset. Fig. 5.7 shows the histograms of euclidean distances between sim-

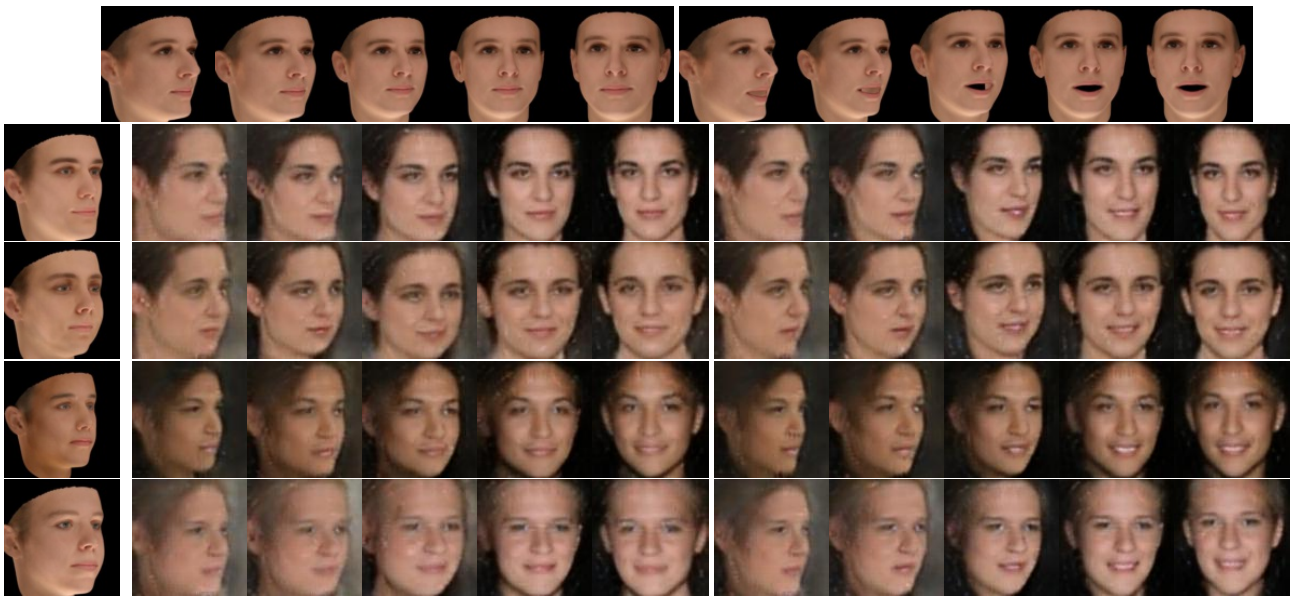


Figure 5.6: Images generated by the proposed approach conditioned by identity variation in the vertical axis, normalized and mouth open expression in left and right blocks and pose variation in the horizontal axis. Images in this figure are not included in the training

ilar and dis-similar images measured in the embedding space for the three datasets. The addition of realism and preservation of identities of the GANFaces can be seen from the comparison of its distribution to the 3DMM synthetic dataset distribution. As the images become more realistic, they become better separable in the pre-trained embedding space. We also observe that the separation of positive and negative pairs of GANFaces is better than that of VGG faces pairs. The probable reason for VGG not achieving a better separation than GANFaces is noisy face labels as indicated in the original study [PVZ15].

### 5.5.3 Face Recognition with GANFaces dataset

We augmented GANFaces with real face dataset i.e. VGG Faces [PVZ15] and trained the VGG19 [SZ15] network and tested its performance on two challenging datasets: LFW [HRBL07] and IJB-A [KKT+15]. We restrict ourselves from limited access to full access of real face dataset and train deep network on different combination of real and GANFaces. Following [MTH+16], we use a pre-trained VGGNet by [SZ15] with 19 layers trained on the ImageNet dataset [RDS+15] and took these parameters as initial parameters. We train the network with different portions of the Oxford VGG Face

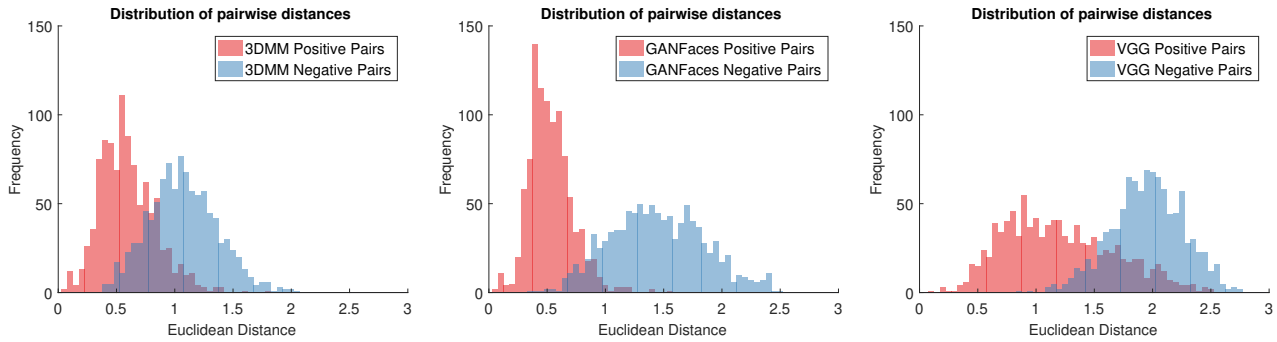



Figure 5.7: Distances of 1000 positive and 1000 negative pairs from three different datasets (GANFaces, 3DMM synthetic images, Oxford VGG) embedded on a NN4 network that is trained with CASIA Face dataset

dataset [PVZ15], augmented with the GANFaces dataset. We remove the last layer of the deep VGGNet and add two soft-max layers to the previous layer, one for each of the datasets. The learning rate is set to 0.1 for the soft-max layers and 0.01 to the pre-trained layers with the ADAM optimizer. Also we halve the gradient coming from the GANFaces soft-max. We decrease the learning rate exponentially and train for 80,000 iterations where all of our models converge well without overfitting. For a given input size of $108 \times 108$, we randomly crop and flip $96 \times 96$ patches and the overall training takes around 9 hours on a GTX 1080TI GPU.

We train 6 models with $20\%$, $50\%$ and $100\%$ of the VGG Face dataset with and without the augmentation of GANFaces-500K. As seen in Fig. 5.8, we evaluate the models on LFW and IJB-A datasets and the benchmark scores are improved with the addition of this dataset even though the image resolution is low. The contribution of GANFaces-500K increases inversely proportional to the number of images included from the VGG dataset, which indicates more synthetic images might improve the

| Method | Real | Synth | Test time Synth | Image size | Acc. (%) | 100% - EER |
|---|---|---|---|---|---|---|
| FaceNet [SKP15] | 200M | - | No | 220×220 | 98.87 | - |
| VGG Face [PVZ15] | 2.6M | - | No | 224×224 | 98.95 | 99.13 |
| Masi et al. [MTH+16] | 495K | 2.4M | Yes | 224×224 | 98.06 | 98.00 |
| Yin et al. [YYS+17a] | 495K | 495K | Yes | 100×100 | 96.42 | - |
| VGG + Recons. Err. | 1.8M | 500K | No | 96×96 | 94.7 | 94.8 |
| VGG + simGAN [SPT+17] | 1.8M | 500K | No | 96×96 | 94.7 | 94.8 |
| VGG + cycleGAN [ZPIE17] | 1.8M | 500K | No | 96×96 | 94.5 | 94.7 |
| VGG(100%) | 1.8M | - | No | 96×96 | 94.8 | 94.6 |
| VGG(100%) + GANFaces-500K | 1.8M | 500K | No | 96×96 | **94.9** | **95.1** |
| VGG(100%) + GANFaces-5M | 1.8M | 5M | No | 96×96 | **95.2** | **95.1** |

Table 5.1: Comparison with state-of-the-art studies on LFW performances

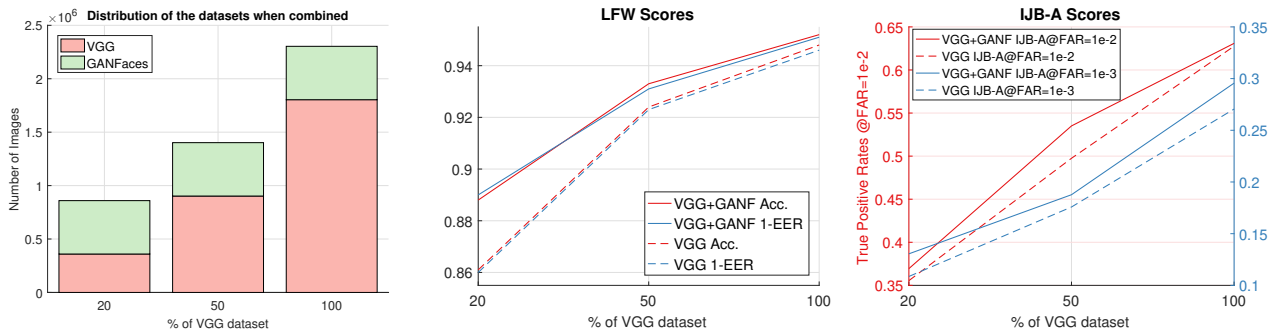



Figure 5.8: Face recognition benchmark experiments. (Left) Number of images used from the two datasets in the experiments. The total number of images in the VGG data set is 1.8M since some images were removed from the URL (Middle) Performance on the LFW dataset with (solid) and without (dashed) the GANFaces-500K dataset. (Right) True Positive Rates on the IJB-A verification task with (solid) and without (dashed) the GANFaces-500K dataset.

results even further.

