SELF-SUPERVISED RECONSTRUCTION AND SYNTHESIS OF FACES

by AYUSH TEWARI

Dissertation zur Erlangung des Grades
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

Saarbrücken, 2021

Date of Colloquium:	July 26, 2021
Dean of the Faculty:	Prof. Dr. Thomas Schuster
Chair of the Committee:	Prof. Dr. Philipp Slusallek
Reviewers:	Prof. Dr. Christian Theobalt
	Dr. Michael Zollhöfer
	Prof. Dr. Peter Wonka
Academic Assistant:	Dr. Lingjie Liu

ABSTRACT

Photorealistic and semantically controllable digital models of human faces are important for a wide range of applications such as movies, virtual reality, and casual photography. Traditional approaches require expensive setups which capture the person from multiple cameras under different illumination conditions. Recent approaches have also explored digitizing faces under less constrained settings, even from a single image of the person. These approaches rely on priors, commonly known as 3D morphable models (3DMMs), which are learned from datasets of 3D scans. This thesis pushes the state of the art in high-quality 3D reconstruction of faces from monocular images. A model-based face autoencoder architecture is introduced which integrates convolutional neural networks, 3DMMs, and differentiable rendering for self-supervised training on large image datasets. This architecture is extended to enable the refinement of a pretrained 3DMM just from a dataset of monocular images, allowing for higher-quality reconstructions. In addition, this thesis demonstrates the learning of the identity components of a 3DMM directly from videos without using any 3D data. Since videos are more readily available, this model can generalize better compared to the models learned from limited 3D scans.

This thesis also presents methods for the photorealistic editing of portrait images. In contrast to traditional approaches, the presented methods do not rely on any supervised training. Self-supervised editing is achieved by integrating the semantically meaningful 3DMM-based monocular reconstructions with a pretrained and fixed generative adversarial network.

While this thesis presents several ideas which enable self-supervised learning for the reconstruction and synthesis of faces, several open challenges remain. These challenges, as well as an outlook for future work are also discussed.

ZUSAMMENFASSUNG

Fotorealistische und semantisch steuerbare digitale Modelle von menschlichen Gesichtern sind wichtig für eine Vielzahl von Anwendungen wie Filme, virtuelle Realität und Gelegenheitsfotografie. Traditionelle Ansätze erfordern teure Setups, die die Person mit mehreren Kameras unter verschiedenen Beleuchtungsbedingungen aufnehmen. Neuere Ansätze haben auch die Digitalisierung von Gesichtern unter weniger strengen Bedingungen untersucht, selbst von einem einzigen Bild der Person. Diese Ansätze stützen sich auf Vorannahmen, sogenannte *3D morphable models* (3DMMs), die aus einer Reihe von 3D-Scans gelernt werden. Diese Dissertation bringt den Stand der Forschung auf dem Gebiet der hochwertigen 3D-Rekonstruktion von Gesichtern aus Einzelaufnahmen voran. Es wird eine modellbasierte Gesichts-Autoencoder-Architektur entwickelt, die neuronale Netze, 3DMMs und differenzierbares Rendern für selbstüberwachtes Training auf großen Bilddatensätzen verbindet. Diese Architektur wird erweitert, um die Verfeinerung eines vortrainierten 3DMMs lediglich anhand eines Datensatzes von monokularen Bildern zu ermöglichen, wodurch qualitativ hochwertigere Rekonstruktionen erzielt werden können. Darüber hinaus demonstriert diese Dissertation das Lernen der Identitätskomponenten eines 3DMM anhand von Videos ohne den Einsatz von 3D-Daten. Da Videos leichter verfügbar sind, kann dieses Modell im Vergleich zu jenen Modellen, die aus begrenzten 3D-Scans gelernt wurden, besser generalisieren.

In dieser Dissertation werden auch Methoden für die fotorealistische Bearbeitung von Porträtbildern vorgestellt. Im Gegensatz zu traditionellen Ansätzen sind die vorgestellten Methoden nicht auf ein überwachtes Training angewiesen. Die selbstüberwachte Bearbeitung wird durch die Verknüpfung der semantisch aussagekräftigen 3DMM-basierten Einzelbildrekonstruktionen mit einem vortrainierten und unveränderlichen generativen adversariellen Netzwerk erreicht.

Während diese Dissertation mehrere Ideen entwickelt, die selbstüberwachtes Lernen für die Rekonstruktion und Synthetisierung von Gesichtern ermöglichen, verbleiben mehrere ungelöste Herausforderungen. Diese Herausforderungen, sowie ein Ausblick auf mögliche zukünftige Forschungsarbeiten werden ebenfalls erörtert.

ACKNOWLEDGMENTS

I would like to thank my advisor Christian Theobalt for his guidance and advice throughout this journey. Christian always tried to provide the perfect research environment with a lot of support and freedom. I have always found his passion for research very inspiring, and I am very glad I could spend so much time in his group. I appreciate Peter Wonka for serving on the thesis committee. His work on Image2StyleGAN (Abdal et al., 2019) was one of the reasons I got interested in exploring StyleGAN for building photorealistic 3D face models.

The work in this thesis was made possible with the help and support of many people. I worked very closely with Michael Zollhöfer, especially in the early projects. Michael is a very passionate researcher and an amazing mentor. I fondly remember our workout sessions at the university gym, where we brainstormed research ideas. The idea behind Chapter 5 came up in these discussions. Pablo Garrido helped a lot during my initial months at MPI, helping me get familiar with existing 3D face reconstruction methods. I also had the pleasure of working closely with Patrick Pérez. Patrick had great suggestions whenever I was stuck in any project. He was also the person behind the names "MoFA" and "StyleRig". Florian Bernard came to my help whenever I struggled with the mathematical side of things. Mohamed Elgharib and Gaurav Bharaj helped with paper writing and evaluations, and Mallikarjun B R helped me with some initial experiments for the work in Chapter 8, even though it required working during the Christmas break.

I had the great opportunity to visit Stanford University and work with Ohad Fried and Maneesh Agrawala. It was a very enjoyable experience and has led to continuing collaborations. I am also thankful to the many other researchers I closely collaborated with and learned from, especially Edgar Tretschk, Hyeongwoo Kim, Qianru Sun, Vladislav Golyanik, Marc Habermann, Gereon Fox, Justus Thies, and Matthias Niessner. I had the pleasure of supervising several Masters students: Hossein Hajipour, Chitra Singh, Tarun Yenamandra, Tianqi Fan, and Linjie Lyu. This was one of the highlights for me and allowed me to explore a wide range of interesting problems.

Much credit goes to the administrative assistants Sabine Budde and Ellen Fries, and the systems administrators Helge Rhodin, Hyeongwoo Kim, Jozef Hladky, and Gereon Fox. I would also like to thank the nice people who proofread the initial draft of this document: Snehaa Seal, Jiayi Wang, Yue Jiang, Franziska Müller, Edgar Tretschk, Lingjie Liu, Michael Zollhöfer, and Dushyant Mehta.

I am deeply indebted to all D4 and D6 members who acted as my family away from home. Finally, I would like to thank my family and friends for their unconditional love and support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Digitizing human faces has a wide range of applications in movies, video games, and virtual reality. Faces of real actors are used to reenact virtual characters in movies and games. Digital avatars of actors have been used in movies, even after they are deceased! (Alexander, 2019; Itzkoff, 2016) Virtual reality applications allow us to interact with a person virtually from our viewpoint which changes as we move around. Such applications require *reconstructing* the face of a person, including detailed 3D geometry and reflectance properties, which describe the appearance of the face under different lighting conditions and viewpoints. For many of these applications, this reconstruction should be controllable by artists or users. For example, reanimating a face requires preserving the identity-specific geometry and only editing of the expressions. After reconstruction and editing, the digitized face can be rendered to create new imagery suitable for the application, see Figure 1.1.

Recovering detailed 3D properties of a person's face traditionally requires expensive capture setups with multiple cameras and lights (Alexander et al., 2009; Beeler et al., 2010). Additionally, manual interventions by artists are required for building controllable rigs suitable for editing (Lewis et al., 2014b). This pipeline is cumbersome, expensive, and not suitable for casual users. Thus, a lot of recent work has explored less-constrained settings, such as reconstruction of 3D faces from just a monocular image or video (Egger et al., 2020; Zollhöfer et al., 2018). However, there is still a large gap in quality, which prevents their usage in many applications which rely on photorealistic rendering. This thesis proposes several methods which improve the state of the art in high-quality reconstruction and controllable synthesis of faces. All methods presented in this thesis rely only on very few observations of a face in *in-the-wild* images, without constraints over the lighting and camera poses in the scenes.

1.1 MONOCULAR RECONSTRUCTION

Monocular 3D reconstruction, i.e., 3D reconstruction of geometry and appearance from a single image is an ill-posed problem due to depth ambiguities, as well as reflectance-shading ambiguities. There are infinitely many 3D points along the camera ray which can project perfectly to a 2D point in an image. This leads to depth ambiguity, as multi-view triangulation cannot be used to obtain the correct face shape in the monocular setting. In addition, the observed colors in the image can be explained by infinitely many combinations of reflectance and illumination (Egger, 2017). The common strategy for constraining the monocular face reconstruction problem is to first capture a training dataset of 3D scans in order to learn a low dimensional representation of 3D faces, referred to as 3D Morphable Models (3DMMs). Constraining the solution to lie within the space of these models allows for plausible monocular 3D reconstruction of the geometry and reflectance of the face. In addition, these models capture the significant modes of deformation along *semantically disentangled components* such as identity and

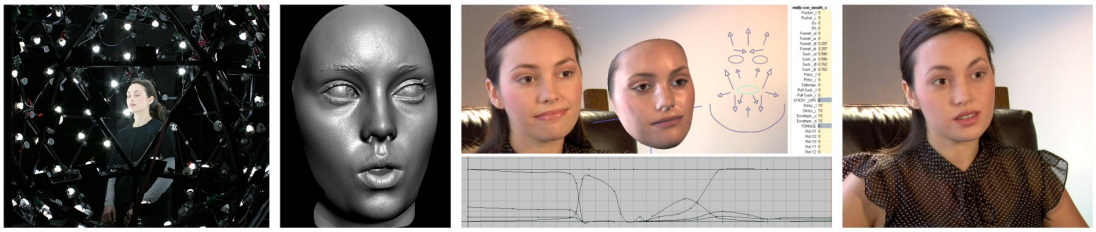

Figure 1.1: Traditional pipeline for creating a digital 3D face model. The person is first captured using multiple cameras and light sources in different expressions. This data is processed to compute a high-quality reconstuction of the geometry and reflectance of the face. This digital face is then rigged such that it can be animated. Finally, the animated digital face is rendered to create the final result. This process leads to very high quality output, but is not suitable for casual users because of the capture setup, and manual effort required in the different steps. Figure taken from Alexander et al. (2009).

expression geometry, and diffuse reflectance. For many applications, it is important for the reconstruction to be parameterized in terms of these semantic meaningful components. For example, facial reenactment primarily depends on the expression component, while facial recognition mainly depends on the identity component.

While the use of 3D Morphable Models has seen wide success with many different monocular reconstruction methods in the literature (Zollhöfer et al., 2018), several limitations exist. Since 3D scanners are not ubiquitous, most models are learned from a low quantity of scans. In addition, it is even difficult for commercial scanners to reconstruct some features such as hair or beards. This limits the quality of generalization of these models, and thus also of the monocular 3D reconstruction methods which use these models. The goal of this thesis is to move towards learning a face morphable model from just images and videos, using as little 3D data as possible. Since monocular images and videos are widely available online, this would allow for learning models which generalize better in-the-wild. One of the main technical contributions of this thesis is the integration of analytically-differentiable 3D rendering and morphable models with convolutional neural networks. This allows for self-supervised learning-based methods without the need for 3D supervision. This new architecture, called model-based face autoencoder (MoFA) is a key component required for learning morphable models from images and videos. The integration of a morphable model in the learning loop also allows for learning the model itself from in-the-wild data. This leads to models which better generalize, compared to the models learned from limited 3D data.

1.2 CONTROLLABLE SYNTHESIS

While rendering a perfect 3D reconstruction of the face should lead to photorealistic results in theory, in practice, this is a very challenging task, see Figure 1.2. This is primarily because most monocular reconstruction methods make several (incorrect) assumptions about the image formation process. The appearance of a human face is the result of complex interactions between the incoming light and the skin, which can be described using the rendering equation (Kajiya, 1986). This leads to several complex effects such as subsurface scattering and specularities. Inverting the rendering equation is infeasible, especially in the monocular setting with very limited constraints. Thus,

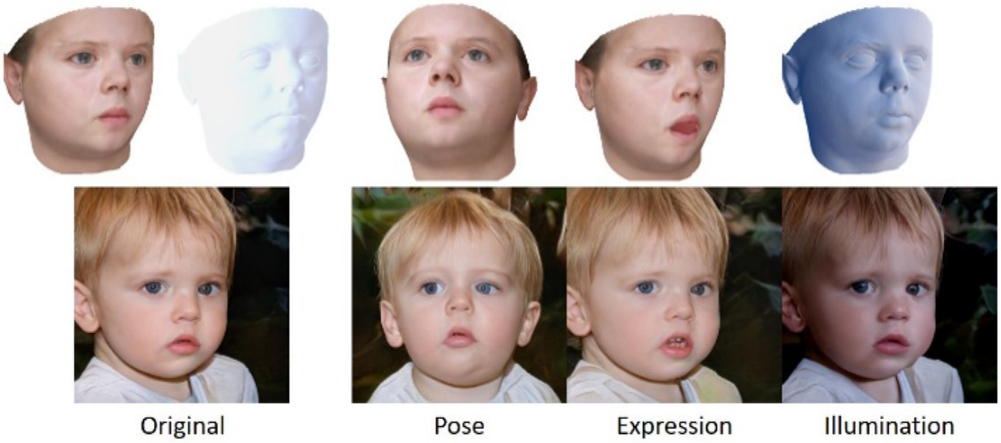

Figure 1.2: Top row shows visualizations of the monocular reconstructions corresponding to the images in the bottom row. 3DMM reconstructions are not photorealistic due to the approximations made in the image formation process. StyleRig (Tewari et al., 2020b) allows for control over the head pose, facial expressions, and scene illumination in a portrait image synthesized by StyleGAN (Karras et al., 2019a), by integrating 3DMM-based reconstruction with the latent space of StyleGAN.

most methods work with a simplified version of the equation, only accounting for the diffuse effects, ignoring specularities and other higher-order effects. In addition, constraining the reconstructions to lie within the low-dimensional 3D Morphable Model space of most existing models does not allow for the reconstruction of the high-frequency details (Figure 1.2). Due to these limitations, rendering monocular 3D reconstructions does not lead to photorealistic results. Finally, most 3D reconstuction methods are limited to reconstructing the frontal face region, ignoring areas such as hair, neck and ears. Synthesizing complete portrait images, including hair, ears, neck, and even upper torso is important for many synthesis applications.

Neural rendering (Tewari et al., 2020c) is an emerging field which provides ways to synthesize photorealistic images and videos from lower-quality reconstructions. The idea is to learn a rendering function, parameterized using a neural network. Semantically meaningful parameters are provided as input to this network. For synthesis of portrait images, these could be the parameters of the 3D Morphable Model, in addition to the scene parameters such as the viewpoint and scene illumination. Neural rendering allows for learning the details missing from the 3D reconstructions directly using the neural network filters, without explicitly modeling the rendering equation. Several portrait editing methods with promising results have been demonstrated in the literature (Kim et al., 2018a; Tewari et al., 2020c; Thies et al., 2019). Such approaches require training data of images/videos of people observed under different poses, expressions and lighting conditions. Since such large-scale datasets at high quality are not readily available, most methods limit themselves to be person- and scene-specific, only being able to synthesize portrait images of a single person in a single environment.