We compare our best model trained by full VGG dataset and GANFaces to the other state of the art methods in Table 5.1. Despite the lower resolution, GANFaces was able to improve our baseline to the numbers comparable to the state-of-the-art. Note that generative methods, such as [MTH+16, YYS+17a], do generation (i.e. pose augmentation and normalization) in the test time whereas we use only given test images. Together with the benefit of low resolution, this makes our models more efficient at test time.

## 5.6 Conclusions

This chapter proposes a novel end-to-end semi-supervised adversarial training framework to generate photorealistic faces of new identities with wide ranges of poses, expressions, and illuminations from 3DMM rendered faces. Our extensive qualitative and quantitative experiments show that the generated images are realistic and identity preserving.

We generated a synthetic dataset of face images closer to a photorealistic domain and combined it with a real face image dataset to train a face recognition CNN and improved the performance in recognition and verification tasks. In the future, we plan to generate millions of high resolution images of thousands of new identities to boost the state-of-the-art face recognition.

The proposed framework helps to avoid some of the common GAN problems such as mode collapse and 3D coherency. It shows how the data generated by 3DMM or any other explicit model can be utilized to improve and control the behaviour of GANs.

# Chapter 6

# Robustifing 3D Expression Reconstruction of Codec Avatars


**6.1 Introduction** . . . . . **124**

**6.2 Methodologies** . . . . . **126**

6.2.1 Baseline . . . . . 126

6.2.2 3D Data Augmentation . . . . . 127

6.2.3 Generic Encoders . . . . . 129

**6.3 Experiments** . . . . . **130**

**6.4 Conclusion** . . . . . **134**


So far, we presented works that put emphasis on facial identity. In this chapter, we change the direction of the thesis slightly to focus on expression synthesis/reconstruction rather than identity. For that, we fix the identity by using personalized models where each model cover one identity with various expressions.

Recently, there has been a few deep learning-based approaches proposed for personalized appearance modeling [LSSS18, WSS+19, NHSW20]. Particularly, Lombardi et al. [LSSS18] proposes an approach to model 3D facial appearance of a person by deep generative networks, called *Codec Avatars*.

The *Codec Avatars* are built by collecting a large amount of data collected in a multi-camera capture setup, while subjects are performing a predefined set of 122 facial expressions. The setup consists of 40 synchronous cameras capturing $5120 \times 3840$ images at 30 frames per second, lying around a hemisphere and pointing towards the subject. A secondary dataset is collected with a similar set of expressions performed by the subjects while they wear a commercially available Virtual Reality (VR) headsets with 3 cameras (i.e. Head-Mounted Cameras (HMCs)) installed which are pointing to the eyes and the mouth. Although these two sequences of expressions are not in temporal correspondence, [LSSS18] proposes a novel approach to exploit the correlation which, eventually, leads to live animation of *Codec Avatars* by VR headsets. A potential application of this system is a dual-way 3D photorealistic telepresence in a VR platform.

In this chapter, we propose a few approaches to improve expression animation performance of *Codec Avatars*. First, we augment training images of expression encoders in 3D. Then, we exploit the whole dataset of multiple subjects to generalize personalized models. The proposed approaches are evaluated under challenging settings and show promising performance.

## 6.1 Introduction

The recent developments of Virtual Reality (VR) technologies has enabled countless new applications. Just like the audio-calls and telegram has been lifted to 2D video-calls in the last decade, human interactions can be lifted into 3D virtual telepresence. The recent studies [LSSS18, WSS+19] on Photorealistic Telepresence in Virtual Reality (VR) environment has shown promising result as a candidate for authentic 3D telecommunication media of the next era.

The aforementioned studies has proposed novel approaches for creating authentic 3D avatars from a multi-camera capture system, namely, *Codec Avatars*, which can be animated by headset mounted cameras (HMC) that are commercially available on VR headsets. More specifically, two variational autoencoders (VAE) are trained on *HMC* images and multi-camera capture scans(called *Mugsy*) as person-specific models of a single subject. Then, the correspondence between the two latent spaces is acquired by self-supervised domain adaptation techniques and the encoder of *HMC* VAE and the

decoder of *Mugsy* VAE is connected end-to-end along with domain transfer layers. Eventually, the *Codec Avatar* of a subject can be animated by transferring the precise expression from *HMC* inputs of a VR headset worn by the same subject. Since, this kind of VR headset allows both receiving and transmitting virtual 3D faces, this approaches enables dual-way 3D telecommunication in VR.

Although the current pipeline of [LSSS18, WSS+19] can acquire precise correspondence between *HMC* and *Mugsy* VAE spaces, it still contains strong reliance on the training data from *HMC* and is sensitive to slight changes in perspective, lighting, background and day-to-day facial attribute changes (e.g. make-up, facial hair etc.). This is mainly due to *HMC* data collection process which has been performed under the same environment conditions in one session. For example, a slight perspective changes can cause large changes in 2D and the encoder fails to embed the expression precisely which was never seen during training. While each subject's data contain fixed background, lighting, pose etc., there is a variation among different subjects on these conditions. Therefore, there is still a room for improvement in robustness to those variations by exploiting the variation other subjects' data.

In this chapter, we investigate various augmentation techniques to robustify the current animation pipeline of *Codec Avatars* by *HMC* input given the correspondence between the two provided by [WSS+19], namely *Rosetta*. First, we attempt to enrich training process of expression encoders by augmenting HMC images in 3D. We basically exploit the geometry and camera parameters optimized by [WSS+19] and make slight random changes to the camera parameters to simulate arbitrarily worn headset conditions. Such 3D augmentation is particularly more helpful for HMC images than 2D augmentation as the cameras are placed very close to the face. Secondly, we leverage the data of multiple subjects to further generalize person-specific models. Basically, we train a common expression encoder network from HMC images of multiple subjects to a common latent presentation by a self-supervised reconstruction loss. The reconstruction loss is calculated based on person-specific Mugsy models of each identity after transferring the common expression representation to person-specific latent representation.

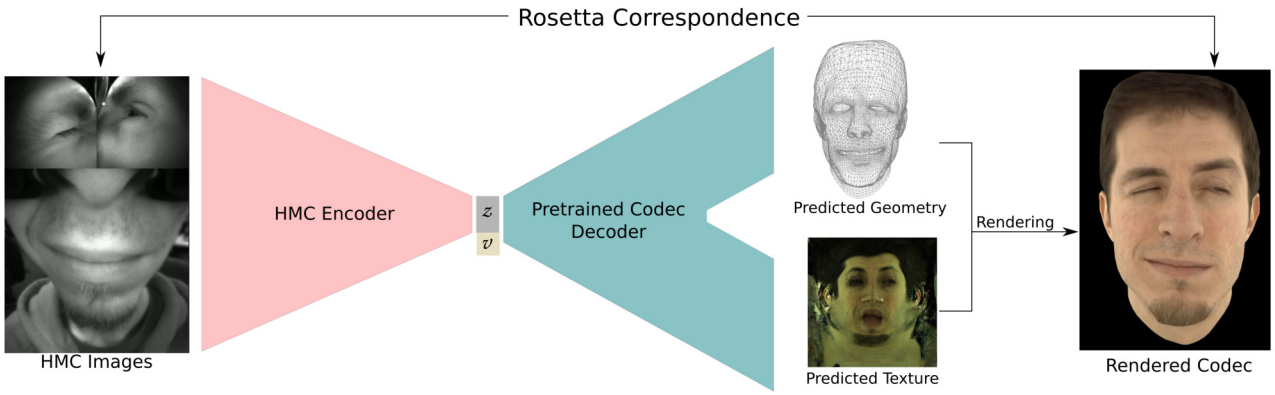



Figure 6.1: Once the correspondence is established by [WSS+19], one can re-train the HMC encoder with different augmentation techniques such as brightness, gamma, 2D translation, rotation, distortion and cropping.

## 6.2 Methodologies

With the availability of correspondence between HMC and Mugsy latent spaces, one can further improve expression reconstruction from HMC to Codec Avatars. Since many attributes of the HMC images are subject to change (such as facial hair, background, camera perspectives), HMC encoder network can be trained to robust to that kind of changes. In the following sections, we first briefly explain Rosetta and the baseline encoder training. And secondly, we explain 3D augmentation of HMC images by Rosetta geometry parameters in order to make HMC encoder less sensitive to perspective changes. Finally, we further robustify HMC encoder by leveraging multiple identity data in a single HMC encoder.

### 6.2.1 Baseline

Previously, [WSS+19] proposed an approach to solve the correspondence between HMC images ($\mathbf{I}^{\text{HMC}}$) and mesh and texture from Mugsy captures ($\mathbf{M} \in \mathbb{R}^{7306\times 3}, \mathbf{T} \in \mathbb{R}^{w\times h}$) by exploiting domain adaptation techniques such as CycleGAN. After training a VAE on Mugsy data, this method exploits cycle consistency between HMC images style-transfered to look like Mugsy images and Codec Avatars rendered in HMC camera parameters by simultaneously fine-tuning camera parameters and training an HMC encoder network ($\mathbf{z} \in \mathbb{R}^{256} \leftarrow \mathcal{E}(\mathbf{I}^{\text{HMC}})$). During the test time, the HMC encoder

network infer codec parameters ($\mathbf{z}$) which was then used to render Codec Avatar by the geometry and texture obtained from the Mugsy-VAE Decoder ($\mathbf{M}' \in \mathbb{R}^{7306\times3}, \mathbf{T}' \in \mathbb{R}^{w\times h} \leftarrow \mathcal{D}(\mathbf{z}, \mathbf{v})$). For more technical details about this approach we refer the reader to the original publications.

Unfortunately, the encoder network overfits to the HMC data used during this training and becomes sensitive to changes in perspective, lighting, background and day-to-day facial attributes (i.e. make-up or facial hair). However once the correspondence is established, one can re-train the HMC encoder ($\mathcal{E}$) with different augmentation techniques as illustrated in Fig. 6.1. For this, we augment HMC images in 2D by common jittering techniques such as brightness, gamma, 2D translation, rotation, distortion and cropping augmentation, denoted as $\mathbf{I}' \leftarrow \mathcal{A}_{2D}(\mathbf{I})$. And then we train a *baseline* HMC encoder by the following loss function.

$$\mathbf{M}', \mathbf{T}' = \mathcal{D}(\mathcal{E}(\mathcal{A}_{2D}(\mathbf{I}^{\text{HMC}})), \mathbf{v}) \tag{6.1}$$

$$\mathcal{L}_{\text{baseline}} = ||\mathbf{M} - \mathbf{M}'||_2 + \lambda||\mathbf{T} - \mathbf{T}'|| \tag{6.2}$$

**Golden Plane Correction :** Another baseline approach to solve the perspective-overfitting problem is by homography correction. Assuming that camera placement of each headset might be different, this approach aims to calibrate them by rectifying HMC images to a mutual perspective, called Golden plane. This rectification process is a standard calibration process among different hardware setups and we do not get into details for brevity.

### 6.2.2 3D Data Augmentation

Head-mounted cameras are subject to significant perspective changes from capture to capture due to its close distance to face. This often causes overfitting to a certain perspective of training set and failure on unseen perspectives. In an ideal scenario, one should perform non-rigid registration of training and testing captures in 3D, but unfortunately, this would require geometry and camera parameters available during the test time which is not the case. Similarly, 2D alignment by the detected keypoints

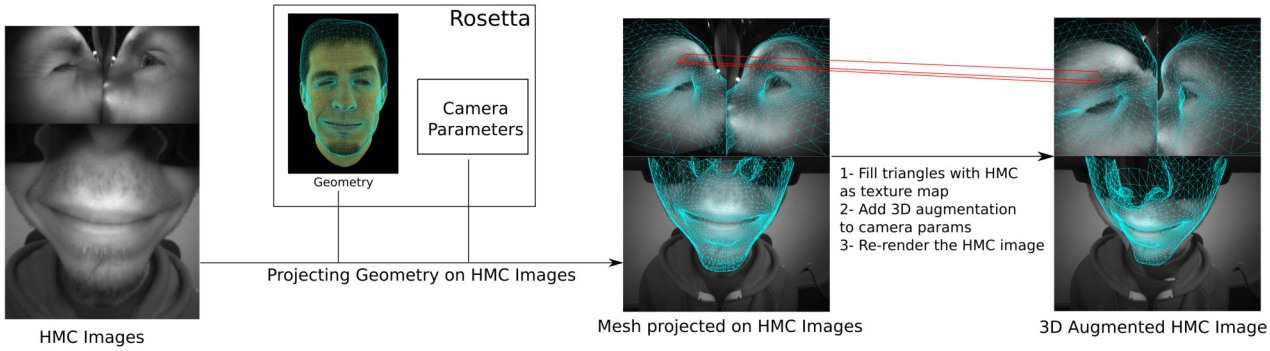



Figure 6.2: In order to 3D augment a given HMC image, we first project 3D mesh on the image with known geometry and camera parameters. After filling triangles with HMC image with the projected 2D vertex coordinates, we render the mesh with a small changes to rotation and translation of the camera parameters. We repeat this procedure randomly during the training.

also fails when the perspective difference is large. While 2D augmentation partially helps to address this issue, in this section, we explore augmenting the training set by rotating and translating HMC images in 3D by the geometry and camera parameters of *Rosetta* [WSS+19].

In order to 3D augment a given HMC image and its corresponding *Rosetta* parameters, we first project 3D mesh on the image with geometry and camera parameters by *Rosetta*. In order to reconstruct texture of 3D mesh for the visible areas, we fill the triangles with the projected 2D vertex coordinates on the HMC image. And then, we render the mesh with barycentric interpolation regarding the vertices and by applying target camera parameters.

This process is similar to UV textured mesh rendering, but this time using HMC images instead of UV texture maps with projected vertex coordinates rather than UV texture coordinates. Since this method is implemented as a PyTorch layer and does not fully unwrap UV map, it is quite efficient as well. Thus we can repeat this augmentation randomly during the training to simulate unseen perspectives of headset cameras with a reasonable range of movement. This approach is illustrated in Fig. 6.2 and denoted as $\mathbf{I}' \leftarrow \mathcal{A}_{3D}(\mathbf{I})$. Finally, a new encoder network is trained by the following:

$$\mathbf{M}', \mathbf{T}' = \mathcal{D}(\mathcal{E}(\mathcal{A}_{3D}(\mathbf{I}^{\text{HMC}})), \mathbf{v}) \tag{6.3}$$

$$\mathcal{L}_{\text{3D}} = ||\mathbf{M} - \mathbf{M}'||_2 + \lambda||\mathbf{T} - \mathbf{T}'|| \tag{6.4}$$

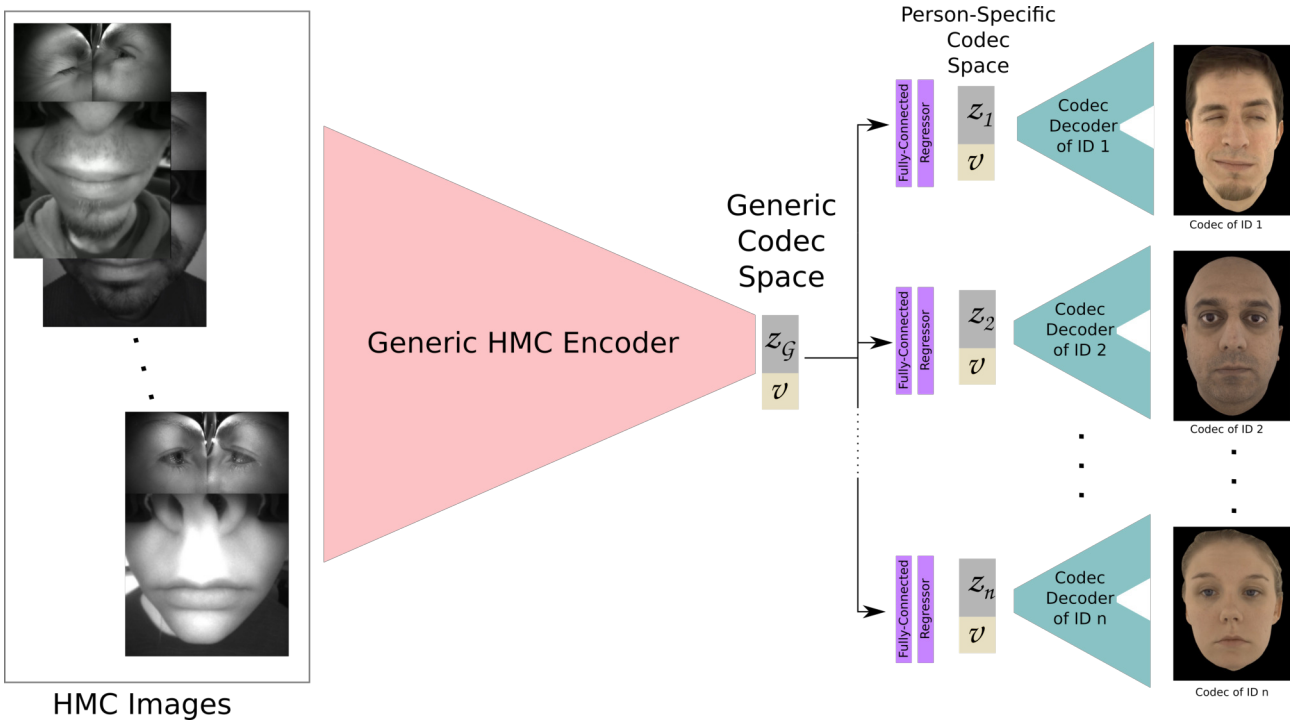



Figure 6.3: Training architecture of the Generic HMC Encoder by the data of multiple subjects. The key insight is to encode all HMC images into a common expression space and then reconstruct them in the corresponding personalized Codec spaces. Aggregation of all subjects provides better generalization to the enviromental conditions that appear in all capture sessions.

### 6.2.3 Generic Encoders

All previous approaches handles pipeline for each identity separately. However, each identity contains one or two capture sessions in our dataset and therefore person-specific HMC encoders gets specialized for the environment conditions of that particular session(s). Consequently, the resulting encoder becomes very sensitive any changes of the input images during test time such as different lighting, facial hair or make-up, pose/perspective, background etc. In order to generalize and robustify expression encoder network to all the conditions appear in the dataset, we propose the architecture in Fig.6.3 to train with person-specific models of multiple identities.