Another goal of this thesis is to provide solutions for neural rendering with fewer restrictions on the training data. A portrait editing method is developed where the training dataset only includes a single image per-identity. This is realized by integrating semantically meaningful 3DMM-based monocular reconstructions with

a pretrained and fixed generative adversarial network (Figure 1.2). This also allows for a generalizable neural rendering method, where the identity at test time does not need to be present in the training dataset.

1.3 STRUCTURE AND CONTRIBUTIONS

This thesis is divided into eight chapters. Chapters 4, 5, 6, 7 and 8 include the main technical contributions and their evaluations.

- Chapter 2 introduces relevant technical background, such as the image formation details used in the later chapters.
- Chapter 3 discusses the relevant previous work.
- Chapter 4 (published as Tewari et al. (2020d, 2017)) introduces a new type of architecture for monocular 3D face reconstruction called model-based face autoencoder (MoFA). This architecture joins forces of state-of-the-art CNN-based regression approaches and 3D morphable models via a deep integration of the two on an architectural level. Unlike previously used CNN-based decoders, the proposed convolutional autoencoder deeply integrates an expert-designed decoder. This decoder layer implements a new generative analytically-differentiable image formation model on the basis of a detailed parametric 3D face model. Model-based autoencoders are trained on large 2D datasets without requiring any corresponding 3D supervision.
- Chapter 5 (published as Tewari et al. (2018)) presents an approach that jointly learns (i) a regressor for face shape, expression, reflectance and illumination on the basis of (ii) a concurrently learned parametric face model. The multi-level face model combines the advantage of 3D Morphable Models for regularization with the out-of-space generalization of a learned corrective space. End-to-end training is performed on in-the-wild images without dense ground truth annotations.
- Chapter 6 (published as Tewari et al. (2019)) proposes multi-frame video-based self-supervised training of a deep network that (i) learns a face identity model both in shape and appearance while (ii) jointly learning to reconstruct 3D faces. The face model is learned using only corpora of in-the-wild video clips collected from the internet. This virtually endless source of training data enables learning of a highly general 3D face model. In order to achieve this, a novel multi-frame consistency loss is proposed that ensures consistent shape and appearance across multiple frames of a subject's face, thus minimizing depth ambiguity. At test time an arbitrary number of frames can be used, so that both monocular as well as multi-frame reconstruction can be performed.
- Chapter 7 (published as Tewari et al. (2020b)) presents a method to provide a face rig-like control over a pretrained and fixed generative adversarial network, namely StyleGAN, via a 3D Morphable Model. A new rigging network, RigNet is trained between the 3DMM's semantic parameters and StyleGAN's input. The network is trained in a self-supervised manner in face images, without the need for manual annotations. At test time, our method generates portrait images

with the photorealism of StyleGAN and provides explicit control over the 3D semantic parameters of the face.

- Chapter 8 (published as Tewari et al. (2020a)) presents the first approach for embedding real portrait images in the latent space of StyleGAN (Karras et al., 2019a), which allows for intuitive editing of the head pose, facial expression, and scene illumination in the image. Semantic editing in parameter space is achieved based on StyleRig, a pretrained neural network that maps the control space of a 3D morphable face model to the latent space of the GAN. A novel hierarchical non-linear optimization problem is solved to obtain the embedding. An identity preservation energy term allows spatially coherent edits while maintaining facial integrity. The approach runs at interactive frame rates and thus allows the user to explore the space of possible edits.
- Chapter 9 discusses important insights as well as opportunities for future work.

1.4 PUBLICATIONS

All the work presented in this thesis was also published in the following publications:

- Ayush Tewari et al. (2017). "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction." In: *The IEEE International Conference on Computer Vision (ICCV)*
- Ayush Tewari et al. (2020d). "High-Fidelity Monocular Face Reconstruction Based on an Unsupervised Model-Based Face Autoencoder." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2, pp. 357–370
- Ayush Tewari et al. (2018). "Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Ayush Tewari et al. (2019). "FML: Face model learning from videos." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Ayush Tewari et al. (2020b). "StyleRig: Rigging StyleGAN for 3D Control over Portrait Images." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Ayush Tewari et al. (2020a). "PIE: Portrait Image Embedding for Semantic Control." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*

In addition, contributions were made to the following publications which are, however, not part of this thesis:

- Hyeongwoo Kim et al. (2018b). "InverseFaceNet: Deep Single-Shot Inverse Face Rendering From A Single Image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Hyeongwoo Kim et al. (2018a). "Deep Video Portraits." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH)*
- Qianru Sun et al. (2018). "A Hybrid Model for Identity Obfuscation by Face Replacement." In: *European Conference on Computer Vision (ECCV)*

- Ohad Fried et al. (2019). "Text-based Editing of Talking-head Video." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH)*
- Ayush Tewari et al. (2020c). "State of the Art on Neural Rendering." In: *Computer Graphics Forum (EG STAR)*
- Bernhard Egger et al. (2020). "3D Morphable Face Models - Past, Present and Future." In: *ACM Transactions on Graphics (TOG)*
- Justus Thies et al. (2020). "Neural Voice Puppetry: Audio-driven Facial Reenactment." In: *European Conference on Computer Vision (ECCV)*
- Mohamed Elgharib et al. (Dec. 2020). "Egocentric Videoconferencing." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*
- Edgar Tretschk et al. (2020b). "DEMEA: Deep Mesh Autoencoders for Non-Rigidly Deforming Objects." In: *European Conference on Computer Vision (ECCV)*
- Edgar Tretschk et al. (2020a). "PatchNets: Patch-Based Generalizable Deep Implicit 3D Shape Representations." In: *European Conference on Computer Vision (ECCV)*
- Mallikarjun B R et al. (2021b). "Learning Complete 3D Morphable Face Models from Images and Videos." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Mallikarjun B R et al. (2021a). "Monocular Reconstruction of Neural Face Reflectance Fields." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Tarun Yenamandra et al. (2021). "i3DMM: Deep Implicit 3D Morphable Model of Human Heads." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

2 BACKGROUND

This chapter introduces the relevant technical background for the thesis. This thesis studies 3D morphable models and their applications in 3D reconstruction and image synthesis. Sec. 2.1 presents an overview of these models. The technical contributions of the thesis rely on a physically-based differentiable renderer, discussed in Sec. 2.2. Generative adversarial networks are also discussed in Sec. 2.3, as they are an important component in Chapters 7 and 8 for controllable synthesis.

2.1 3D MORPHABLE MODELS

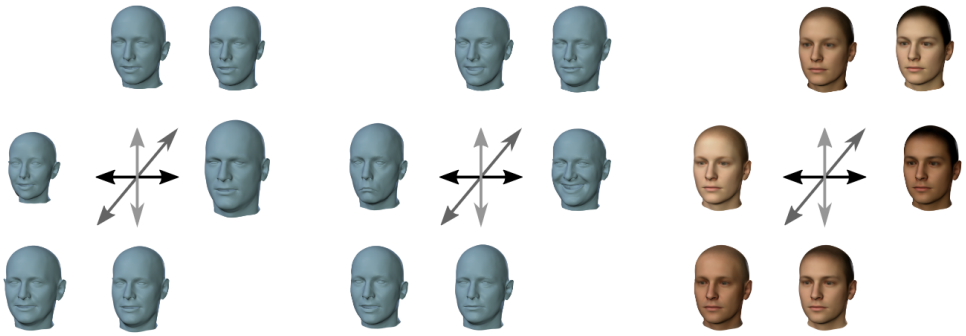

Figure 2.1: 3D morphable models represent the 3D geometry and appearance of faces using separate models for the identity geometry (left), expressions (middle), and appearance (right). The BFM 2019 model (Gerig et al., 2018) is visualized here. Figure taken from Egger et al. (2020).

3D morphable models (3DMMs) are generative models of the geometry and appearance of faces in 3D. These models are commonly used as priors for the underconstrained problem of monocular 3D reconstruction (Zollhöfer et al., 2018). The 3DMMs used in this thesis are defined using a template mesh with N vertices, and consist of semantically disentangled identity geometry, expression, and diffuse skin reflectance components, see Fig. 2.1. The expression component is also commonly referred to as a blendshape model (Lewis et al., 2014a). The parameters of a 3DMM include the facial expression parameters $\boldsymbol{\delta} \in \mathbb{R}^{N_e}$, identity shape parameters $\boldsymbol{\alpha} \in \mathbb{R}^{N_s}$, and skin reflectance parameters $\boldsymbol{\beta} \in \mathbb{R}^{N_r}$, where N_e, N_s, and N_r define the size of the models. This thesis studies linear models, i.e., each component of the model can be represented using a matrix. Thus, a 3DMM consists of an expression model $\mathbf{E}_e \in \mathbb{R}^{3N \times N_e}$, an identity shape model $\mathbf{E}_s \in \mathbb{R}^{3N \times N_s}$, and a skin reflectance model $\mathbf{E}_r \in \mathbb{R}^{3N \times N_r}$. It also includes an average face shape $\mathbf{A}_s \in \mathbb{R}^{3N}$, and an average face reflectance $\mathbf{A}_e \in \mathbb{R}^{3N}$. Given a 3DMM and its parameters, the face geometry and reflectance can be computed as:

$$\mathbf{V} = \mathbf{A}_s + \mathbf{E}_s\boldsymbol{\alpha} + \mathbf{E}_e\boldsymbol{\delta} \ , \tag{2.1}$$

$$\mathbf{R} = \mathbf{A}_r + \mathbf{E}_r\boldsymbol{\beta} \ . \tag{2.2}$$

Here, the x-, y-, z- coordinates of all vertices are stacked in the vector $\mathbf{V} \in \mathbb{R}^{3N}$. Similarly, the r-, g-, b- reflectance values of all vertices are stacked in the vector $\mathbf{R} \in \mathbb{R}^{3N}$.

Most existing 3DMMs are learned from datasets of 3D scans. The captured scans are first brought into dense correspondences. Principal component analysis of these processed scans is commonly used to learn the different components of the model. Unlike existing approaches, this thesis will introduce methods to learn 3DMMs directly from 2D data such as videos and images. This allows for better generalization of the models, compared to models learned only from 3D data.

2.2 DIFFERENTIABLE RENDERING

The thesis tackles the task of inverse rendering, where physical 3D parameters such as geometry, reflectance, and light are inferred from 2D image observations. A differentiable renderer is a crucial component for this task. This renderer describes how light interacts with the objects in a scene and how the scene is projected onto the image plane. It is easy to define a loss function for 3D face reconstruction using this renderer: we want the estimated 3D reconstruction to match the input image after the rendering process, see Fig. 2.2. Since this renderer is differentiable, the gradients from this loss function can be used for training. Throughout the thesis, 3D faces are represented as triangle meshes described using vertices, and triangles connecting these vertices. The albedo is represented per-vertex. The terms reflectance and albedo are used interchangeably.

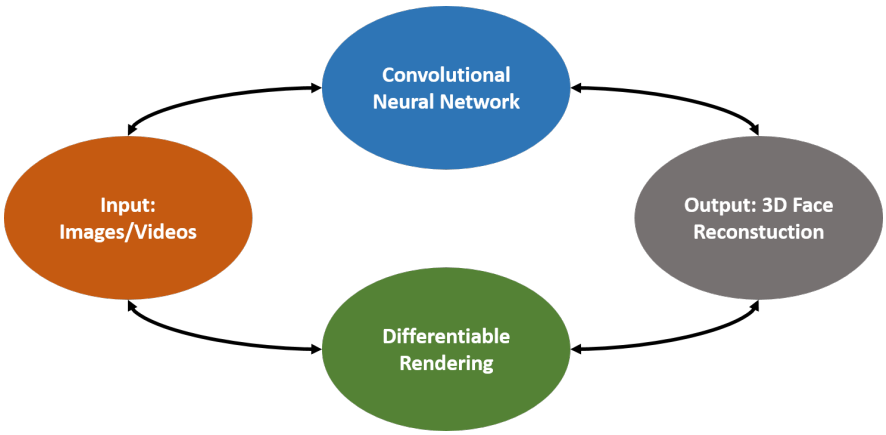

Figure 2.2: This thesis proposes a self-supervised training loop where a dataset of images or videos is used to train a convolutional neural network for 3D reconstruction. A differentiable physically-based renderer transforms the 3D reconstructions into synthetic images which can be used for defining self-supervised loss functions.

2.2.1 *Light Transport*

Light transport describes how the light in the scene interacts with the surface of the objects in order to obtain the final appearance. It is formulated using the rendering equation (Kajiya, 1986):

$$L_o(\mathbf{p}, \omega_o) = L_e(\mathbf{p}, \omega_o) + L_r(\mathbf{p}, \omega_o), \tag{2.3}$$

where $L_o(\mathbf{p}, \omega_o)$ is the outgoing radiance at point $\mathbf{p}$ on the surface in direction ω_o. $L_e(\mathbf{p}, \omega_o)$ is the light emitted by the point $\mathbf{p}$ and $L_r(\mathbf{p}, \omega_o)$ is the outgoing radiance due to the interaction of the surface with incoming light. We are interested in rendering human faces which do not emit any light, with $L_e(\mathbf{p}, \omega_o) = 0$ for all $\mathbf{p}$ and ω_o. The reflected component can be described as

$$L_r(\mathbf{p}, \omega_o) = \int_\Omega f_r(\mathbf{p}, \omega_i, \omega_o)\, L(\mathbf{p}, \omega_i)\, A(\mathbf{n}, \omega_i)\, \mathrm{d}\,\omega_i. \tag{2.4}$$

$L(\mathbf{p}, \omega_i)$ is the incoming light at $\mathbf{p}$ in direction ω_i, and $A(\mathbf{n}, \omega_i) = \max(\omega_i \cdot \mathbf{n}, 0)$, where $\mathbf{n}$ is the normal at point $\mathbf{p}$, see Fig. 2.3. The reflectance of the surface $f_r(\mathbf{p}, \omega_i, \omega_o)$ describes how the surface interacts with the incoming light. The integration is performed over the sphere of directions Ω. The rendering equation does not consider subsurface scattering effects, which play an important role in the appearance of skin. While it is possible to extend the equation in order to consider these effects (Jensen et al., 2001), the solution become intractable in our setting.

The rendering equation in Eq. 2.4, even without subsurface scattering cannot be trivially solved and does not have any closed form solution in the general case. Ray-tracing techniques (Pharr et al., 2016) are commonly used where camera rays are bounced around in the scene recursively to compute the outgoing radiance at a point. This thesis uses a differentiable renderer which is integrated with convolutional neural networks during training. Ray-tracing is computationally expensive and impractical in such settings. Thus, we simplify the rendering equation in the following paragraphs.

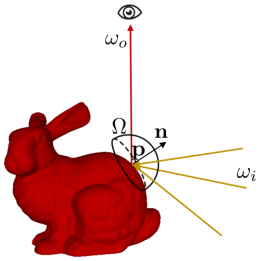

Figure 2.3: The rendering equation computes the outgoing radiance at a point by modeling how incoming light interacts with the surface.

LAMBERTIAN ASSUMPTION First, we assume that the surface of the face is lambertian, i.e., the light reflected at any point does not depend on the outgoing direction. We can then approximate $f_r(\mathbf{p}, \omega_i, \omega_o)$ as a constant diffuse albedo $a_\mathbf{p}$ for each point:

$$L_r(\mathbf{p}) = a_\mathbf{p} \int_\Omega L(\mathbf{p}, \omega)\, A(\mathbf{n}, \omega)\, \mathrm{d}\,\omega. \tag{2.5}$$

The incoming light direction is denoted with ω here.

DISTANT LIGHT ASSUMPTION The following chapters also solve for the scene lighting using image observations. This is a very challenging task in the general setting which requires several simplifications. The scene is approximated with only distant illumination where $L(\mathbf{p}, \omega)$ is independent of ω:

$$L_r(\mathbf{p}) = a_\mathbf{p} \int_\Omega L(\omega)\, A(\mathbf{n}, \omega)\, \mathrm{d}\,\omega. \tag{2.6}$$

While this does not allow us to capture inter-reflection and cast shadows, this is a crucial step required to solve the rendering equation in closed form.