The idea here is to exploit the fact that the complete expression spectrum is mostly overlap for each identity. The Codec's parametric spaces ($\mathbf{z}_i$), however, are totally different for each identity as the VAEs in *Rosetta* are trained separately. Therefore, we propose to employ a deep common encoder network accompanied by shallow transformation networks for each identity. So that all HMC images

are encoded into a common expression space and then transformed into personalized Codec spaces. Since we can now use the original Codec Decoders by *Rosetta*, common expression network and the latent transformation networks can be supervised by Codec reconstruction losses in Eq.6.1.

More specifically, we train a *Generic HMC Encoder* network ($\mathcal{E}_\mathcal{G}$) for multiple identities simultaneously. For $n$ number of identities, we have combined HMC images of all identities such as $\{\mathbf{I}_i^{\text{HMC}} | i \in 1, 2, \ldots, n\}$ and encode them by $\mathcal{E}_\mathcal{G}$ network to a common *Generic Codec Space* ($\mathbf{z}_\mathcal{G}$). For each identity $i$, a separate person-specific multi-layer perceptron network ($\mathbf{z}_i \leftarrow \mathcal{T}_i(\mathbf{z}_\mathcal{G})$) transforms the common latent parameters ($\mathbf{z}_\mathcal{G}$) to person-specific Codec parameters ($\mathbf{z}_i$). And then the corresponding pre-trained Codec decoder generates predicted geometry and texture by:

$$\mathbf{M}'_i, \mathbf{T}'_i \leftarrow \mathcal{D}_i(\mathbf{z}_i, \mathbf{v}) \tag{6.5}$$

Finally, *Generic HMC Encoder* network ($\mathcal{E}_\mathcal{G}$) and all person-specific transformation networks ($\mathcal{T}_i$) are trained by the cumulative loss for all identities:

$$\mathcal{L}_{\text{id}} = \sum_i^n (||\mathbf{M}_i - \mathbf{M}'_i||_2 + \lambda ||\mathbf{T}_i - \mathbf{T}'_i||) \tag{6.6}$$

## 6.3 Experiments

In order to evaluate the effectiveness of the proposed approaches, we have collected three new sessions of HMC captures for two subjects. Compared to the training set captures, subjects are placed in a different environment with different clothing and background and changed their facial styles, i.e. facial hair. The first capture (Subj.A, Cond.1) contains mainly perspective changes, the second one (Subj.A, Cond.2) includes especially extreme lighting conditions on mouth and some leakage to the eyes as well. The subject in the third capture (Subj.B) has a significantly different facial hair. Moreover, all of these captures are performed with a different background and clothing. Below we examine the performance of our approaches under these challenging unseen conditions.

In the experiments, we train 6 models under 6 different settings as listed below:

Table 6.1: Mean Absolute Errors of different settings for different parts of the face. The comparison is done on UV Texture and mesh geometry compared to Ground truth provided by [WSS+19] correspondence. Best accuracy of each line is typed in bold.

| | MAE | Baseline | Golden | 3D Aug. | 3D Aug.+ Golden | Multi-ID | 3D Aug.+ Multi-ID |
|---|---|---|---|---|---|---|---|
| Sbj. A, Cond. 1 | Face Texture | 4.544 | 3.975 | **3.377** | 3.380 | 3.468 | 3.743 |
| | Eye Texture | 8.549 | 5.489 | 4.318 | 4.418 | **4.314** | 5.107 |
| | Mouth Texture | 5.248 | 4.681 | **3.842** | 3.845 | 4.062 | 4.171 |
| | Face Geometry | 4.908 | 4.287 | **3.123** | 3.261 | 3.274 | 3.449 |
| | Eye Geometry | 0.676 | 0.486 | **0.376** | 0.384 | 0.414 | 0.446 |
| | Mouth Geometry | 1.741 | 1.664 | **1.063** | 1.164 | 1.129 | 1.095 |
| | *Avg. Improvement* | - | 12.59% | **31.03%** | 29.59% | 28.49% | 23.68% |
| Sbj. A, Cond. 2 | Face Texture | 6.16 | 6.108 | 5.568 | **5.472** | 5.532 | 5.766 |
| | Eye Texture | 7.705 | 7.481 | 5.525 | 5.525 | **5.030** | 5.802 |
| | Mouth Texture | 7.14 | 6.995 | 6.705 | **6.320** | 6.401 | 6.59 |
| | Face Geometry | 6.848 | 6.716 | 7.176 | 6.474 | **5.903** | 6.099 |
| | Eye Geometry | 0.636 | 0.517 | 0.606 | **0.524** | 0.529 | 0.583 |
| | Mouth Geometry | 2.993 | 2.859 | 3.349 | 2.956 | 2.586 | **2.459** |
| | *Avg. Improvement* | - | 1.39% | 2.41% | 8.32% | **12.00%** | 8.67% |
| Sbj. B, Cond. 1 | Face Texture | 3.026 | - | **2.497** | - | **2.497** | 2.724 |
| | Eye Texture | 6.219 | - | **6.005** | - | 6.385 | 5.634 |
| | Mouth Texture | 4.72 | - | 3.333 | - | **3.304** | 3.692 |
| | Face Geometry | 5.376 | - | 3.955 | - | **3.819** | 4.853 |
| | Eye Geometry | 0.479 | - | 0.505 | - | **0.487** | 0.619 |
| | Mouth Geometry | 2.068 | - | 1.234 | - | **1.207** | 1.43 |
| | *Avg. Improvement* | - | - | 21.96% | - | **23.22%** | 9.85% |
| | **Overall Average** | - | 6.99% | 18.46% | 18.95% | **21.24%** | 14.07% |

- Baseline
- Golden-plane correction
- 3D-Augmentation
- 3D-Augmentation & Golden-plane
- Multiple Identities
- 3D-Augmentation & Multiple Identities

Please note that, all the settings includes standard 2D Augmentation. The only difference is that in some settings, number of iterations are adjusted to the point where the models are converged. After training the models, we test them on the three test captures where the ground truth is provided by *Rosetta*.

As quantitative evaluation, we demonstrate *Mean Absolute Errors* (MAE) for texture and geometry of the full face. In order measure regional performance of the models, we additionally measure the errors for eyes and mouth regions. The detailed results and the improvements of the models are shown in Table 6.1. The results show that, compared to the baseline, Golden-plane rectification improves by $12.6\%$, 3D Augmentation improves by $36.0\%$ on geometry.

Although quantitative results gives a very good idea of the performances, qualitative observations are key to measure precise expression reconstruction as the ultimate goal of this project is to give real-like feeling of appearance and motion to *Codec Avatars*. Nevertheless, comparing only with ground truth by *Rosetta* might be misleading as we have observed that its reconstructions are slightly bit off from the actual expression in some instances. Figures 6.4, 6.5 and 6.6 demonstrate the qualitative results for our three test captures.

In both quantitative and qualitative results, we see a slight improvement by Golden-plane rectifier. Yet that improvement is surpassed by 3D augmentation with a significant margin. 3D Augmentation seems to be very successful for Subject A, Cond. 1 (in Fig.6.4 and Table6.1 top block) where the most difference to the training set is about perspective. That is reasonable as we would expect 3D Augmentation could only help for this aspect of the problem. For other test scenarios where there is also a huge illumination and facial hair changes, 3D Augmentation naturally fails.