SPHERICAL HARMONICS PROJECTION Since the surface is assumed to be diffuse, the distant lighting can be approximated using spherical harmonics. Spherical har-

monics (SH) are orthonormal basis functions, the analog of Fourier transform on the spherical domain. Any spherical function $f(\omega)$ can be projected onto the SH bases as

$$\boldsymbol{f}_i = \int_\Omega f(\omega) y_i(\omega) d\omega. \tag{2.7}$$

Here, $\boldsymbol{f}_i$ is the i-th SH coefficient and $y_i(\omega)$ is the i-th SH basis function. The indices are linearized, with $i = l^2 + l + m$, where l is the index of the SH band, and m is the index within the band, where $-l \leq m \leq l$. The original signal can then be approximated with n SH bands as

$$\tilde{f}(\omega) = \sum_{i=1}^{n^2} \boldsymbol{f}_i y_i(\omega). \tag{2.8}$$

Numerical integration is used for projecting general functions but a closed form solution exists for the transfer function $A(\mathbf{n}, \omega_\mathrm{i})$. Let $\boldsymbol{L}_\mathrm{i}$ and $\boldsymbol{A}(\mathbf{n})_\mathrm{i}$ denote the SH projections of the incoming light and the transfer functions. Spherical harmonics allow for an easy way to compute product integrals using a dot product of the SH coefficients:

$$L_\mathrm{r}(\mathbf{p}) \;=\; a_\mathbf{p} \int_\Omega \tilde{L}(\omega)\, \tilde{A}(\mathbf{n}, \omega)\; \mathrm{d}\,\omega \;=\; a_\mathbf{p} \sum_{i=1}^{n^2} \boldsymbol{L}_i \boldsymbol{A}(\mathbf{n})_i. \tag{2.9}$$

The radiance of diffuse surfaces can be represented with high accuracy using only 3 bands of SH coefficients. $L_\mathrm{r}(\mathbf{p})$ can be analytically computed as

$$\begin{aligned} L_\mathrm{r}(\mathbf{p}) = a_\mathbf{p} \Big(& c_4 \boldsymbol{L}_0 \; + 2c_2 (\boldsymbol{L}_3\, x + \boldsymbol{L}_1\, y + \boldsymbol{L}_2\, z) + c_1 \boldsymbol{L}_8\, (x^2 - y^2) \\ & + c_3 \boldsymbol{L}_6\, z^2 - c_5 \boldsymbol{L}_6 \; + 2c_1 (\boldsymbol{L}_4\, xy + \boldsymbol{L}_7\, xz + \boldsymbol{L}_5\, yz) \Big), \end{aligned} \tag{2.10}$$

where $c_1 = 0.42903$, $c_2 = 0.511664$, $c_3 = 0.743125$, $c_4 = 0.886227$, $c_5 = 0.247708$, and $\mathbf{n} = (x, y, z)$. Please refer to Ramamoorthi and Hanrahan (2001b) for the details. Since this thesis is interested in the inverse problem of estimating the incident light from images, we do not need to use the original light representation $L(\omega)$. Instead, we only estimate the SH coefficients $\boldsymbol{L}$. This allows us to skip the expensive numerical integration step required to project an environment map onto the SH bases.

2.2.2 *Camera Model and Projection*

Given the geometry and albedo of the face, and the scene illumination, we can now compute the outgoing radiance of each point on the surface. We now need to transform the scene to an image in the renderer.

PINHOLE CAMERA Several camera models exist in the computer graphics literature. This thesis uses a pinhole camera. While this model ignores several effects of real cameras such as lens distortions, it offers a good balance between simplicity and correctness. The camera is described using its intrinsics, i.e., focal length and principal point. For a 3D point $\mathbf{v} = (x, y, z)$, its corresponding image coordinates $\mathbf{p} = (u, v)$ can be computed as

$$u = f_x \frac{x}{z} + c_x,\; v = f_y \frac{x}{z} + c_y. \tag{2.11}$$

The focal length and principal points of the camera are represented using (f_x, f_y) and (c_x, c_y). This thesis works with in-the-wild images, where the camera intrinsics are not known. Estimating the intrinsics along with the other scene parameters just using image observations is very challenging, due to the ambiguities between the geometry and intrinsics. Thus, all cameras in the world are assumed to have the same intrinsics. The 3D reconstructions obtained can thus only be correct up to the errors due to this assumption.

POINT-BASED RENDERING Since our scenes are diffuse and the outgoing radiance of every point can be computed analytically, rasterization is the most suitable rendering technique because of its speed. If the renderer had to accommodate more complex global illumination effects, ray-tracing would be the more suitable method. The rasterization process first projects each triangle of the mesh onto the image plane. Each pixel is then colored with the corresponding projected triangle. The meshes used in this thesis are very dense with around $60,000$ vertices. The image observations are low resolution (240×240 pixels) in most cases. In this setting, the rendering process can be further simplified by considering the mesh as a point cloud and ignoring the mesh connectivity during rendering. Here, each vertex is projected to the corresponding point on the image. The points inside the mesh triangles are not considered for rendering. Note that this is used only for the differentiable renderer in the learning process. Rasterization with triangles is used for all mesh visualizations.

VISIBILITY The visibility of each vertex needs to be considered during rendering, such that occluded vertices are not rendered. Rasterization relies on techniques like z-buffering for this computation where the depth of each candidate triangle is compared for a pixel. Since this thesis deals with the frontal face region which is rather convex, visibility can be approximated using backface culling. If a vertex has normals facing towards the camera (front-facing), it is considered visible. While this offers a fast approximation for visibility, it is not always accurate. A front-facing vertex that is occluded by another surface in the scene would be considered visible by this technique.

2.2.3 *Gradients*

The only computation in the renderer which is not differentiable is the visibility component. The gradients at a pixel are only passed to the visible vertex, and not through the visibility computation. More details on the computation of gradients can be found in Sec. 4.4. The renderer is implemented on the GPU in a data-parallel manner and integrated with the deep-learning frameworks of Caffe (Jia et al., 2014) and Tensorflow (Abadi et al., 2015).

2.3 GENERATIVE NEURAL NETWORKS

In contrast to physically-based rendering, machine learning offers a different way of synthesizing images. The machine learning methods learn the distribution of real images from large datasets and use this learned distribution to sample realistic images. Most methods do not need any 3D assets such as geometry, light, etc. for synthesis. Several types of generative models have been proposed, such as variational

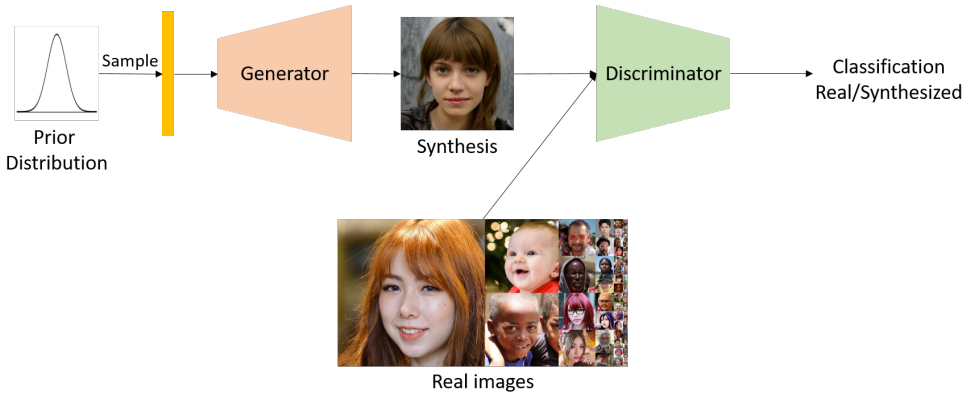

Figure 2.4: Generative Adversarial Networks (GANs) consist of a generator and a discriminator. Random samples from a prior distribution are given as input to the generator. The discriminator is trained to classify synthesized images from real images. A minimax optimization enables the generator to synthesize realistic images.

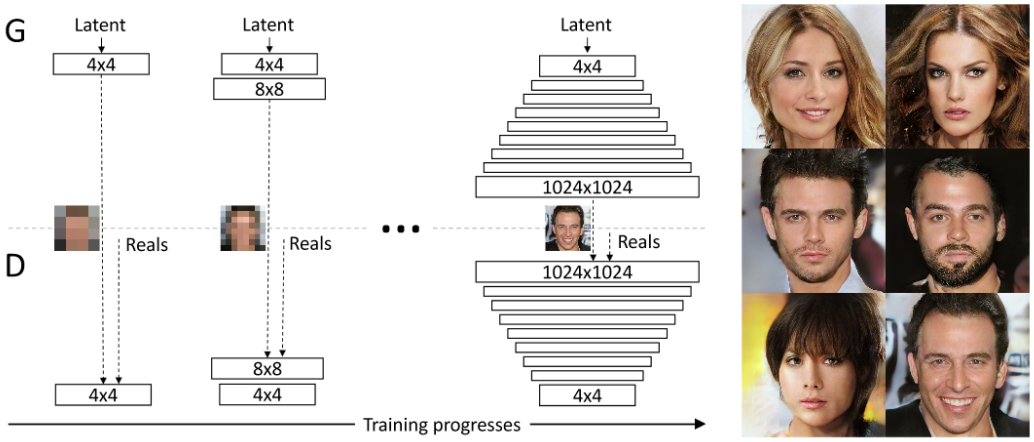

Figure 2.5: Progressive growing of the generator and discriminator networks allows for the synthesis of high-resolution images. Figure taken from Karras et al. (2018).

autoencoders (Doersch, 2016), autoregressive models (Oord et al., 2016), normalizing flow-based models (Kingma and Dhariwal, 2018) and adversarial models (Goodfellow et al., 2014). Here, we will look at the background of generative adversarial networks, which are used in Chapters 7 and 8.

2.3.1 *Generative Adversarial Networks*

A generative adversarial network consists of two sub-networks, a generator G and a discriminator D, see Fig. 2.4. The input to the generator is a latent noise vector z sampled from a prior distribution $p_z(z)$. The discriminator learns to distinguish between samples x from the real unknown distribution $p_d(x)$ and samples synthesized from the generator. A minimax optimization problem is designed with value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_d(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{2.12}$$

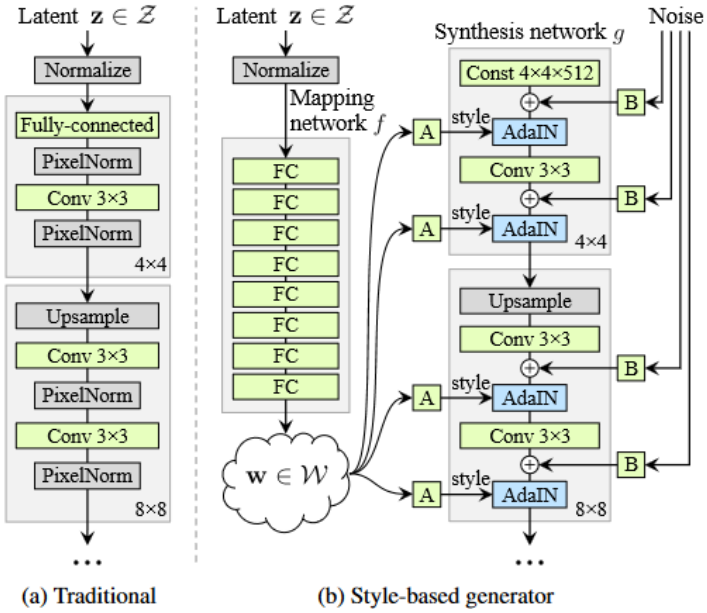

Figure 2.6: The StyleGAN architecture uses a mapping network to non-linearly transform the input latent vector. The transformed vector is broadcasted to each convolutional layer. Figure taken from Karras et al. (2019a).

While this original formulation was widely used, several improvements for more stable training have been introduced such as least squares GAN (Mao et al., 2017) and Wasserstein GAN (Arjovsky et al., 2017). Please see Mescheder et al. (2018) for a detailed discussion on the loss functions.

Orthogonal to the loss functions and regularizers, the network architecture of the generator and discriminator also play an important role. The original formulation of Goodfellow et al. (2014) used MLPs for these networks. Radford et al. (2015) introduced deep convolutional networks which led to higher quality and more stable training. Progressive growing of the networks was proposed by Karras et al. (2018) for synthesizing high-resolution images, see Fig. 2.5. Here, the network is learned in a coarse-to-fine manner, starting with very low resolutions. New layers are added for synthesizing higher resolution images. StyleGAN (Karras et al., 2019a,b) also uses this progressive growing strategy, but changes the nature of the input latent vector, see Fig. 2.6. While traditional generator architectures use the latent vector as input only to the first layer, StyleGAN broadcasts this latent vector at different resolutions. After training, the inputs to the different resolutions can be independently modified, leading to independent control over different scales of features in the image. For example, the latent vectors for the lower resolution layers change the global information like head pose, expressions, and background, while the latent vectors for the higher resolutions change the fine appearance and other high-frequency details. However, semantic components such as head pose or expressions cannot be independently controlled. A mapping network is also introduced in StyleGAN, which first transforms the randomly sampled latent vector non-linearly into a new latent space. The transformed latent space is more disentangled and leads to higher quality synthesis. Since generative models learn from data, these networks need to be trained on high-quality and high-

resolution images. When the goal is to synthesize portrait images, CelebA-HQ (Karras et al., 2018) and FFHQ (Karras et al., 2019a) are the commonly used image datasets.

While StyleGAN can randomly sample realistic portrait images, it does not offer semantic control akin to the computer graphics rendering pipeline. Chapters 7 and 8 present methods for semantic control over a pretrained and fixed StyleGAN network. This is done by building connections between the image synthesis pipelines in machine learning and computer graphics.

3

RELATED WORK

This chapter discusses the existing works for 3D reconstruction and controllable synthesis of faces. 3D morphable models learned from 3D scans are discussed first, followed by monocular 3D reconstruction, and joint reconstruction and model learning approaches. Deep generative models and portrait editing methods are also discussed in detail. A deeper discussion on these methods can be found in the recent state-of-the-art reports (Egger et al., 2020; Tewari et al., 2020c).

3.1 PARAMETRIC FACE MODELS FROM 3D SCANS

Active Appearance Models (AAMs) use a linear model for jointly capturing shape and texture variation (Cootes et al., 2001). Matching an AAM to an image is a registration problem, usually tackled via energy optimization. A closely related approach to AAMs is the 3D morphable model of faces (3DMM), introduced in Chapter 2, which represent 3D deformations in a low-dimensional subspace and are often built from scanner data (Blanz and Vetter, 1999; Bogo et al., 2014; Li et al., 2017). The most widely used face model is the one by Blanz and Vetter (1999), which is an affine parametric model of face geometry and texture that is learned from high-quality scans. Similar models, which also include facial animations, are presented in Blanz et al. (2003) and Gerig et al. (2018). Recently, Booth et al. (2016) created a Large-scale Facial Model (LSFM) from around 10,000 facial scans, which represents a richer shape distribution. Multilinear models generalize statistical models by capturing a set of mutually orthogonal variation modes (for example, global and local deformations) via a tensor decomposition (Bolkart and Wuhrer, 2015, 2016; Cao et al., 2013; Vlasic et al., 2005). However, unstructured subspaces or even tensor generalizations are incapable of modeling localized deformations from limited data. In this respect, Neumann et al. (2013) and Bernard et al. (2016) devised methods for computing sparse localized deformation components directly from mesh data. Lüthi et al. (2018) proposed the so-called Gaussian Process morphable models (GPMMs), which are modeled with arbitrary non-linear kernels, to handle strong non-linear shape deformations. Ranjan et al. (2018) built a non-linear model using a deep mesh autoencoder with fast spectral convolution kernels. Garrido et al. (2016b) trained radial basis function networks to learn a corrective 3D lip model from multiview data. Li et al. (2017) built a hybrid model that combines a linear shape space with articulated motions and semantic blendshapes. Although 3DMMs are highly efficient priors, they limit face reconstruction to a restricted low-dimensional subspace, for example, beards or characteristic noses cannot be reconstructed. The largest 3D scan datasets are still much smaller compared to image and video datasets. In this thesis, we take the first steps towards building morphable models directly from 2D data, which allows for better generalization.

3.2 OPTIMIZATION-BASED 3DMM RECONSTRUCTION

Many approaches for reconstruction from a single image (Garrido et al., 2016a; Romdhani and Vetter, 2005) and from image collections (Roth et al., 2016) are based on energy optimization. Here, the estimate in each iteration is rendered and compared to the input image(s). Such analysis-by-synthesis optimization is a widely studied problem. The discussion here is structured based on the modality of the input (monocular vs. multi-view).

3.2.1 *Monocular Reconstruction*

3DMMs have been widely used for monocular reconstruction using analysis-based-synthesis optimization (Blanz and Vetter, 1999; Paysan et al., 2009; Romdhani and Vetter, 2005). Many approaches require the computation of the 3D face silhouette in order to constrain it with the image silhouette. While most approaches use an incorrect fixed silhouette, some approaches allow the 3D silhouette to slide over a predefined path (for example, isolines) (Cao et al., 2014; Zhu et al., 2015) or iterate over a fixed vertex set to find 3D contour correspondences (Fried et al., 2016). Garrido et al. (2016a) obtained high-quality 3D face rigs from monocular RGB video based on a multi-layer model. Even real-time facial reconstruction and reenactment has been demonstrated (Huber et al., 2016; Thies et al., 2016b). Face tracking methods only reconstruct the per-frame expressions, pose, and/or illumination. The identity components are computed in a precomputation step. While real-time face tracking is in general feasible, optimization-based complete face reconstruction is computationally expensive and not feasible at real-time rates. Moreover, optimization-based approaches are sensitive to initialization, thus requiring 2D landmark detection (Jin and Tan, 2016; Wang et al., 2014). Chapter 4 of this thesis will introduce a combination of learning- and optimization-based approaches, which allows for addressing these limitations.

3.2.2 *Muti-Image Reconstruction*

Face reconstruction is also possible by fitting a template model to photo collections. Kemelmacher-Shlizerman and Seitz (2011) reconstructed an average shape and appearance model from a person-specific photo-collection via low-rank matrix factorization. Suwajanakorn et al. (2014) used this model to track detailed facial motion from unconstrained videos. Kemelmacher-Shlizerman (2013) built a 3DMM from a large photo collection of people, grouped into a fixed set of semantic labels. Liang et al. (2016) also leveraged multi-view person-specific photo-collections to reconstruct the full head. In a different line of research, Thies et al. (2015) fit a coarse parametric model to user-selected views to recover personalized face shape and albedo. Roth et al. (2016) personalized an existing morphable model to an image collection by using a coarse-to-fine photometric stereo formulation. Note that most of these methods do not learn a general face model, for example, a shape basis that spans the range of facial shapes of an entire population, but instead, they obtain a single person-specific 3D face instance. Besides, these methods require curated photo collections. This thesis, on the contrary, builds a 3DMM representation that generalizes across multiple face identities and imposes weaker assumptions on the training data.

Similar to photo-collections, multi-frame reconstruction techniques exploit either temporal information or multiple views to better estimate 3D geometry. Shi et al. (2014) globally fit a multilinear model to 3D landmarks at multiple keyframes and enforced temporal consistency of in-between frames via interpolation. Garrido et al. (2016a) obtained a person-specific facial shape by averaging per-frame estimates of a parametric face model. Ichim et al. (2015) employed a multi-view bundle adjustment approach to reconstruct facial shape and refine expressions using actor-specific sequences. Piotraschke and Blanz (2016) combined region-wise reconstructions of a 3DMM from many images using a normal distance function. Garg et al. (2013) proposed a model-free approach that globally optimizes for dense 3D geometry in a non-rigid structure from motion framework. Beyond faces, Tulsiani et al. (2017) trained a CNN to predict single-view 3D shape (represented as voxels) using multi-view ray consistency. Chapter 6 of this thesis will introduce a simple and intuitive approach to obtain higher-quality reconstructions using multiple images of a person.

3.3 SHAPE-FROM SHADING

Recovering fine-scale surface structure is a long-standing and well-researched problem in computer vision. Refinement techniques for general surfaces (Delaunoy and Prados, 2011; Li et al., 2016; Tylecek and Sara, 2010; Vu et al., 2012; Wu et al., 2011) are normally based on multi-view imagery. A variety of techniques exist in the context of facial detail estimation. Data-driven approaches (Cao et al., 2015; Huynh et al., 2018; Richardson et al., 2017) learn a mapping from the input image to the fine-scale geometric structure. While these approaches are in general fast, the recovered detail does not necessarily match the input. Some approaches produce details directly based on intensity variation (Beeler et al., 2010, 2011; Sela et al., 2017). While the obtained results look visually plausible, they are not physically accurate. Optimization-based refinement techniques (Garrido et al., 2013, 2016a; Shi et al., 2014) try to invert physical image formation models. Although the recovered detail in general matches the input, these approaches are computationally quite expensive, normally requiring several minutes to process a single frame. Chapter 4 of this thesis will leverage the data-parallel power of modern graphics cards to accelerate optimization-based mesh refinement.

3.4 LEARNING-BASED RECONSTRUCTION

Learning-based approaches regress 3D face geometry from a single image by learning an image-to-parameter or image-to-geometry mapping (Kim et al., 2018b; Olszewski et al., 2016; Richardson et al., 2017; Sela et al., 2017; Tewari et al., 2018, 2017; Tuan Tran et al., 2017). Supervised methods require ground truth face geometry (Laine et al., 2017; Tuan Tran et al., 2017). Since expensive multi-view capture setups are required to acquire such ground truth, some methods use a morphable model from which synthetic training images are generated (Dou et al., 2017; Kim et al., 2018b; Richardson et al., 2016, 2017; Sela et al., 2017). Some approaches use a mixture of synthetic and real supervised data (Klaudiny et al., 2017; McDonagh et al., 2016). Jackson et al. (2017) trained a CNN that directly regresses a volumetric 3D face representation from a single image. Trigeorgis et al. (2017) used a CNN to estimate surface normals from a given

input image. However, synthetic renderings usually lack realistic features, which has a negative impact on the reconstruction accuracy. Recently, some approaches allow for training networks on real images without 3D supervision, using analysis-by-synthesis loss functions. This was first introduced in Tewari et al. (2017), presented in Chapter 4 of this thesis. Deng et al. (2019) added a face recognition loss for higher quality reconstructions. Genova et al. (2018) demonstrated high-quality reconstructions of the identity components using unsupervised cycle-consistent and recognition losses.

3.5 LEARNING PARAMETRIC FACE MODELS FROM 2D DATA

Some approaches try to learn or refine 3D morphable models directly from 2D data. Personalized face models are extracted from monocular video by first refining an existing parametric model in a coarse-to-fine manner (for example, as in Roth et al. (2016)) and then learning a mapping from coarse semantic deformations to finer non-semantic detail layers (Bouaziz et al., 2013; Garrido et al., 2016a; Hsieh et al., 2015; Ichim et al., 2015). A number of works have been proposed for in-the-wild general 3DMM learning (Bas and Smith, 2018; Booth et al., 2017; Sengupta et al., 2018; Tran and Liu, 2018b). Most approaches here initialize the model with an existing 3DMM (Lin et al., 2020; Tran et al., 2019; Tran and Liu, 2018b). Some methods learn a general corrective space on top of an exiting 3DMM (Chaudhuri et al., 2020; Tewari et al., 2018). Tewari et al. (2018), presented in Chapter 5 of this thesis, was the first approach to obtain high-quality generalized corrective models. Tewari et al. (2019), presented in Chapter 6, was the first method for learning the identity components of a 3DMM from scratch. Learning without a good initial model is a more challenging task due to the ambiguities in the monocular setting. Thus, the approach in Chapter 6 uses multiple frames of a video to constrain the reconstructions. This model still uses a pre-trained expression model. B R et al. (2021b) proposed a method for learning all components of a 3DMM from videos, only using a single face scan.

3.6 DEEP GENERATIVE MODELS

Generative adversarial networks (GANs), introduced in Chapter 2, are networks that learn the manifold of real images in a training dataset. These networks consist of two main blocks: a generator and a discriminator (Goodfellow et al., 2014). The generator takes a random noise vector from a prior distribution as an input to produce an image. The discriminator tries to distinguish between the real and synthesized images. These networks are trained adversarially with complementary objectives. Karras et al. (2018) showed that such a network can generate high-resolution photorealistic images of human faces. To achieve this, they employed a progressive strategy of slowly increasing the size of the generator and the discriminator by adding more layers during training. This enables a more stable training, and in turn, helps learn high-resolution images of faces. StyleGAN (Karras et al., 2019a) can synthesize highly photorealistic images while allowing for more control over the output, compared to Karras et al. (2018). However, StyleGAN still suffers from a clear entanglement of semantically different attributes. Therefore, it does not provide an interpretable control over the image synthesis process. Exploring the latent space of GANs for image editing has been recently explored in Jahanian et al. (2019). They can only

achieve simple transformations, such as zoom and 2D translations, as they need ground truth images for each transformation during training. For faces, concurrent efforts have been made in controlling images synthesized by GANs (Abdal et al., 2019; Shen et al., 2020), but they lack explicit rig-like 3D control of the generative model. Chapter 7 will present an approach for such rig-like control by a combination of a 3DMM (Paysan et al., 2009) and StyleGAN (Karras et al., 2019a). Isola et al. (2017) used conditional GANs to produce image-to-image translations. Here, the input is a conditional image from a source domain, which is translated to the target domain by the generator. Their approach, however, requires paired training data between the source and target domains. CycleGAN (Zhu et al., 2017) and UNIT (Liu et al., 2017) learn to perform image-to-image translation with unpaired data using cycle-consistency losses. GAUGAN (Park et al., 2019) shows interactive semantic image synthesis based on spatially adaptive normalization. The remarkable quality achieved by GANs has inspired the development of several neural rendering applications for faces (Egger et al., 2020; Tewari et al., 2020c; Zollhöfer et al., 2018) and others objects (Chan et al., 2019; Martin-Brualla et al., 2018; Yu and Smith, 2019).

3.6.1 *Person-specific Video Editing Techniques*

There has been a lot of research on person-specific techniques that require a large training corpus of the target person as input (Bansal et al., 2018; Kim et al., 2019; Kim et al., 2018a; Thies et al., 2016a, 2019; Wiles et al., 2018). These approaches can be classified into model-based (Kim et al., 2019; Kim et al., 2018a; Thies et al., 2016a, 2019) and image-based (Bansal et al., 2018) techniques. Model-based techniques employ a parametric face model to represent the head pose, facial expression, and incident scene illumination. The semantic parameter space spanned by the model can be used to either perform intuitive edits or transfer parameters from a source to a target video. On the other end of the spectrum are image-based techniques that can transfer parameters but do not provide intuitive semantic control.

MODEL-BASED VIDEO EDITING TECHNIQUES Facial reenactment approaches (Thies et al., 2016a, 2019) change the facial expressions in a target video to the expressions in a driving source video. These approaches achieve impressive results but require a video of the target person as input and do not enable editing of the head pose and incident illumination. Kim et al. (2018a) proposed the first full head reenactment approach that is able to edit the head pose as well as the facial expression. A conditional deep generative model is leveraged as a neural rendering engine. While these approaches (Kim et al., 2018a; Thies et al., 2016a, 2019) produce exciting results, they do not preserve the speaking style of the target. Kim et al. (2019) proposed an approach for editing the expressions of a target subject while maintaining his/her speaking style. This is made possible by a novel style translation network that learns a cycle-consistent mapping in blendshape space. In contrast to the approach presented in Chapter 8, all these techniques require a long video of the target as input and cannot edit a single image of an arbitrary person.

IMAGE-BASED VIDEO EDITING TECHNIQUES Image-based techniques enable control of a target face through a driving video. The approach of Bansal et al. (2018)

allows them to modify the target video while maintaining the speaking style. A novel recycle loss is defined in the spatio-temporal video domain. This approach obtains high-quality results for expressions and pose transfer. In contrast to the approach presented in Chapter 8, image-based approaches do not provide intuitive control via a set of semantic control parameters and have to be trained in a person-specific manner. Thus, they cannot be employed to edit a single given image.

3.6.2 *Few-shot Editing Techniques*

Few-shot editing techniques (Wang et al., 2019b; Wiles et al., 2018; Zakharov et al., 2019) require only a small set of images of the target person as input. Given multiple frames showing a target person, X2Face (Wiles et al., 2018) drives a frontalized face embedding by a regressed warp field that is estimated by an encoder-decoder network. The approach can also drive faces based on audio. Wang et al. (2019b) presented a few-shot video editing approach and demonstrated the reenactment of a target face via a source video. A novel network weight generation module is proposed that is based on an attention mechanism. To animate faces, the network is trained to transfer image sketches to photo-realistic face images. The network is trained on a large multi-identity training corpus and can be applied to new unseen still images. Zakharov et al. (2019) presented a few-shot technique for animating faces. Their solution has three components: 1) a generator network that translates landmark positions to photo-realistic images, 2) an embedding network that learns an intermediate representation for conditioning the generator, and 3) a discriminator. The network is trained on a large corpus of face images across multiple identities and generalizes to new identities at test time. Impressive results are shown in animating images, including legacy photos and even paintings. The learned models of few-shot techniques (Wang et al., 2019b; Wiles et al., 2018; Zakharov et al., 2019) can be improved by fine-tuning on a few example images of the target person, such as images taken from different viewpoints or at different time instances. While these methods can also be used in the single-shot setting where only a single image of the target person is available, a detailed discussion of methods that operate in such a setting follows next.

3.6.3 *Single-shot Editing Techniques*

Several works (Averbuch-Elor et al., 2017; Geng et al., 2018; Nagano et al., 2018) exist for controlling the expression and head pose in an image without any other image/video of the person in the image. Nagano et al. (2018) presented *paGAN*, an approach for creating personalized avatars from just a single image of a person. However, the work does not synthesize photo-realistic hair. The approach of Averbuch-Elor et al. (2017) brings portrait images to life by animating their expression and pose. The target image is animated through a 2D warp computed from the source video's movement. The mouth interior is copied from the source and blended into the warped target image. The approach of Geng et al. (2018) employs deep generative models to synthesize more realistic facial details and a higher quality mouth interior. First, a dense spatial motion field is used to warp the target image. Afterward, the first network corrects the warped target image and synthesizes important skin detail. Finally, the second network synthesizes the mouth interior, including realistic teeth.

Siarohin et al. (2019) proposed a method for animating a single image based on a driving sequence. The method uses a neural network to compute a dense warping field by detecting keypoints in both the target image and the driving frames. Based on this information, a second network produces high-quality output frames. Since keypoint extraction is also learned during training, the method is applicable for any input category, particularly for face and full-body images. While these methods can only be controlled via a driving video, the approach presented in Chapter 8 enables intuitive editing of the head pose, facial expression, and incident illumination in a portrait image through intuitive parametric control.