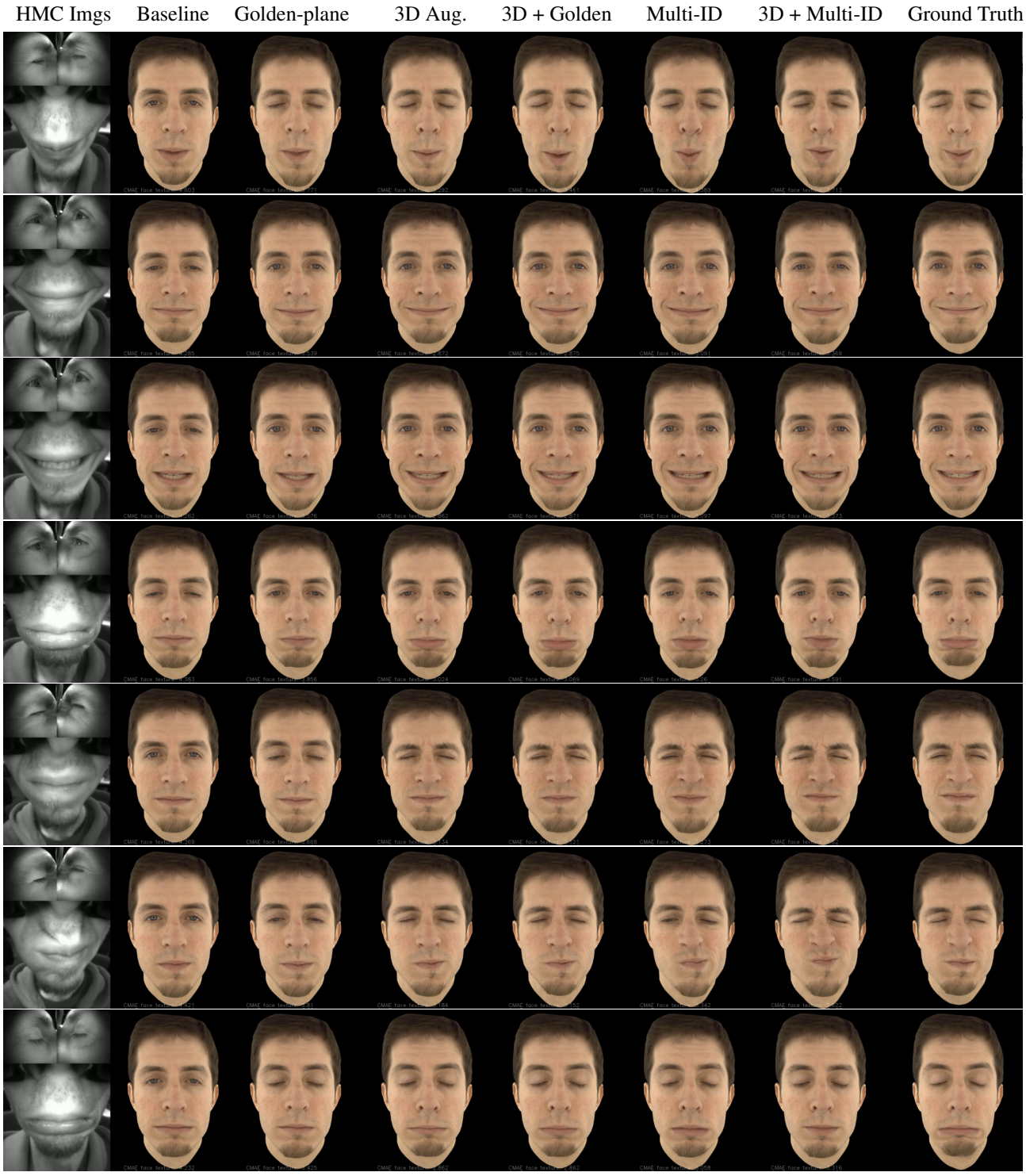



Figure 6.4: Qualitative results of the test capture of subject A in condition 1. Each row shows results of a different expression reconstruction of HMC images by the Codec Avatars. The experiment setting of each column is displayed at the top.

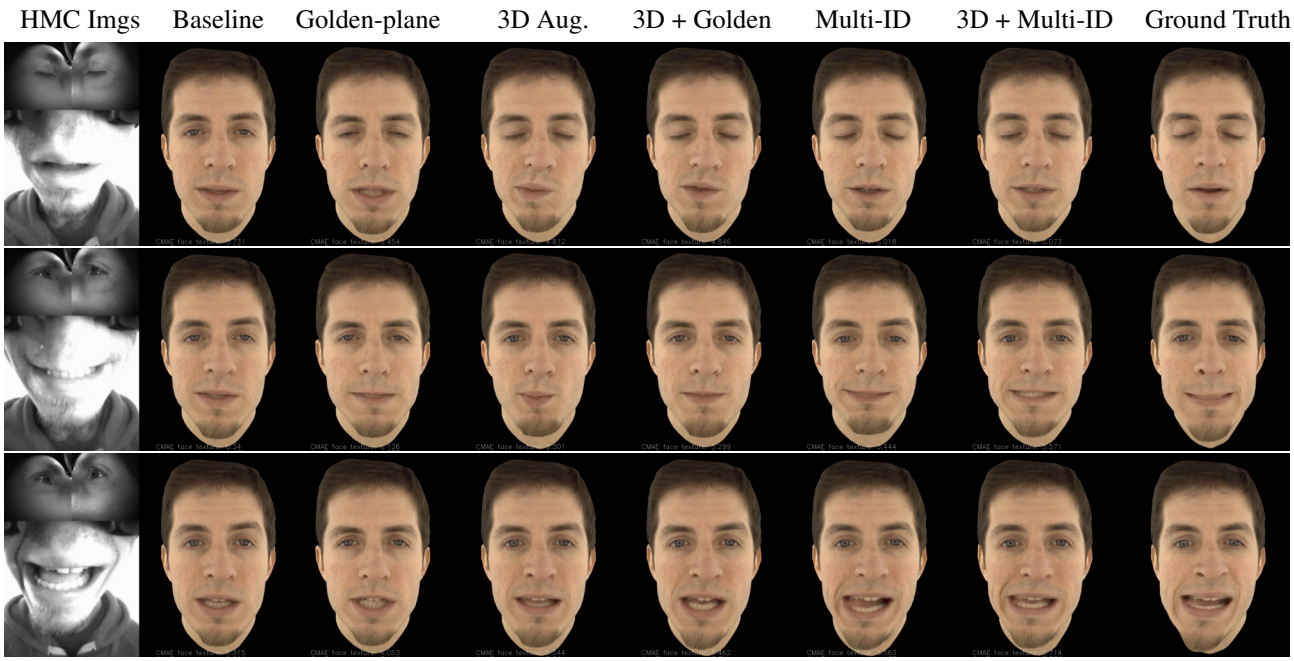



Figure 6.5: Qualitative results of the test capture of subject A in condition 2. Each row shows results of a different expression reconstruction of HMC images by the Codec Avatars. The experiment setting of each column is displayed at the top.

Identity Augmentation (Multi-Id) approach perform the best considering the all the three test cases. Since complete training set consist off variation of lighting, facial-hair/make-up, perspective, background etc., the resulting generic encoder becomes invariant to those variations in order to able to reconstruct each and every identity precisely. Nevertheless, 3D + Multi-Id performs slightly better than Multi-Id in Fig.6.5, even seems to be better than ground truth. When we look at HMC images rather than ground truth, we see ground truth is not so accurate after all. For example tongue is not placed towards left in Fig.6.5 3th row and the teeth is supposed to be visible in Fig.6.5 2nd row. In Fig. 6.6, 3D + Multi-Id results show less sleepy eyes than Multi-Id. The reason why 3D Augmentation is still improving qualitatively might be because that they training is limited with 20 identities due to memory restriction.

## 6.4 Conclusion

In this chapter, we presented a few approaches to make *Codec Avatar* animation by VR headset robust to different variation such as pose, lighting, facial hair, make-up, clothing, background etc.




Figure 6.6: Qualitative results of the test capture of subject B in condition 1. Each row shows results of a different expression reconstruction of HMC images by the Codec Avatars. The experiment setting of each column is displayed at the top.

First, we proposed a method to augment head-mounted camera images by first modeling them in 3D and apply random camera perturbations during training. So that, the trained model becomes less sensitive to different perspectives. Secondly, we proposed a generic expression encoder architecture and a novel approach to train it with multiple identities simultaneously. That way, we aimed to robustify personalized *Codec Avatars* by generalizing encoder network to all kinds of variations appears in the dataset of multiple subjects.

Results indicated that, 3D Augmentation offers a better solution for perspective invariance than Golden-plane correction by both quantitative and qualitative performance results. Moreover, since Golden-plane rectification special calibration for each hardware, it is also less convenient compared to 3D Augmentation approach which is generic. Generic Encoders trained with multiple identities also improves over the baselines with a significant margin thanks to the variation in multiple capture sessions. Moreover, we also concluded that employing 3D Augmentation and Generic Encoders together improves the performance of each of them, resulting in a satisfactory expression reconstruction in our benchmarks.

# Chapter 7

# Conclusion

## 7.1 Summary of Thesis Achievements

This thesis has explored important milestones for 3D face modeling and reconstruction to meet deep learning – by (1) revisiting optimization-based 3D face reconstruction with GAN-based non-linear texture model, (2) and deep identity features as energy function, (3) non-linear modeling of 3D face shape, texture, normals, and expression jointly for photorealistic face synthesis, (4) reboust expression and identity reconstruction.

Firstly, linear texture models are replaced by a GAN-based model trained on high-resolution UV maps as our statistical representation of the facial texture in **Chapter 3**. **Chapter 3** also propose to use deep face recognition features for 3D face reconstruction from a single image in an end-to-end reconstruction pipeline.

Secondly, **Chapters 4 and 5** presents two approaches for photorealistic face synthesis by 3DMMs. The former proposes to train a joint GAN as a 3D face model for shape, texture, normals, and expression and the latter proposes to sample from 3DMM and image-transfer to the realistic domain by an image-to-image GAN.

Finally, **Chapter 6** studies a few approaches to robustify *Codec Avatars* expression reconstruction from head-mounted camera inputs. The proposed approaches are based on augmenting training data

in 3D perspective and training for multiple identities together to improve generalization.

Many of the presented studies are published in peer-reviewed top conferences/journals, contain strong technical novelty and provide qualitative, and quantitative experiments in comparison with recent state-of-the-art approaches.

This thesis has presented a series of works that show deep generative and discriminative learning can improve the existing 3D facial modeling approaches. Particularly, it has shown significant steps towards building high-resolution and high-fidelity appearance and shape models by the recently emerged Generative Adversarial Networks. Moreover, it has also shown the strength and effectiveness of deep features for reconstruction. Therefore we can say that it accomplished an important milestone in the renovation of the existing linear models and low-level features for 3D face reconstruction and synthesis. In general, the thesis brought well-studied deep learning and 3D face modeling literature together to show the potentials of incorporating them.