3.6.4 *Portrait Relighting*

The discussion above focussed on editing the expressions and pose of the person. Editing the appearance of the image by relighting them is also important for casual photography applications. Relighting approaches modify the incident illumination on the face (Meka et al., 2019; Peers et al., 2007; Shu et al., 2017a; Sun et al., 2019; Zhou et al., 2019). Earlier works (Peers et al., 2007; Shu et al., 2017a) require an exemplar portrait image that has been taken under the target illumination conditions. More recent techniques use deep generative models (Meka et al., 2019; Sun et al., 2019; Zhou et al., 2019) and could relight images based on an environment map. Zhou et al. (2019) trained a relighting technique based on a large corpus of synthetic images. Relighting is performed in the luminance channel, which simplifies the learning task. Sun et al. (2019) used light stage data to train their relighting approach. At test time, the network produces high-quality relighting results, even for in-the-wild images. While training with light stage data leads to high-quality results, their scarcity and careful recording protocol can limit their adaptation. Meka et al. (2019) showed that the 4D reflectance field can be estimated from two color gradient images captured in a light stage. This allows for capturing relightable videos.

3.6.5 *Image Editing using SyleGAN*

Several recent methods have been proposed to edit images generated with StyleGAN. Most approaches linearly change the StyleGAN latent codes for editing (Härkönen et al., 2020; Shen et al., 2020; Tewari et al., 2020b). Non-linear editing has been shown in Abdal et al. (2020b). Image2StyleGAN (Abdal et al., 2019, 2020a) is a popular approach for embedding real images into the StyleGAN latent space with very high fidelity. InterFaceGAN (Shen et al., 2020) and StyleFlow (Abdal et al., 2020b) demonstrated editing of real images using these embeddings. Very recently, Zhu et al. (2020) introduced a domain-guided embedding method that allows for higher-quality editing, compared to Image2StyleGAN. However, they did not demonstrate results at the highest resolution for StyleGAN.

4

HIGH-FIDELITY MONOCULAR FACE RECONSTRUCTION BASED ON AN UNSUPERVISED MODEL-BASED FACE AUTOENCODER

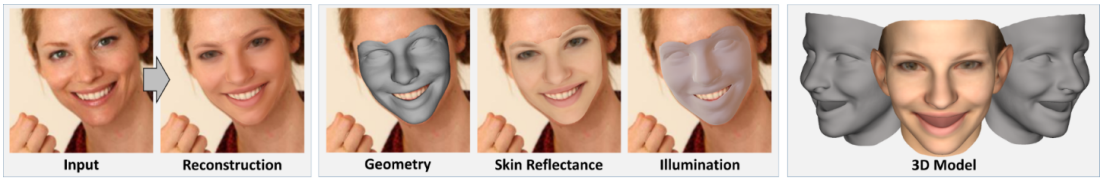

Figure 4.1: This chapter presents a model-based deep convolutional face autoencoder which enables unsupervised learning of semantic pose, shape, expression, reflectance and lighting parameters. The trained encoder predicts these parameters from a single monocular image, all at once.

This chapter presents a novel model-based deep convolutional autoencoder that addresses the highly challenging problem of reconstructing a 3D human face from a single in-the-wild color image (published as Tewari et al. (2020d, 2017)), see Fig. 4.1. To this end, a convolutional encoder network is combined with an expert-designed generative model that serves as decoder. The core innovation is the differentiable parametric decoder that encapsulates image formation analytically based on a generative model. The decoder takes as input a code vector with exactly defined semantic meaning that encodes detailed face pose, shape, expression, skin reflectance and scene illumination. Due to this new way of combining CNN-based with model-based face reconstruction, the CNN-based encoder learns to extract semantically meaningful parameters from a single monocular input image. For the first time, a CNN encoder and an expert-designed generative model can be trained end-to-end in an unsupervised manner, which renders training on very large (unlabeled) real world datasets feasible. The obtained reconstructions compare favorably to current state-of-the-art approaches in terms of quality and richness of representation. This chapter also presents a stochastic vertex sampling technique for faster training of our networks, and moreover, analysis-by-synthesis and shape-from-shading refinement approaches to achieve higher-fidelity reconstructions.

4.1 INTRODUCTION

Detailed, dense 3D reconstruction of the human face from image data is a longstanding problem in computer vision and computer graphics. Previous approaches have tackled this challenging problem using calibrated multi-view data or uncalibrated photo collections (Kemelmacher-Shlizerman and Seitz, 2011; Roth et al., 2016). Robust and detailed three-dimensional face reconstruction from a single arbitrary in-the-wild image, for example, downloaded from the Internet, is still an open research problem due to the high degree of variability of uncalibrated photos in terms of resolution and employed imaging device. In addition, in unconstrained photos, faces show a high variability in global pose, facial expression, and are captured under diverse and

difficult lighting. Detailed 3D face reconstruction is the foundation for a broad scope of applications, which range from robust face recognition, over emotion estimation, to complex image manipulation tasks. In many applications, faces should ideally be reconstructed in terms of meaningful low-dimensional model parameters, which facilitates interpretation and manipulation of reconstructions (Thies et al., 2016b).

Recent monocular reconstruction methods broadly fall into two categories: Generative and regression-based. Generative approaches fit a parametric face model to image and video data, for example, Blanz et al. (2003), Blanz and Vetter (1999), and Fried et al. (2016), by optimizing the alignment between the projected model and the image (Garrido et al., 2016a; Kemelmacher-Shlizerman et al., 2010; Suwajanakorn et al., 2014, 2015; Thies et al., 2016b). State-of-the-art generative approaches capture very detailed and complete 3D face models on the basis of semantically meaningful low-dimensional parameterizations (Garrido et al., 2016a; Thies et al., 2016b). Unfortunately, the fitting energies are usually highly non-convex. Good results thus require an initialization close to the global optimum, which is only possible with some level of control during image capture or additional input data, for example, detected landmarks.

Only recently, the first regression-based approaches for dense 3D face reconstruction based on deep convolutional neural networks were proposed. Richardson et al. (2016) use iterative regression to obtain a high quality estimate of pose, shape and expression, and finer scale surface detail (Richardson et al., 2017) of a face model. The expression-invariant regression approach of Tuan Tran et al. (2017) obtains high-quality estimates of shape and skin reflectance. Based on an image-to-image translation network, Sela et al. (2017) obtain the facial geometry from a single image by translating the input image to a depth map. Unfortunately, these approaches can only be trained in a supervised fashion on corpora of densely annotated facial images whose creation poses a major obstacle in practice. In particular, the creation of a training corpus of photo-realistic synthetic facial images that include facial hair, parts of the upper body and a consistent background is challenging. While the refinement network of Richardson et al. (2017) can be trained in an unsupervised manner, their coarse shape regression network requires synthetic ground truth data for training. Also, the quality and richness of representation (e.g., illumination and colored reflectance in addition to geometry) of these methods does not match the best generative ones. However, trained networks are efficient to evaluate and can be trained to achieve remarkable robustness under difficult real world conditions.

This chapter contributes a new type of model-based face autoencoder (MoFA) that joins forces of state-of-the-art generative and CNN-based regression approaches for dense 3D face reconstruction via a deep integration of the two on an architectural level. The network architecture is inspired by recent progress on deep convolutional autoencoders, which, in their original form, couple a CNN encoder and a CNN decoder through a code-layer of reduced dimensionality (Hinton and Salakhutdinov, 2006; Masci et al., 2011; Zhao et al., 2016). Unlike previously used CNN-based decoders, the proposed convolutional autoencoder deeply integrates an expert-designed decoder. This layer implements, in closed form, an analytically-differentiable image formation model on the basis of a detailed parametric 3D face model (Blanz and Vetter, 1999). Some previous fully CNN-based autoencoders tried to disentangle (Grant et al., 2016; Kulkarni et al., 2015), but could not fully guarantee the semantic meaning of code layer parameters. In the new network, exact semantic meaning of the code vector,

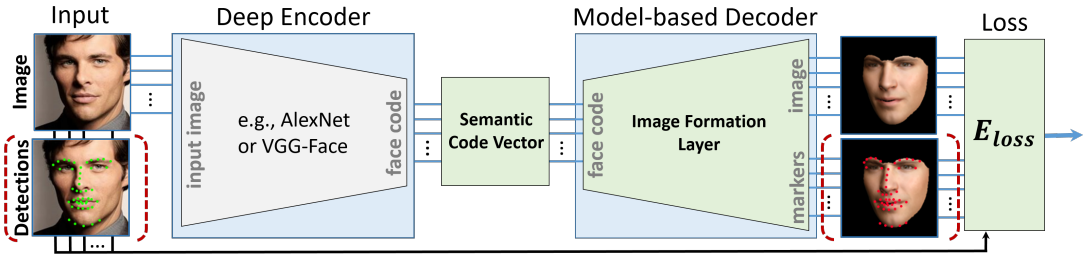

Figure 4.2: The proposed deep model-based face autoencoder enables unsupervised end-to-end learning of semantic parameters, such as pose, shape, expression, skin reflectance and illumination. An optional landmark-based surrogate loss enables faster convergence and improved reconstruction results, see Sec. 4.5. Both scenarios require no supervision of the semantic parameters during training.

i.e., the input to the decoder, is ensured by design. Moreover, the proposed decoder is compact and does not need training of enormous sets of unintuitive CNN weights.

Unlike previous CNN regression-based approaches for face reconstruction, a single forward pass of the network estimates a much more complete face model, including pose, shape, expression, skin reflectance, and illumination, at a high quality. The new network architecture allows, for the first time, combined end-to-end training of a sophisticated model-based (generative) decoder and a CNN encoder, with error backpropagation through all layers. It also allows, for the first time, unsupervised training of a network that reconstructs dense and semantically meaningful faces on unlabeled in-the-wild images via a dense photometric training loss. In consequence, the network generalizes better to real world data compared to networks trained on synthetic face data (Richardson et al., 2016, 2017). This chapter further introduces a stochastic vertex sampling strategy to train the networks faster. Since learning-based approaches have limited capacity, they have to trade-off the quality of individual reconstructions in order to work on a diverse range of images. Therefore, optimization-based techniques that can be added as refinement steps to further improve the quality of the results are also explored. The focus is on a fast data-parallel implementation of these two additional steps.

4.2 OVERVIEW

The proposed novel deep convolutional model-based face autoencoder enables unsupervised end-to-end learning of a network which estimates meaningful semantic face and rendering parameters, see Fig. 4.2. To this end, convolutional encoders are combined with an expert-designed differentiable model-based decoder that analytically implements image formation. The decoder generates a realistic synthetic image of a face and enforces semantic meaning by design. Rendering is based on an image formation model that enforces full semantic meaning via a parametric face prior, see Chapter 2 for details. More specifically, pose, shape, expression, skin reflectance and illumination are independently parameterized. The synthesized image is compared to the input image using a robust photometric loss E_{loss} that includes statistical regularization of the face. In combination, this enables unsupervised end-to-end training of our networks. 2D facial landmark locations can be optionally provided to add a surrogate loss for faster convergence and improved reconstructions, see Sec. 4.5. Note,

both scenarios require no supervision of the semantic parameters. After training, the encoder part of the network enables regression of a dense face model and illumination from a single monocular image, without requiring any other input, such as landmarks.

4.3 SEMANTIC CODE VECTOR

The semantic code vector $\mathbf{x} \in \mathbb{R}^{257}$ parameterizes the facial expression $\boldsymbol{\delta} \in \mathbb{R}^{64}$, shape $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, skin reflectance $\boldsymbol{\beta} \in \mathbb{R}^{80}$, camera rotation $\mathbf{T} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$, and the scene illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$ in a unified manner:

$$\mathbf{x} = (\underbrace{\boldsymbol{\alpha},\ \boldsymbol{\delta},\ \boldsymbol{\beta}}_{\text{face}},\ \underbrace{\mathbf{T},\ \mathbf{t},\ \boldsymbol{\gamma}}_{\text{scene}}) \ . \tag{4.1}$$

In the following, the parameters that are associated with the employed face model are described. The parameters that govern image formation are described in Sec. 4.4.

The face is represented as a manifold triangle mesh with $N = 24$k vertices $\mathbf{V} = \{\mathbf{v}_i \in \mathbb{R}^3 | 1 \leq i \leq N\}$. The associated vertex normals $\mathbf{N} = \{\mathbf{n}_i \in \mathbb{R}^3 | 1 \leq i \leq N\}$ are computed using a local one-ring neighborhood. The spatial embedding $\mathbf{V}$ is parameterized by an affine face model:

$$\mathbf{V} = \hat{\mathbf{V}}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \mathbf{A}_s + \mathbf{E}_s\boldsymbol{\alpha} + \mathbf{E}_e\boldsymbol{\delta} \ . \tag{4.2}$$

Note that, by abuse of notation, here the point-set $\mathbf{V}$ is represented as a $3N$-dimensional vector. Here, the average face shape $\mathbf{A}_s$ has been computed based on 200 (100 male, 100 female) high-quality face scans (Blanz and Vetter, 1999). The linear PCA bases $\mathbf{E}_s \in \mathbb{R}^{3N\times 80}$ and $\mathbf{E}_e \in \mathbb{R}^{3N\times 64}$ encode the modes with the highest shape and expression variation, respectively. The expression basis is obtained by applying PCA to the combined set of blendshapes of Alexander et al. (2009) and Cao et al. (2013), which have been re-targeted to the face topology of Blanz and Vetter (1999) using deformation transfer (Sumner and Popović, 2004). The PCA basis covers more than 99% of the variance of the original blendshapes.

In addition to facial geometry, per-vertex skin reflectance is parameterized as $\mathbf{R} = \{\mathbf{r}_i \in \mathbb{R}^3 | 1 \leq i \leq N\}$ based on an affine parametric model:

$$\mathbf{R} = \hat{\mathbf{R}}(\boldsymbol{\beta}) = \mathbf{A}_r + \mathbf{E}_r\boldsymbol{\beta} \ . \tag{4.3}$$

Here, the average skin reflectance $\mathbf{A}_r$ has been computed based on Blanz and Vetter (1999) and the orthogonal PCA basis $\mathbf{E}_r \in \mathbb{R}^{3N\times 80}$ captures the modes of highest variation. Note, all basis vectors are already scaled with the appropriate standard deviations $\sigma_k^\bullet$ such that $\mathbf{E}_\bullet^T\mathbf{E}_\bullet = \text{diag}(\cdots, [\sigma_k^\bullet]^2, \cdots)$.

4.4 PARAMETRIC MODEL-BASED DECODER

Given a scene description in the form of a semantic code vector $\mathbf{x}$, the parametric decoder generates a realistic synthetic image of the corresponding face. Since the image formation model is fully analytical and differentiable, an efficient backward pass is implemented that inverts image formation via standard backpropagation. This enables unsupervised end-to-end training of the network. The image formation model employed is described in the following.

PERSPECTIVE CAMERA Realistic facial imagery are rendered using a pinhole camera model under a full perspective projection $\Pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ that maps camera space coordinates onto screen space coordinates. The position and orientation of the camera in world space is given by a rigid transformation, which we parameterize based on a rotation $\mathbf{T} \in \mathbf{SO}(3)$ and a global translation $\mathbf{t} \in \mathbb{R}^3$. Hence, the functions $\Phi_{\mathbf{T},\mathbf{t}}(\mathbf{v}) = \mathbf{T}^{-1}(\mathbf{v} - \mathbf{t})$ and $\Pi \circ \Phi_{\mathbf{T},\mathbf{t}}(\mathbf{v})$ transform an arbitrary point $\mathbf{v}$ from world space into camera space and further into screen space, respectively.

ILLUMINATION MODEL Scene illumination is represented using Spherical Harmonics (SH) (Müller, 1966). Here, distant low-frequency illumination and a purely *Lambertian* surface reflectance are assumed. Thus, the radiance is evaluated at vertex $\mathbf{v}_i$ with surface normal $\mathbf{n}_i$ and skin reflectance $\mathbf{r}_i$ as follows:

$$C(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma}) = \mathbf{r}_i \cdot \sum_{b=1}^{B^2} \boldsymbol{\gamma}_b \mathbf{H}_b(\mathbf{n}_i) \ . \tag{4.4}$$

The $\mathbf{H}_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ are SH basis functions and the $B^2 = 9$ coefficients $\boldsymbol{\gamma}_b \in \mathbb{R}^3$ ($B = 3$ bands) parameterize colored illumination using the red, green and blue channel. Please refer to Chapter 2 for details.

IMAGE FORMATION Realistic images of the face are rendered using the presented camera and illumination model. To this end, in the forward pass $\mathcal{F}$, the screen space position $\mathbf{u}_i(\mathbf{x})$ and the associated pixel color $\mathbf{c}_i(\mathbf{x})$ is computed for each $\mathbf{v}_i$:

$$\begin{aligned} \mathcal{F}_i(\mathbf{x}) &= [\mathbf{u}_i(\mathbf{x}), \mathbf{c}_i(\mathbf{x})]^T \in \mathbb{R}^5 \ , \\ \mathbf{u}_i(\mathbf{x}) &= \Pi \circ \Phi_{\mathbf{T},\mathbf{t}}\big(\hat{\mathbf{V}}_i(\boldsymbol{\alpha}, \boldsymbol{\delta})\big) \ , \\ \mathbf{c}_i(\mathbf{x}) &= C\big(\hat{\mathbf{R}}_i(\boldsymbol{\beta}), \mathbf{T}\mathbf{n}_i(\boldsymbol{\alpha}, \boldsymbol{\delta}), \boldsymbol{\gamma}\big) \ . \end{aligned} \tag{4.5}$$

Here, $\mathbf{T}\mathbf{n}_i$ transforms the world space normals into camera space and γ models illumination in camera space.

BACKPROPAGATION To enable training, a backward pass is implemented that inverts image formation:

$$\mathcal{B}_i(\mathbf{x}) = \frac{\mathrm{d}\mathcal{F}_i(\mathbf{x})}{\mathrm{d}(\boldsymbol{\alpha},\ \boldsymbol{\delta},\ \boldsymbol{\beta},\ \mathbf{T},\ \mathbf{t},\ \boldsymbol{\gamma})} \in \mathbb{R}^{5\times 257} \ . \tag{4.6}$$

This requires the computation of the gradients of the image formation model (see Eq. (4.5)) with respect to the face and scene parameters. For high efficiency during training, the gradients are evaluated in a data-parallel manner, see Sec. 4.5.

4.5 LOSS LAYER

A robust dense photometric loss function is employed that enables efficient end-to-end training of our networks. The loss is inspired by recent optimization-based approaches (Garrido et al., 2016a; Thies et al., 2016b) and combines three terms:

$$E_{\text{loss}}(\mathbf{x}) = \underbrace{w_{\text{land}} E_{\text{land}}(\mathbf{x}) + w_{\text{photo}} E_{\text{photo}}(\mathbf{x})}_{\text{data term}} + \underbrace{w_{\text{reg}} E_{\text{reg}}(\mathbf{x})}_{\text{regularizer}} \ . \tag{4.7}$$

Here, the loss enforces sparse landmark alignment E_{land}, dense photometric alignment E_{photo} and statistical plausibility E_{reg} of the modeled faces. Note, E_{land} is optional and implements a surrogate loss that can be used to speed up convergence, see Sec. 4.7. The binary weight $w_{\text{land}} \in \{0, 1\}$ toggles this constraint. The constant weights $w_{\text{photo}} = 1.92$ and $w_{\text{reg}} = 2.9 \times 10^{-5}$ are empirically determined.

DENSE PHOTOMETRIC ALIGNMENT LOSS The goal of the encoder is to predict model parameters that lead to a synthetic face image that matches the provided monocular input image. To this end, a loss enforcing dense photometric alignment is employed, similar to Thies et al. (2016b), on a per-vertex level using a robust $\ell_{2,1}$-norm:

$$E_{\text{photo}}(\mathbf{x}) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left\| \mathcal{I}\big(\mathbf{u}_i(\mathbf{x})\big) - \mathbf{c}_i(\mathbf{x}) \right\|_2 . \tag{4.8}$$

Here, $\mathcal{I}$ is an image of the training corpus and for occlusion awareness, all visible vertices are iterated over, approximate as the set of front-facing vertices $\mathcal{V}$.

SPARSE LANDMARK ALIGNMENT In addition to dense photometric alignment, an optional surrogate loss is proposed based on the detected facial landmarks Saragih et al. (2011). A subset of 46 landmarks (out of 66) is used, see Fig. 4.2. Given the subset $\mathcal{L} = \{(\mathbf{s}_j, c_j, k_j)\}_{j=1}^{46}$ of detected 2D landmarks $\mathbf{s}_j \in \mathbb{R}^2$, with confidence $c_j \in [0, 1]$ (1 confident) and corresponding model vertex index $k_j \in \{1, ..., N\}$, projected 3D vertices are enforced to be close to the 2D detections:

$$E_{\text{land}}(\mathbf{x}) = \sum_{j=1}^{46} c_j \cdot \left\| \mathbf{u}_{k_j}(\mathbf{x}) - \mathbf{s}_j \right\|_2^2 . \tag{4.9}$$

Please note, this surrogate loss is optional. The networks can be trained fully unsupervised without supplying these sparse constraints. After training, landmarks are never required.

STATISTICAL REGULARIZATION During training, the optimization problem is further constrained using statistical regularization (Blanz and Vetter, 1999) on the model parameters:

$$E_{\text{reg}}(\mathbf{x}) = \sum_{k=1}^{80} \boldsymbol{\alpha}_k^2 + w_{\boldsymbol{\beta}} \sum_{k=1}^{80} \boldsymbol{\beta}_k^2 + w_{\boldsymbol{\delta}} \sum_{k=1}^{64} \boldsymbol{\delta}_k^2 . \tag{4.10}$$

This constraint enforces plausible facial shape $\boldsymbol{\alpha}$, expression $\boldsymbol{\delta}$ and skin reflectance $\boldsymbol{\beta}$ by preferring values close to the average (the basis of the linear face model is already scaled by the standard deviations). The parameters $w_{\boldsymbol{\beta}} = 1.7 \times 10^{-3}$ and $w_{\boldsymbol{\delta}} = 0.8$ balance the importance of the terms. Note, pose $(\mathbf{T}, \mathbf{t})$ and illumination $\boldsymbol{\gamma}$ are not regularized.

BACKPROPAGATION The gradient of the robust loss is passed backward to the model-based decoder and is combined with $\mathcal{B}_i(\mathbf{x})$ using the chain rule. This enables training via stochastic gradient descent during backpropagation,

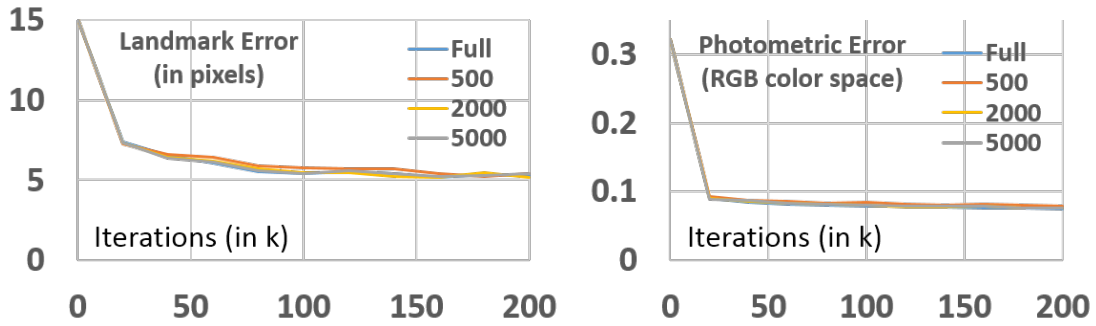

Figure 4.3: Quantitative evaluation of stochastic sampling on real data. Even drastic sampling of $\approx 2\%$ of vertices only marginally reduces the quality of the reconstruction results.

DATA-PARALLEL GPU IMPLEMENTATION Eq. (4.8) is implemented in an iteratively reweighted fashion as follows:

$$E_{\text{photo}}(\mathbf{x}) = \frac{1}{N} \sum_{i \in \mathcal{V}} \frac{1}{C_i} \left\| \mathcal{I}\big(\mathbf{u}_i(\mathbf{x})\big) - \mathbf{c}_i(\mathbf{x}) \right\|_2^2, \tag{4.11}$$

where $C_i = \|\mathcal{I}(\mathbf{u}_i(\mathbf{x}^{\text{old}})) - \mathbf{c}_i(\mathbf{x}^{\text{old}})\|_2$. Here, $\mathbf{x}^{\text{old}}$ is the estimate for the code vector in the current iteration. Moreover, since the computation of the number of visible vertices $|\mathcal{V}|$ is expensive (since it would require an additional pass over the vertices), here it is approximated with N. The loss function can now be represented as a sum of squares of individual residuals, i.e., $E_{\text{loss}}(\mathbf{x}) = \mathbf{F}^T(\mathbf{x})\mathbf{F}(\mathbf{x})$, where $\mathbf{F} : \mathbb{R}^{257} \rightarrow \mathbb{R}^M$ is a vector-valued function such that $\mathbf{F}(\mathbf{x})$ contains all the $M = |V|+46+80+80+64$ residuals of the energy (Eq. 4.7). For obtaining high performance, the computation of $\mathbf{F}$ is parallelized to exploit the data-parallel computing power of modern graphics cards, i.e., all elements of the vector $\mathbf{F}$ are computed fully in parallel (each entry by a dedicated thread). In the forward pass, $E_{\text{loss}} = \mathbf{F}^T\mathbf{F}$ is computed using block reductions. The local dot product in each block is computed using shared memory and thread synchronization. Results from different blocks are added on the CPU. In the backward pass, the gradients of E_{loss} can be calculated as

$$\frac{\mathrm{d}E_{\text{loss}}(\mathbf{x})}{\mathrm{d}\mathbf{x}} = 2\mathbf{J}^T(\mathbf{x})\mathbf{F}(\mathbf{x}) \ , \tag{4.12}$$

where $\mathbf{J}(\mathbf{x}) \in \mathbb{R}^{M \times 257}$ is the Jacobian of $\mathbf{F}$ at $\mathbf{x}$. $\mathbf{J}$ is computed similarly to $\mathbf{F}$ by using one thread per entry of the matrix. The dense matrix-vector multiplication can be interpreted as computing a dot product for each element of $\mathbf{x}$, which is done similarly to the forward pass. The updated mesh (geometry and albedo) for the next forward-backward pass is computed based on a matrix-vector multiplication and we use one thread per entry of the output vector.

4.6 STOCHASTIC SAMPLING

Since MoFA depends on several parameters (the weighting of the individual energy terms, the relative learning rates for different output parameters, and other network hyper-parameters), finding a good configuration is a repetitive task that requires several (user-guided) iterations. The bottleneck of this procedure is the relatively long training time of the network. In order to speed-up this process, we make use of a new

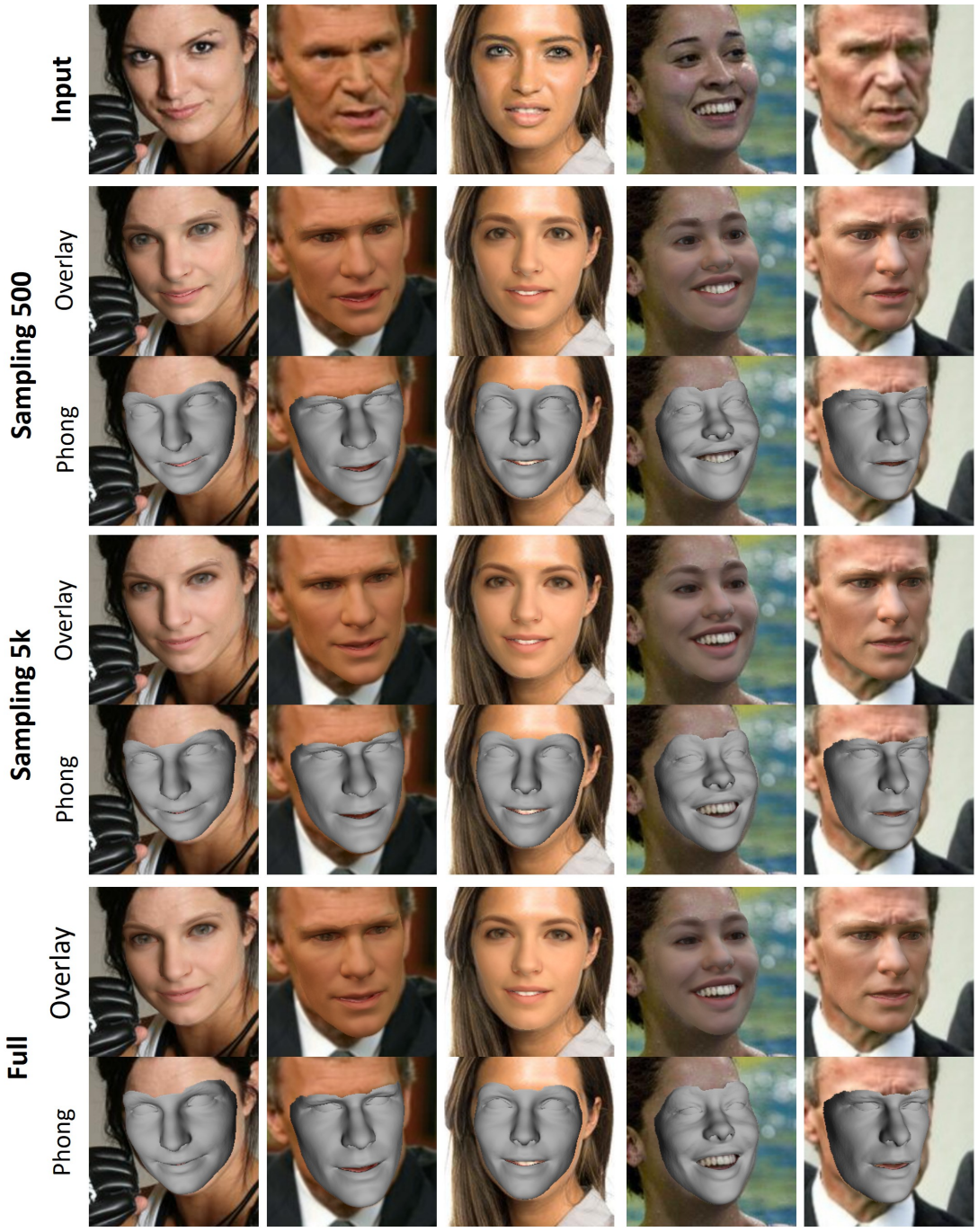

Figure 4.4: Qualitative comparison of MoFA with and without stochastic sampling. The stochastic sampling of vertices lets us train networks much faster with comparable results to networks trained using all vertices.

stochastic sampling strategy. The basic idea is to randomly sample a small subset of vertices for each input image and then merely backpropagate the error for this small set. To be more specific, the energy E_{photo} in (4.7) os defined for a subset of sampled vertices $\mathbf{S} \subseteq \mathbf{V}$. Let $\mathcal{S} \subseteq \mathbf{S}$ be the subset of visible vertices. The loss is then defined as the sum of model-vertex-specific energy terms E^i_{photo}, i.e.,

$$E_{\text{photo}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} E^i_{\text{photo}}\,, \tag{4.13}$$

where

$$E^i_{\text{photo}} := \|\mathcal{I}(\mathbf{u}_i(\mathbf{x})) - \mathbf{c}_i(\mathbf{x})\|_2. \tag{4.14}$$

This energy is implemented similarly as in Eq. 4.11. Using this sampling strategy for training can be interpreted as stochastic gradient descent not only over the set of images in the training set, but also over the face vertices. Note that since the semantically defined code vector has global influence, i.e., each vertex of the reconstruction influences all parameters of the network through E^i_{photo}, this is a valid sampling strategy.