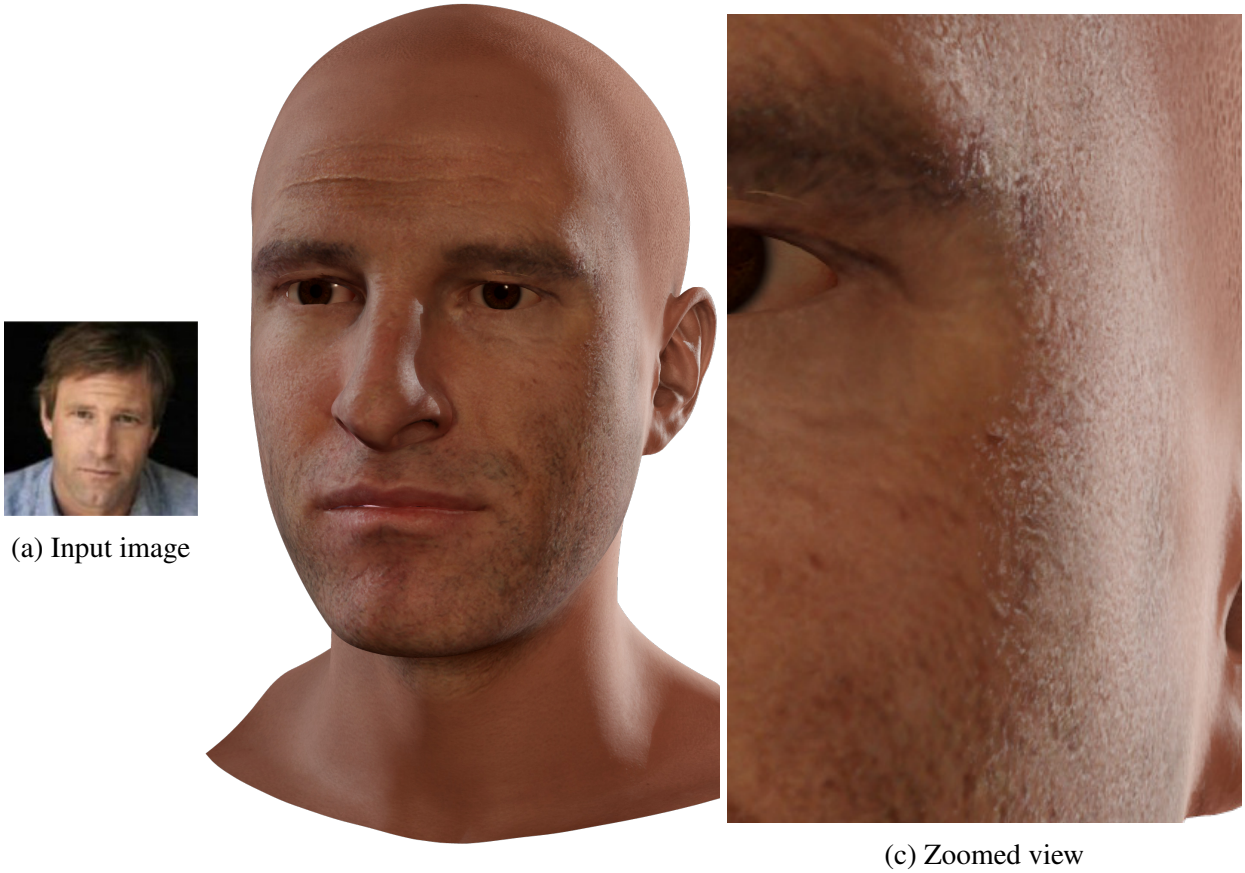
## 7.2 Limitations & Future Work

### 7.2.1 Higher Resolution and Full-head Completion

We have extended FastGANFit approach (**Chapter 3**) to enhance the reconstructed texture [LMG+20] and complete the rest of the head [PVOS+20] with other co-authors.

The data used to construct GANFit texture models contains baked illumination that appeared during data collection. But since the lighting throughout the collection process was the same, this illumination is consistent across the dataset. Therefore, [LMG+20] proposed to remove this baked illumination by simulating the lighting that appeared during the collection. This clean diffuse albedo is then enhanced by a super-resolution network and transferred into specular albedo, diffuse and specular normals with the help of additional light stage data.

The second study [PVOS+20] also utilizes an additional head model by cross-correlating with the face model in order to complete 3D full-heads. Also, it improves the reconstruction of eyes, ears, and



(a) Input image

(b) Reconstruction

(c) Zoomed view

Figure 7.1: 3D face reconstruction from a single in-the-wild image to high-resolution BRDF renderable full-head reconstruction. Initial reconstruction is done by [GPKZ19, GPKZ20] as shown in **Chapter 3**. Then the reconstruction is enhanced by [LMG+20] to higher resolution BRDF renderable texture and normals maps. Finally, its full-head is completed by [PVOS+20].

surroundings of the eyes by training part-specific models. As can be seen in Figure 7.1, the resulting reconstructions are very promising to bridge the uncanny valley.

### 7.2.2 Semi-personalized 3D Face Models

Now that, we have shown photorealistic 3D reconstruction from a single image and touched Codec Avatars which can generate realistically animatable personalized models, it is time to set the future directions. Monocular 3D reconstruction approaches are generic enough for any identity given a single image, yet the resulting reconstruction often looks like without soul. On the other hand, person-specific models are extremely successful to model motion and high range of expressions, but it requires collecting tens of thousands of images of target individuals in a light stage. One potential

future direction might be combining the best of both worlds by developing a semi-personalized model that effectively fine-tunes generic 3D face models by guided image collection of a target subject.

The main focus of such an approach would be to improve motion-realism of 3D avatars during an animation, driven either by camera input, speech, or text. In all cases, photorealistic and precise expression modeling is key. This means expression needs to be accurately received and well-reflected on shape, texture, and normals. Previously we attempted to generate expressive 3D faces by driving them with expression encoding from a CNN trained for expression recognition. Such a model can be personalized by the person-specific data collected by a smartphone camera with guidance on the screen to perform a particular set of expressions. Such an application would be very exciting to put into practice in modern smartphones.

# References


[AAB+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. 31

[AKV08] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, pages 1–6. IEEE, 2008. 36, 86

[ARV07] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 36, 89

[ASW93] Takaaki Akimoto, Yasuhito Suenaga, and Richard S. Wallace. Automatic creation of 3D facial models. *IEEE Computer Graphics and Applications*, 13(5):16–22, 1993. 86

[BAP+17] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3D face morphable models "In-the-Wild". In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

*2017*, volume 2017-January, pages 5464–5473, 2017. 25, 28, 30, 45, 46, 47, 49, 51, 52, 57, 60, 61, 78, 115, 116

[BBC11] Martin Breidt, Heinrich H. Biilthoff, and Cristobal Curio. Robust semantic analysis by synthesis of 3D facial motion. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pages 713–719. IEEE, 2011. 36, 86

[BCRC17] Ankan Bansal, Carlos Castillo, Rajeev Ranjan, and Rama Chellappa. The do's and don'ts for CNN-based face verification. *ICCVW*, 2017. 104

[BCW+18] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 40, 87

[BDBM11] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2D/3D hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11, pages 79–80, Scottsdale, Arizona, USA, December 2011. Association for Computing Machinery. 74

[BDS19] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GaN training for high fidelity natural image synthesis. *7th International Conference on Learning Representations, ICLR 2019*, 2019. 38, 39, 42, 84, 87

[BJC18] Shabab Bazrafkan, Hossein Javidnia, and Peter Corcoran. Face synthesis with landmark points from generative adversarial networks and inverse latent space mapping. *arXiv preprint arXiv:1802.00390*, 2018. 40, 87

[BM92] Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 14, pages 239–256, 1992. 78

[BRP+18] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3D morphable models. *International Journal of Computer Vision*, 126(2-4):233–254, 2018. 86

[BRV+18] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou. 3D reconstruction of "In-the-Wild" faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2638–2652, 2018. 28, 45, 46, 47, 51

[BRZ+16] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniahy, and David Dunaway. A 3D morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 5543–5552, 2016. 17, 20, 27, 35, 36, 37, 56, 71, 89, 90, 116

[BSBW14] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. Review of statistical shape spaces for 3D data with comparative analysis for human faces. *Computer Vision and Image Understanding*, 128:1–17, 2014. 36, 86

[BSBW17] Anil Bas, William A.P. Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10117 LNCS, pages 377–391, 2017. 28, 45, 46, 51, 61

[BSD+17] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:95–104, 2017. 22, 108, 117, 118

[BSM17] David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative adversarial networks. *IEEE Access*, 6:11342–11348, 2017. 21, 25, 38, 105, 111, 112, 115

[BTS+16] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016. 108

[BV99] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 25, 26, 36, 45, 46, 47, 51, 52, 56, 83, 86, 105, 110

[BV03] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *TPAMI*, 25(9):1063–1074, 2003. 25, 28, 60

[BW15] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 International Conference on Computer Vision, ICCV 2015, pages 3604–3612, 2015. 38, 39

[BWP13] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics*, 32(4):40, 2013. 26, 36, 86

[BWS+18] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 41

[CBK+17] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T. Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 3386–3395, 2017. 26, 57, 65, 113

[CBZ+19] Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. MeshGAN: Non-linear 3D morphable models of faces. *arXiv preprint arXiv:1903.10384*, 2019. 89

[CCZ+19] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International*

*Conference on Computer Vision*, volume 2019-October, pages 9428–9438, October 2019. 86

[CDH+16] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2180–2188, 2016. 109

[CET98] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1407(6):484–498, 1998. 87

[CGM+18] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, 2018. 105, 107

[CHZ14] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4), July 2014. 26

[CK17] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1520–1529, 2017. 107

[CKPZ18] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4DFAB: A large scale 4D database for facial expression analysis and biometric applications. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5117–5126, 2018. 25, 71

[CWW+16] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics*, 35(4), 2016. 41

[CWZ+14] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 25, 83, 116

[DBP+19] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2019. 105

[DCX+18] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2018. 58, 88, 100, 101

[DFI+15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 104

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4685–4694, 2019. 25, 59, 61, 62, 65, 69, 71, 98, 100

[DGZ+19] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage Dense Face Localisation in the Wild. *arXiv:1905.00641 [cs]*, May 2019. 25

[DLHT16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016. 104

[DMS98] Douglas DeCarlo, Dimitris Metaxas, and Matthew Stone. An anthropometric face model using variational techniques. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1998*, volume 98, pages 67–74, 1998. 38, 86

[DPSD17] Hang Dai, Nick Pears, William Smith, and Christian Duncan. A 3D morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 3104–3112, 2017. 36, 56

[DSVG11] Michael De Smet and Luc Van Gool. Optimal regions for linear model-based 3D face reconstruction. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6494 LNCS, pages 276–289, 2011. 89

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pages 886–893, 2005. 49

[DZCZ18] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou. Cascade multi-view hourglass model for robust 3D face alignment. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 399–403. IEEE, 2018. 63, 67, 69, 71

[DZZ17] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2017-July, pages 2006–2014, 2017. 99, 100

[EST+20] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D Morphable Face Models – Past, Present, and Future. *ACM Transactions on Graphics*, 39(5):157:1–157:38, June 2020. 25, 26, 51

[GAA+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 2017-December, pages 5768–5778, 2017. 38, 90

[GAM+18] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G. Shapiro, Donald L. Weaver, and Joann G. Elmore. Detection and classification of cancer in whole slide breast

histopathology images using deep convolutional networks. *Pattern Recognition*, 84:345–356, December 2018. 104

[GBK17] Baris Gecer, Vassileios Balntas, and Tae-Kyun Kim. Learning Deep Convolutional Embeddings for Face Representation Using Joint Sample- and Set-Based Supervision. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1665–1672, October 2017. 25, 100, 106, 114

[GBK+18] Baris Gecer, Binod Bhattarai, Josef Kittler, Tae-Kyun Kim, Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model. In *European Conference on Computer Vision (ECCV)*, volume 11215, pages 230–248. Springer International Publishing, 2018. 98

[GCM+18] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 19, 26, 45, 47, 50, 53, 57, 58, 64, 65, 68, 69, 71, 72, 73, 74, 77, 78, 79

[Gec16] Barıs Gecer. *Detection and Classification of Breast Cancer in Whole Slide Histopathology Images Using Deep Convolutional Networks*. PhD thesis, 2016. 104

[GGW+05] Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Fredrick Pighin. Making faces. In *ACM SIGGRAPH 2005 Courses, SIGGRAPH 2005*, page 10. ACM, 2005. 41

[GLP+20] Baris Gecer, Alexander Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, 2020. 25

[Goo16] Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 30, 38, 63, 105, 107

[Gow75] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 90

[GPKZ19] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, Long Beach, CA, USA, June 2019. IEEE. 23, 29, 139

[GPKZ20] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction. *Sumitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 23, 139

[GPM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 3, pages 2672–2680, 2014. 38, 39, 84, 87, 107

[GUA+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 107

[GVS+15] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt. VDub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Computer Graphics Forum*, 34(2):193–204, May 2015. 26

[GZC+16] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics*, 35(3):28, 2016. 26, 38, 57, 60

[GZC+19] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. CNN-Based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1294–1307, 2019. 26, 57

[HB19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2019. 108

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 104

[HMLL12] Gary B. Huang, Marwan A. Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems*, volume 1, pages 764–772, 2012. 78

[HRBL07] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 119

[HSW+17] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics*, 36(6):195, 2017. 26, 83

[HW77] Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977. 52

[HWY+18] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8398–8406, 2018. 88

[HYK+17] Guosheng Hu, Fei Yan, Josef Kittler, William Christmas, Chi Ho Chan, Zhenhua Feng, and Patrik Huber. Efficient 3D morphable face model fitting. *Pattern Recognition*, 67:366–379, 2017. 51, 61

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Com-*

*puter Vision and Pattern Recognition*, volume 2016-December, pages 770–778, 2016. 100

[IZZE17] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:5967–5976, 2017. 40, 106, 107, 115

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9906 LNCS, pages 694–711. Springer, 2016. 115

[JCZ19] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11-12):1767–1779, 2019. 40, 87

[JMLH01] Henrik Wann Jensen, Stephen R Marschner, Marc Levoy, and Pat Hanrahan. A practical model for subsurface light transport. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 511–518, 2001. 45

[JZD+18] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3D Face Reconstruction With Geometry Details From a Single Image. *IEEE Transactions on Image Processing*, 27(10):4756–4770, October 2018. 25, 56

[KALL18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018. 17, 25, 30, 38, 39, 63, 71, 84, 87, 90, 105, 109

[KB15] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. 71, 116

[KKT+15] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, pages 1931–1939, 2015. 119

[KLA19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4396–4405, 2019. 25, 28, 38, 39, 42, 84, 87, 109, 110

[KLA+20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]*, March 2020. 25, 28, 38, 39, 109

[KRZ20] Mohammad Rami Koujan, Anastasios Roussos, and Stefanos Zafeiriou. DeepFaceFlow: In-the-Wild Dense 3D Facial Motion Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6618–6627, 2020. 25

[KTC+19] Minsuk Kahng, Nikhil Thorat, Duen Horng Chau, Fernanda B. Viégas, and Martin Wattenberg. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):310–320, January 2019. 17, 39

[KW14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014. 38, 40, 87

[KZT+18] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4625–4634, 2018. 53

[LBB+17] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194, 2017. 36, 86

[LBBM18] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1886–1895, 2018. 89

[LBK17] Ming Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in Neural Information Processing Systems*, 2017-December:701–709, 2017. 108

[LDW+14] Kai Li, Qionghai Dai, Ruiping Wang, Yebin Liu, Feng Xu, and Jue Wang. A data-driven approach for facial expression retargeting in video. In *IEEE Transactions on Multimedia*, volume 16, pages 299–310. IEEE, 2014. 40, 88

[LKA+17] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–10, 2017. 52

[LLCL19] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 48

[LMAR14] JP Lewis, Zhenyao Mo, Ken Anjyo, and Taehyun Rhee. Probable and improbable faces. In *Mathematical Progress in Expressive Image Synthesis I*, pages 21–30. Springer, 2014. 50

[LMG+20] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically Renderable 3D Facial Reconstruction "in-the-wild". In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 760–769, 2020. 23, 25, 85, 138, 139

[Low99] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, September 1999. 49, 52

[LPG17] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:853–862, 2017. 107

[LSEJ18] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation Letters*, 3(1):387–394, 2018. 108

[LSSS18] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 37(4):68, 2018. 26, 29, 41, 58, 123, 124, 125

[LTT17] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional CycleGAN for attribute guided face image generation. *arXiv*, pages 1–9, 2017. 109

[LWP10] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010*, 29(4):32, 2010. 36, 86

[LWY+17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 104

[Mar19] Marmoset LLC. Marmoset toolbag, 2019. 92

[MMS08] Iordanis Mpiperis, Sotiris Malassiotis, and Michael G Strintzis. Bilinear models for 3-D face and facial expression recognition. *IEEE Transactions on Information Forensics and Security*, 3(3):498–511, 2008. 38, 39

[MO14] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]*, November 2014. 40

[MPK09] Umar Mohammed, Simon J.D. Prince, and Jan Kautz. Visio-lization: Generating novel facial images. *ACM Transactions on Graphics*, 28(3):57, 2009. 38, 87

[MPN+19] Stylianos Moschoglou, Stylianos Ploumpis, Mihalis Nicolaou, Athanasios Papaioannou, and Stefanos Zafeiriou. 3DFaceGAN: Adversarial nets for 3D face representation, generation, and translation. *arXiv preprint arXiv:1905.00307*, 2019. 40, 85, 87

[MPS+17] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2017-July, pages 1997–2005, 2017. 99

[MTH+16] Iacopo Masi, Anh Tuân Trân, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9909 LNCS, pages 579–596. Springer, 2016. 88, 108, 119, 120, 121

[NHSW20] J Naruniec, L Helminger, C Schroers, and RM Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184. Wiley Online Library, 2020. 123

[NKH19] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 25

[NRH+92] Fred E Nicodemus, Joseph C Richmond, Jack J Hsia, IW Ginsberg, and T Limperis. Geometrical considerations and nomenclature for reflectance. *NBS monograph*, 160:4, 1992. 45

[NSX+19] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. paGAN: Real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37(6):258, 2019. 41

[OLSL16] Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016. 52

[OOS17] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 2642–2651, Sydney, NSW, Australia, August 2017. JMLR.org. 91

[PAM+18] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11214 LNCS, pages 835–851, 2018. 40, 87

[PB81] Stephen M. Platt and Norman I. Badler. Animating facial expressions. In *Proceedings of the 8th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1981*, volume 15, pages 245–252. ACM, 1981. 86

[PB16] Marcel Piotraschke and Volker Blanz. Automated 3D face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 3418–3427, 2016. 30, 52, 61