EVALUATION OF THE STOCHASTIC SAMPLING The sampling strategy is quantitatively evaluated in Fig. 4.3 for different numbers of samples used while training. As can be seen, sampling fewer vertices only marginally reduces the quality of the results, while enabling us to train the networks much faster (time taken to train the network with 500, 2000, 5000 and all the vertex samples are 4, 5.2, 7.7 and 23.7 hours, respectively, using a GeForce TitanX graphics card). Qualitative results are shown in Fig. 4.4, where it can be seen that the resulting rendered images have similar visual quality.

4.7 RESULTS OF MOFA

In this section unsupervised learning of the model-based autoencoder, and the improvement in accuracy due to a surrogate loss are demonstrated in-the-wild. Encoders based on AlexNet (Krizhevsky et al., 2012) and VGG-Face (Parkhi et al., 2015) are tested, where the last fully connected layer is modified to output the 257 model parameters. The reported results have been obtained using AlexNet (Krizhevsky et al., 2012) as encoder. Note that the surrogate loss is not employed and all the vertices of the mesh are used (i.e., no stochastic sampling) unless stated otherwise. After training, the encoder regresses pose, shape, expression, skin reflectance and illumination at once from a single image, see Fig. 4.5. An image corpus (see Fig. 4.6) is used for training, which is a combination of four datasets: CelebA (Liu et al., 2015), LFW (Huang et al., 2007), Facewarehouse (Cao et al., 2013), and 300-VW (Chrysos et al., 2015; Shen et al., 2015; Tzimiropoulos, 2015). The corpus is automatically annotated using facial landmark detection (see Sec. 4.5) and cropped to a bounding box using Haar Cascade Face Detection (Bradski, 2000). Frames with bad detections are pruned. The crops are scaled to a resolution of 240×240 pixels. In total, 147k images are collected, which are randomized and split into 142k for training and 5k for evaluation. The network is trained using the *Caffe* (Jia et al., 2014) deep learning framework. For efficiency, the

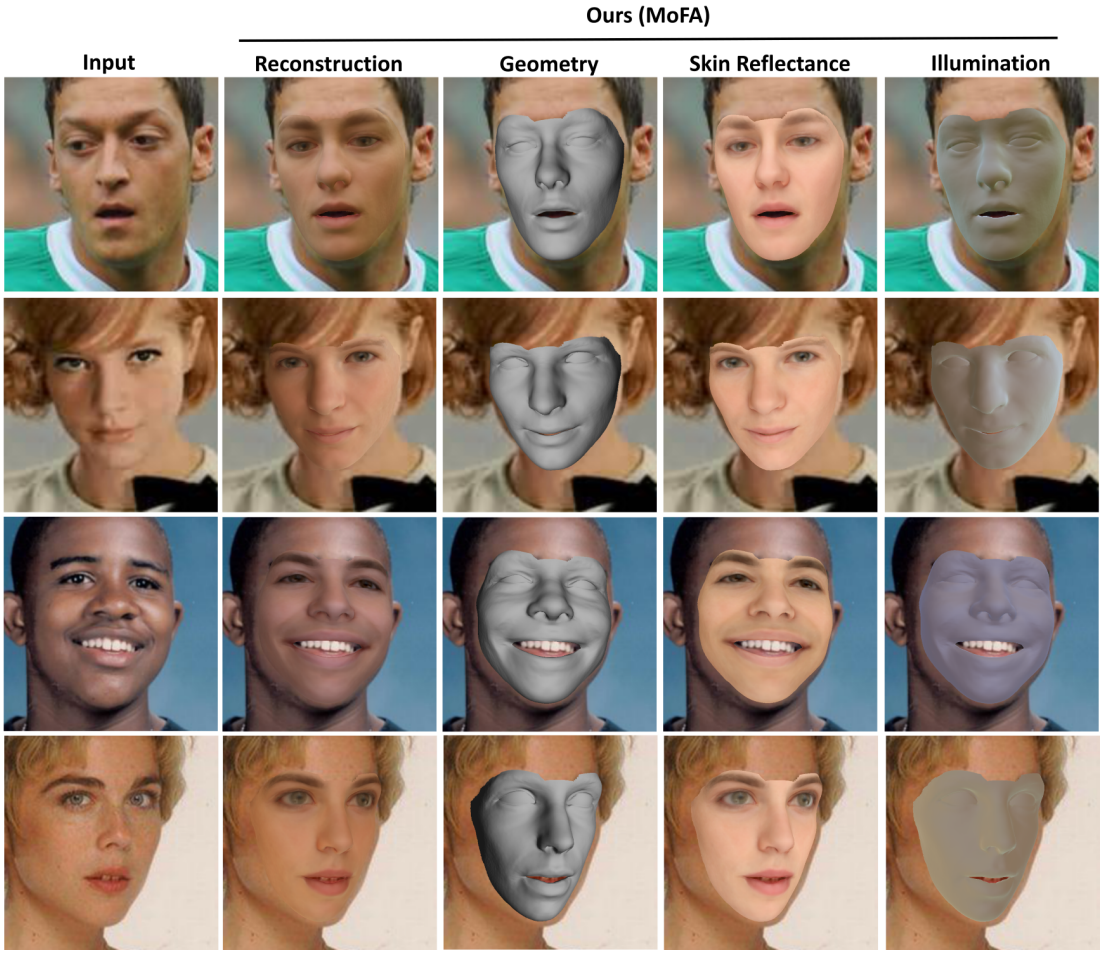

Figure 4.5: The proposed approach enables the regression of high quality pose, shape, expression, skin reflectance and illumination from just a single monocular image (images from CelebA (Liu et al., 2015)).

Figure 4.6: Sample images of the real world training corpus.

Figure 4.7: Comparison to Richardson et al. (2016, 2017) on 300-VW (Chrysos et al., 2015; Shen et al., 2015; Tzimiropoulos, 2015) (left) and LFW (Huang et al., 2007) (right). MoFA obtains higher reconstruction quality and provides estimates of colored reflectance and illumination. Note, in Richardson et al. (2016, 2017) the grayscale reflectance is not regressed but obtained via optimization. MoFA on the other hand regresses all parameters (including reflectance) at once.

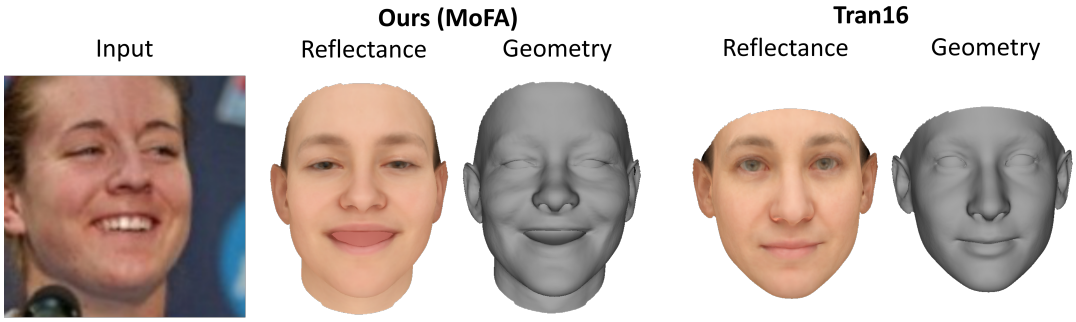

Figure 4.8: Comparison to Tuan Tran et al. (2017) on LFW (Huang et al., 2007). MoFA obtains visually similar quality. Here, the full face model is shown, but training only uses the frontal part (cf. Fig 4.2, right).

model-based decoder and the robust photometric loss are implemented in a single CUDA (NVIDIA, 2008) layer. The networks are trained using *AdaDelta* with 200k batch iterations (batch size of 5). The base learning rate is 0.1 for all parameters, except for the Z-translation that was set to 0.0005. At test time, regressing all parameters using a TitanX Pascal graphics card is fast and takes only 4ms (AlexNet) or 14ms (VGG-Face). Training takes 13 hours (AlexNet) or 20 hours (VGG-Face). The encoder is initialized based on the provided pre-trained weights. All weights in the last fully connected layer are initialized to zero. This guarantees that the initial prediction is the average face placed in the middle of the screen and lit by ambient light, which is a good initialization. Note, the ambient coefficients of our renderer have an offset of 0.7 to guarantee that the scene is initially lit. Next, the method is compared to the state-of-the-art optimization- and learning-based monocular reconstruction approaches, and all its components are evaluated.

COMPARISON TO RICHARDSON ET AL. (2016, 2017) The approach is compared to the CNN-based iterative regressor of Richardson et al. (2016, 2017). The results are compared qualitatively (Fig. 4.7) and quantitatively (Fig. 4.16) to their coarse regression network. Note, the refinement layer of Richardson et al. (2017) is orthogonal to the proposed approach. Unlike Richardson et al. (2016, 2017), the proposed network is trained completely unsupervised on real images, while they use a synthetic training corpus that lacks realistic features. In contrast to Richardson et al. (2016, 2017), the method also regresses colored skin reflectance and illumination, which is critical for many applications, for example, relighting. Note, the grayscale reflectance of Richardson et al. (2016, 2017) is not regressed, but obtained via optimization.

COMPARISON TO TUAN TRAN ET AL. (2017) Fig. 4.8 qualitatively compares to the CNN-based identity regression approach of Tuan Tran et al. (2017). The reconstructions of the proposed method are of visually similar quality; however, with additional high quality estimates of the facial expression and illumination. A face verification test on LFW if also performed. The proposed approach obtains an accuracy of 77%, which is higher than the monocular 3DMM baseline (Romdhani and Vetter, 2005) (75%). Tuan Tran et al. (2017) report an accuracy of 92%. The proposed approach is not designed for this scenario, since it is trained unsupervised on in-the-wild images. Tuan Tran et al. (2017) require more supervision (photo collection) to train their network.

COMPARISON TO THIES ET AL. (2016B) MoFA is compared qualitatively (Fig. 4.9) and quantitatively (Fig. 4.16) to the state-of-the-art optimization-based monocular reconstruction approach of Thies et al. (2016b). MoFA obtains similar or even higher quality, while being orders of magnitude faster (4ms vs. $\approx$ 500ms). Note, while Thies et al. (2016b) tracks at real-time frame rates after identity estimation, it requires half a second to fit all parameters starting from the average model. While MoFA only requires face detection at test time, Thies et al. (2016b) require detected landmarks.

COMPARISON TO GARRIDO ET AL. (2016A) MoFA is compared to our own implementation (no detail refinement and shape correctives, photometric + landmark + regularization terms, 50 Gauss-Newton steps) of the high quality off-line monocular reconstruction approach of Garrido et al. (2016a), which requires landmarks as input. MoFA obtains comparable quality, while requiring no landmarks, see Fig. 4.10 and

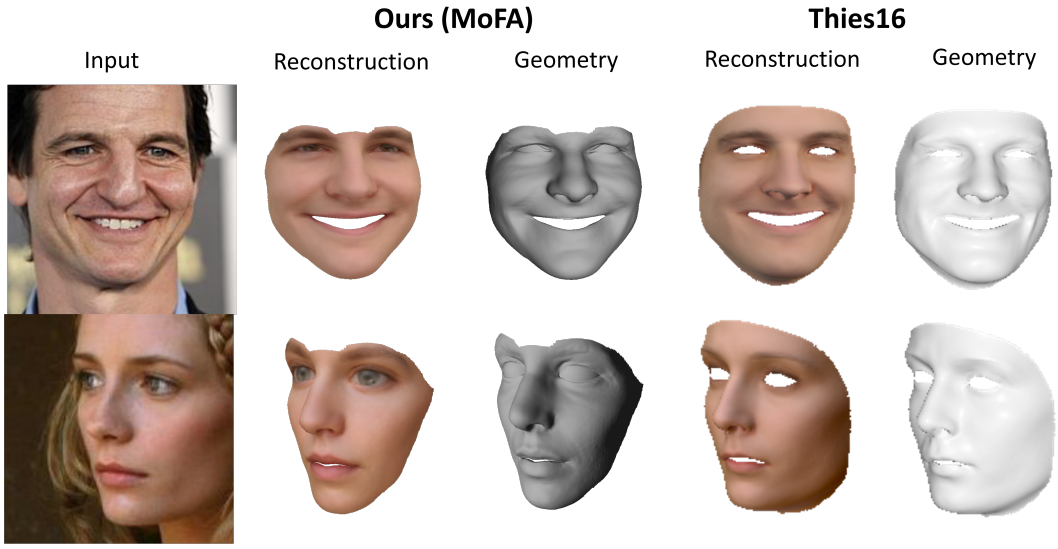

Figure 4.9: Comparison to the monocular reconstruction approach of Thies et al. (2016b) on CelebA (Liu et al., 2015). MoFA obtains similar or higher quality, while being orders of magnitude faster (4ms vs. ≈ 500ms).

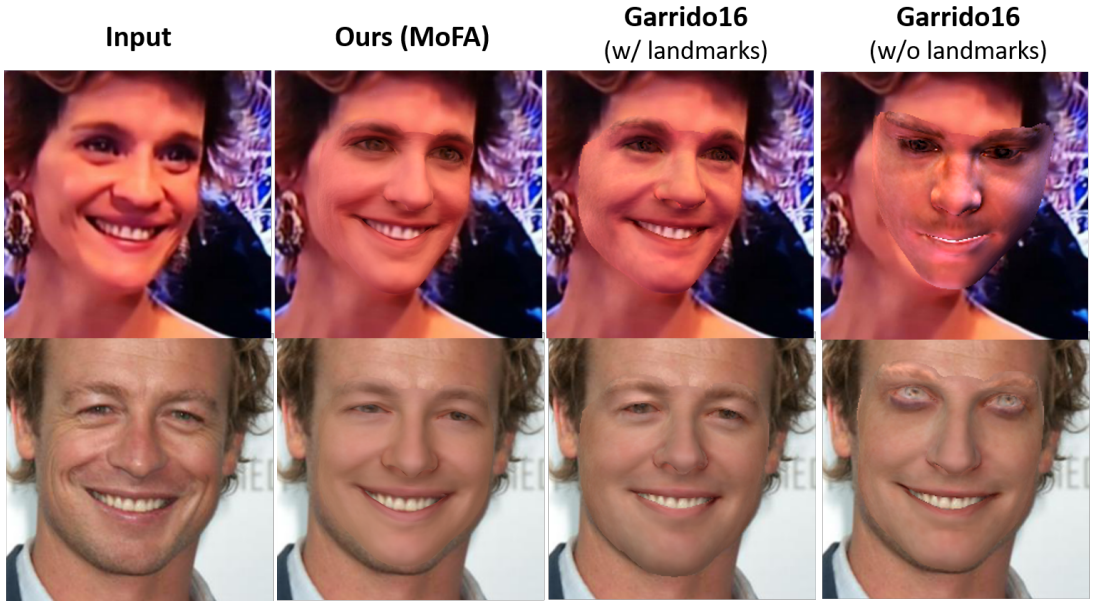

Figure 4.10: Comparison to our implementation of the high quality offline monocular reconstruction approach of Garrido et al. (2016a). MoFA obtains similar quality without requiring landmarks as input. Without landmarks, Garrido et al. (2016a) often gets stuck in a local minimum.