[PCKO15] Sung Woo Park, Jong Woo Choi, Kyung S. Koh, and Tae Suk Oh. Mirror-Imaged Rapid Prototype Skull Model and Pre-Molded Synthetic Scaffold to Achieve Optimal Orbital Cavity Reconstruction. *Journal of Oral and Maxillofacial Surgery*, 73(8):1540–1553, August 2015. 26

[PGLC15] Vishal M. Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 105

[PKA+09] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, September 2009. 36, 56, 86, 116

[PS09] Ankur Patel and William A.P. Smith. 3D morphable face models revisited. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, volume 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1327–1334. IEEE, 2009. 36, 86

[PS12] Ankur Patel and William AP Smith. Driving 3D morphable models using shading cues. *Pattern Recognition*, 45(5):1993–2004, 2012. 50

[PS16] Ankur Patel and William AP Smith. Manifold-based constraints for operations in face space. *Pattern recognition*, 52:206–217, 2016. 50

[PVOS+20] Stylianos Ploumpis, Evangelos Ververas, Eimear O' Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William A. P. Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3D morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2020. 23, 96, 138, 139

[PVZ15] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, pages 41.1–41.12, 2015. 19, 25, 62, 78, 79, 104, 116, 118, 119, 120

[PWP+19] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A.P. Smith, and Stefanos Zafeiriou. Combining 3D morphable models: A large scale face-and-head model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 10926–10935, 2019. 96

[RAY+16] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *33rd International Conference on Machine Learning, ICML 2016*, 3:1681–1690, 2016. 21, 40, 108, 112

[RBSB18] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. *Lecture Notes in Computer Science (in-*

*cluding subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11207 LNCS:725–741, 2018. 25, 41, 85, 87, 89

[RBV02] Sami Romdhani, Volker Blanz, and Thomas Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 2353, pages 3–19, 2002. 60

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 119

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 104

[RMC16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016. 38, 87, 105, 107, 112

[RPDB16] Oren Rippel, Manohar Paluri, Piotr Dollar, and Lubomir Bourdev. Metric learning with adaptive density discrimination. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016. 109, 114

[RSK16] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 460–467, 2016. 25, 26, 53, 57, 104, 108

[RSOK17] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 5553–5562, 2017. 26, 50, 53, 57

[RV05] Sami Romdhani and Thomas Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume II, pages 986–993, 2005. 28, 30, 45, 46, 51, 52, 56, 61

[SCC+16] Soumyadip Sengupta, Jun Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016. 99

[SEMV17] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, 123(2):160–183, 2017. 52

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, pages 815–823, 2015. 25, 62, 104, 115, 118, 120

[SLY+18] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2018. 88

[SPH20] Gwen R.J. Swennen, Lies Pottel, and Piet E. Haers. Custom-made 3D-printed face masks in case of pandemic crisis situations with a lack of commercially available FFP2/3 masks. *International Journal of Oral and Maxillofacial Surgery*, 49(5):673–677, May 2020. 26

[SPT+17] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:2242–2251, 2017. 105, 108, 111, 120

[SRK17] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 1585–1594, 2017. 25, 26, 40, 53, 57, 84

[SS15] Baochen Sun and Kate Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, 2015. 107

[SSK19a] Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Transactions on Multimedia Computing Communications and Applications*, 15(3s), October 2019. 25

[SSK19b] Ron Slossberg, Gil Shamai, and Ron Kimmel. High quality facial surface and texture synthesis via generative adversarial networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11131 LNCS:498–513, 2019. 25, 29, 40, 58, 85, 87

[STZP13a] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 115, 116

[STZP13b] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, 2013. 53, 115, 116

[SWH+17] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 2326–2335, 2017. 19, 57, 73, 75

[SWL16] Leon Sixt, Benjamin Wild, and Tim Landgraf. Rendergan: Generating realistic labeled data. *arXiv preprint arXiv:1611.01331*, 2016. 105, 108

[SYH+17] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 5444–5453, 2017. 19, 74, 113

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, abs/1409.1556, 2015. 119

[TBG+19] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, and Christian Theobalt. FML: Face Model Learning From Videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10804–10814, Long Beach, CA, USA, June 2019. IEEE. 45

[TEB+20] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. *arXiv:2004.00121 [cs]*, March 2020. 109

[THDS15] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 107

[THM+18] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gerard Medioni. Extreme 3D Face Reconstruction: Seeing Through Occlusions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, Salt Lake City, UT, June 2018. IEEE. 25, 86

[THMM17] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 1493–1502, 2017. 19, 26, 53, 57, 58, 68, 71, 72, 73, 74, 78, 79, 113

[THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:2962–2971, 2017. 107

[TL18] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018. 26, 38, 45, 57, 58, 60, 71, 72, 73

[TL19] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 45

[TMH18] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. Generating photo-realistic training data to improve face recognition accuracy. *arXiv preprint arXiv:1811.00112*, 2018. 88

[TYL17] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 1283–1292, 2017. 106, 109, 113

[TYL19] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3007–3021, 2019. 88

[TZB+18] Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 50

[TZG+18] Ayush Tewari, Michael Zollhofer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. *Proceedings of the IEEE Computer Society*

*Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 26, 38, 45, 57, 58, 60, 71, 72, 73

[TZK+17] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, volume 2018-January, pages 1274–1283, 2017. 25, 26, 28, 45, 53, 57, 58, 71, 72, 73, 113

[TZN+15] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics*, 34(6):181–183, 2015. 25, 36, 86

[TZS+19] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Communications of the ACM*, volume 62, pages 96–104, 2019. 25, 26, 51, 57

[VA19] Raviteja Vemulapalli and Aseem Agarwala. A Compact Embedding for Facial Expression Similarity. *arXiv:1811.11283 [cs]*, January 2019. 25

[VK20] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. *arXiv:2007.03898 [cs, stat]*, July 2020. 38, 40, 41

[VRM+17] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:4627–4635, 2017. 104, 108

[VT05] M Alex O Vasilescu and Demetri Terzopoulos. Multilinear independent components analysis. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 547–553. IEEE, 2005. 38

[vVk17] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6306–6315. Curran Associates, Inc., 2017. 41

[WBM+16] Erroll Wood, Tadas Baltrušaitis, Louis Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Eye Tracking Research and Applications Symposium (ETRA)*, volume 14, pages 131–138. ACM, 2016. 104, 108

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. 36, 60

[WFJ+17] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1211–1220, Venice, October 2017. IEEE. 19, 74, 77

[WSS+19] Shih En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. VR facial animation via multiview image translation. *ACM Transactions on Graphics*, 38(4):67, 2019. 15, 22, 26, 29, 32, 41, 123, 124, 125, 126, 128, 131

[WTP17] Lior Wolf, Yaniv Taigman, and Adam Polyak. Unsupervised creation of parameterized avatars. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1539–1547, 2017. 107

[WZLQ16] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9911 LNCS, pages 499–515. Springer, 2016. 106, 109, 114

[WZX+16] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016. 40

[XLZ+16] Chao Xiong, Luoqi Liu, Xiaowei Zhao, Shuicheng Yan, and Tae-Kyun Kim. Convolutional fusion network for face verification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):517–528, 2016. 104

[XZT+15] Chao Xiong, Xiaowei Zhao, Danhang Tang, Karlekar Jayashree, Shuicheng Yan, and Tae-Kyun Kim. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 2015. 104

[YDXZ20] Yuxin Yao, Bailin Deng, Weiwei Xu, and Juyong Zhang. Quasi-Newton Solver for Robust Non-Rigid Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7600–7609, 2020. 36

[YLLL14] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 99, 115, 118

[YMW+11] Fei Yang, Dimitri Metaxas, Jue Wang, Eli Shechtman, and Lubomir Bourdev. Expression flow for 3D-Aware face component transfer. *ACM Transactions on Graphics*, 30(4):1–10, 2011. 36, 86

[YSN+18] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics*, 37(4):162, 2018. 19, 57, 74

[YYS+17a] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 4010–4019, 2017. 53, 88, 106, 107, 109, 113, 120, 121

[YYS+17b] Shanxin Yuan, Qi Ye, Björn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *CVPR*, 2017. 104

[ZCR+17] Stefanos Zafeiriou, Grigorios G. Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3D Menpo Facial Landmark Tracking

Challenge. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2503–2511, Venice, Italy, October 2017. IEEE. 53

[ZDKZ19] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3D face decoding over 2500FPS: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 1097–1106, 2019. 41

[ZLGS03] Qingshan Zhang, Zicheng Liu, Baining Guo, and Harry Shum. Geometry-driven photorealistic facial expression synthesis. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2003*, 12(1):48–60, 2003. 40, 87

[ZLL+16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 146–155, 2016. 53, 56, 98, 115, 116

[ZPIE17] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 2242–2251, 2017. 21, 22, 105, 106, 108, 109, 110, 111, 116, 117, 118, 120

[ZSFB15] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, pages 4511–4520, June 2015. 104, 108

[ZTG+18] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum*, 37(2):523–550, May 2018. 51

[ZWC+20] Wenbin Zhu, HsiangTao Wu, Zeyu Chen, Noranart Vesdapunt, and Baoyuan Wang. ReDA:Reinforced Differentiable Attribute for 3D Face Reconstruction. In *Proceed-*

*ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4958–4967, 2020. 48

[ZXJ+17] Jian Zhao, Lin Xiong, Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Sugiri Pranata, Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems*, volume 2017-December, pages 66–76, 2017. 88

[ZXL+17] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 107

[ZXL+19] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019. 38

[ZZLQ16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 25, 116

[ZZY17] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *ICCV*, 2017. 108