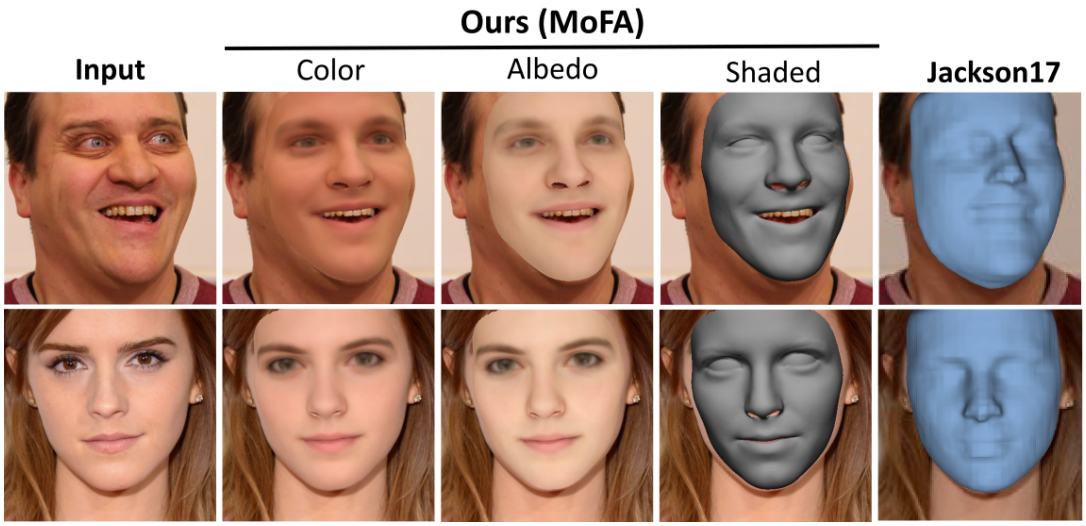

Figure 4.11: Comparison to Jackson et al. (2017). MoFA obtains higher quality reconstructions while also estimating the reflectance and incident scene illumination.

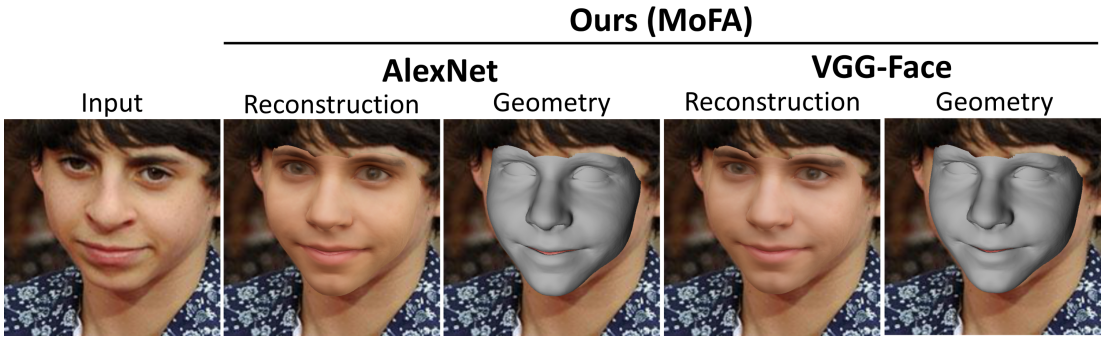

Figure 4.12: Different encoders are evaluated in combination with our model-based decoder. Overall, VGG-Face (Parkhi et al., 2015) leads to slightly better results than AlexNet (Krizhevsky et al., 2012), though the results are comparable.

Fig. 4.16. Without sparse constraints as input, optimization-based approaches often get stuck in a local minimum.

MoFA is also compared to the monocular CNN-based approach of Jackson et al. (2017) (Fig. 4.11). It obtains qualitatively better alignments and higher quality results.

EVALUATION OF DIFFERENT ENCODERS The impact of different encoders is also evaluated. VGG-Face (Parkhi et al., 2015) leads to slightly better results than AlexNet (Krizhevsky et al., 2012), see Fig. 4.12. On average, VGG-Face (Parkhi et al., 2015) has a slightly lower landmark (4.9 pixels vs. 5.3 pixels) and photometric error (0.073 vs. 0.075, color distance in RGB space, each channel in $[0, 1]$), see Fig. 4.13.

QUANTITATIVE EVALUATION OF UNSUPERVISED TRAINING Unsupervised training decreases the dense photometric and landmark error (on a validation set of 5k real images), even when landmark alignment is not part of the loss function, see Fig. 4.13. The landmark error is computed based on 46 detected landmarks (Saragih et al., 2011). Training with our surrogate loss improves landmark alignment (AlexNet: 3.7 pixels

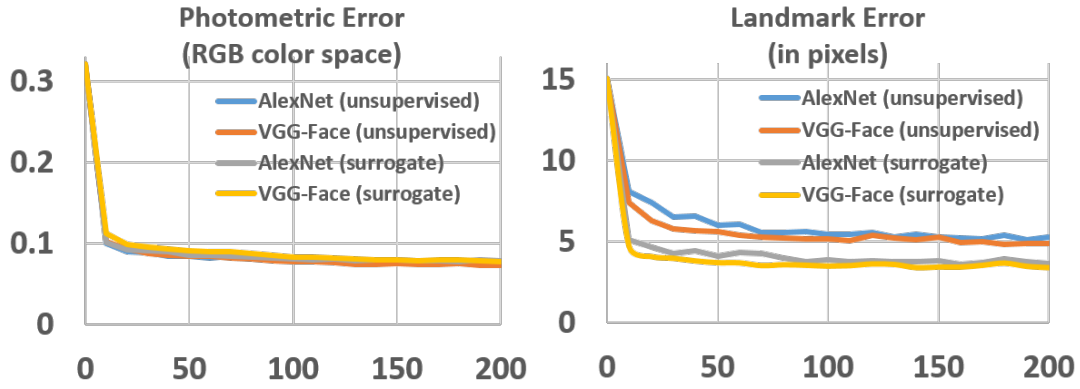

Figure 4.13: Quantitative evaluation of MoFA on real data: Both landmark and photometric errors are decreased during unsupervised training, even though landmark alignment is not part of the loss function.

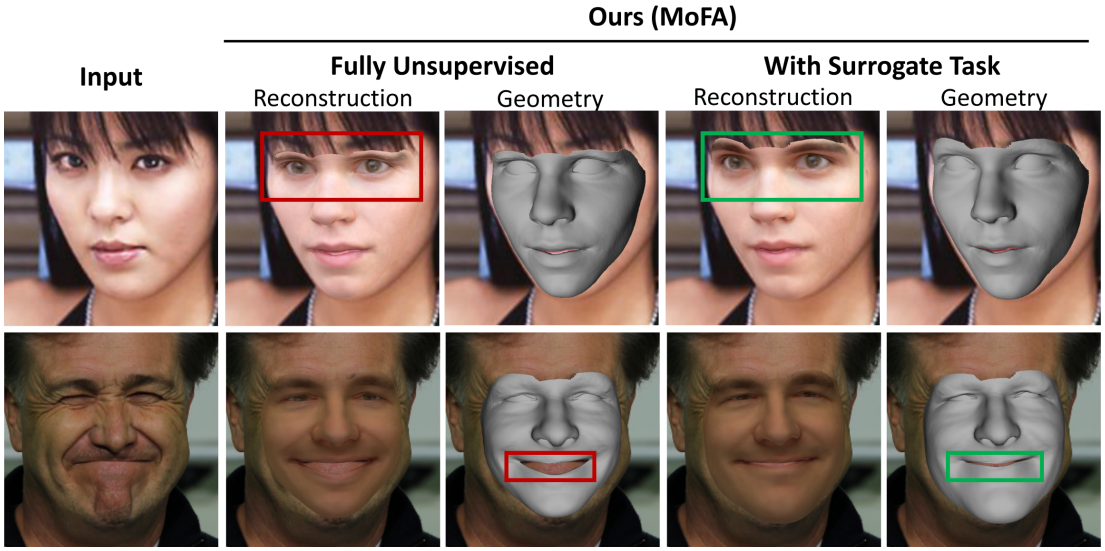

Figure 4.14: Evaluation of the influence of the proposed surrogate task. The surrogate task leads to improved reconstruction quality and increases robustness to occlusions and strong expressions.

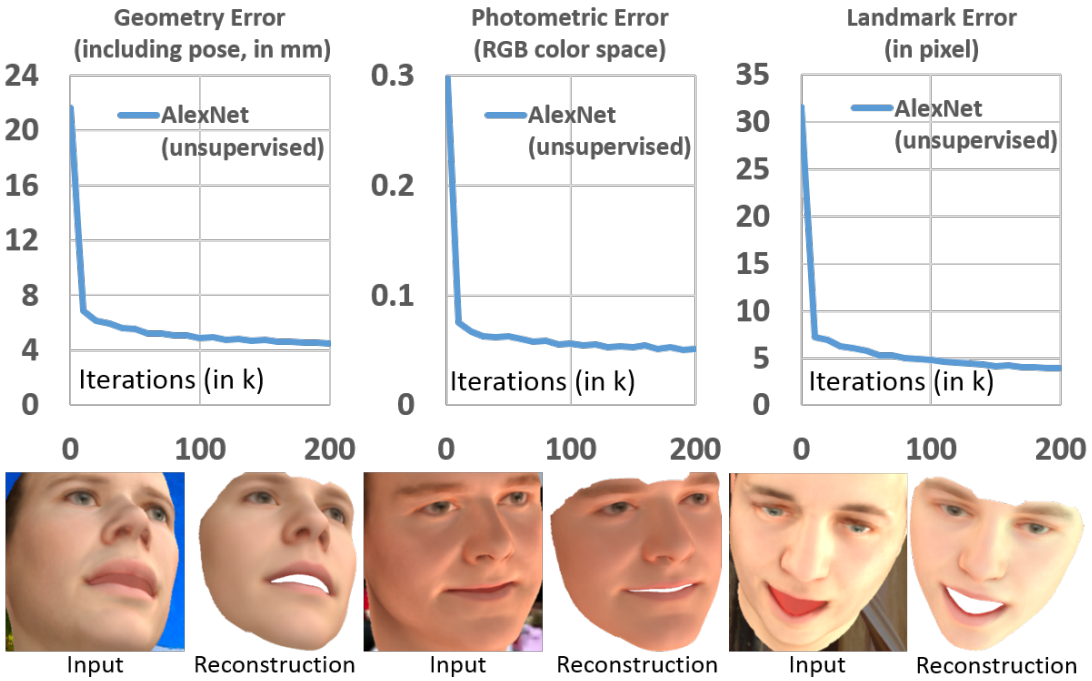

Figure 4.15: Quantitative evaluation of MoFA on synthetic ground truth data: Training decreases the geometric, photometric and landmark errors.

Table 4.1: Quantitative evaluation on real data. Average Hausdorff distance to the ground truth for different approaches.

	Geometry	Photometric	Landmark
Ours (MoFA w/o surrog.)	1.9mm	0.065	5.0px
Ours (MoFA w/ surrog.)	1.7mm	0.068	3.2px
Garrido et al. (2016a)	1.4mm	0.052	2.6px

vs. 5.3 pixels, VGG-Face: 3.4 pixels vs. 4.9) and leads to a similar photometric error (AlexNet: 0.078 vs. 0.075, VGG-Face: 0.078 vs. 0.073, color distance in RGB space, each channel in $[0, 1]$). The influence of our landmark-based surrogate loss is also evaluated qualitatively, see Fig. 4.14. Training with landmarks helps to improve robustness to occlusions and the quality of the predicted expressions. Note that both scenarios do not require landmarks at test time.

QUANTITATIVE EVALUATION A ground truth evaluation is performed based on 5k rendered images with known parameters. The model-based autoencoder (AlexNet, unsupervised) is trained on a corpus of 100k synthetic images with background augmentation (cf. Fig. 4.15). The geometric error is measured as the point-to-point 3D distance (including the estimated rotation, translation and isotropic scale are compensated for) between the estimate and the ground truth mesh. This error drops from 21.6mm to 4.5mm. The photometric error in RGB space also decreases (0.33 to 0.05) and so does the landmark error (31.6 pixels to 3.9 pixels). Overall, we obtain good fits. A quantitative comparison is also performed for 9 identities (180 images) on Facewarehouse, see Table 4.1 and Fig. 4.16. MoFA obtains low errors and on par with optimization-based techniques in terms of Hausdorff distance, but it is much

Table 4.2: Geometric error on 180 meshes of the FaceWarehouse (Cao et al., 2013) dataset. Surface-to-surface error (including sliding) based on a precomputed dense correspondence map between the employed test set and our mesh topology.

	Ours		Others			
	MoFA (surrogate)	Opt	Tewari et al. (2018) (Fine)	Tewari et al. (2018) (Coarse)	Kim et al. (2018b)	Garrido et al. (2016a) (Coarse)
Mean	2.19 mm	1.87 mm	1.84 mm	2.03 mm	2.11 mm	1.59 mm
SD	0.54 mm	0.42 mm	0.38 mm	0.52 mm	0.46 mm	0.30 mm
Time	4 ms	110 ms	4 ms	4 ms	4 ms	$>$ 1 min

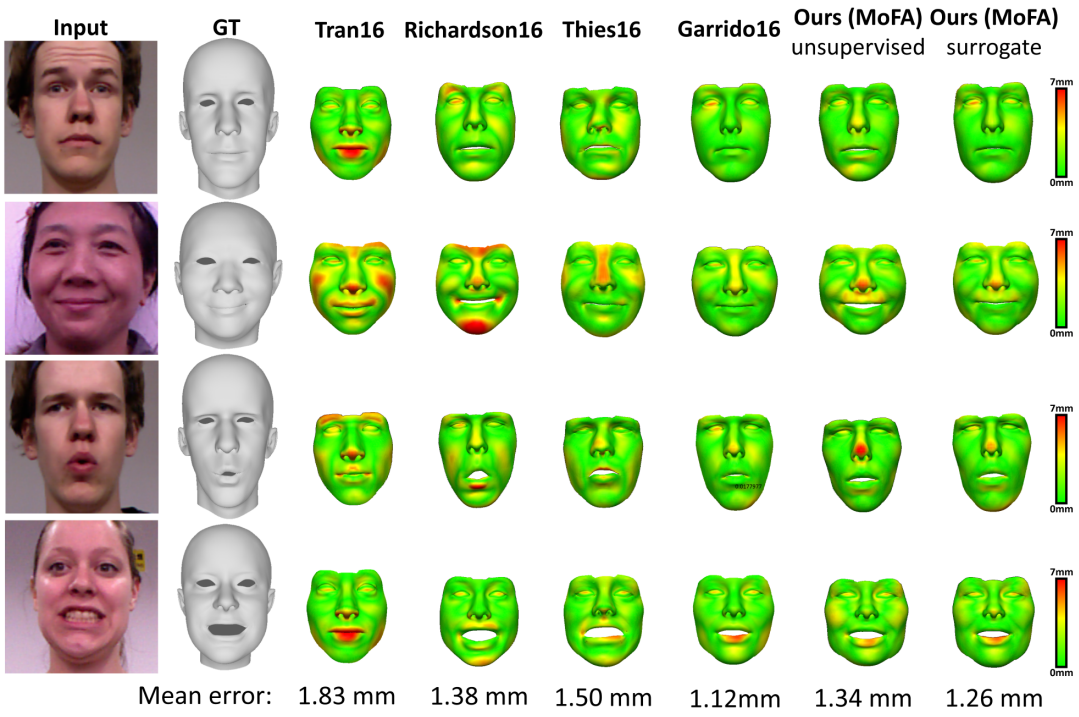

Figure 4.16: Quantitative evaluation on Facewarehouse (Cao et al., 2013): MoFA obtains a low error that is comparable to optimization-based approaches. For this test, the network is trained using the intrinsics of the Kinect.

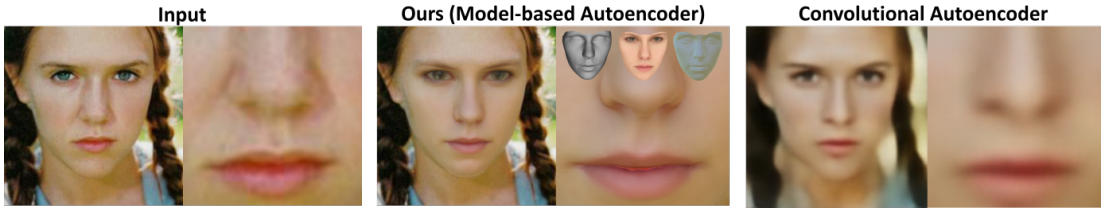

Figure 4.17: MoFA gives results of higher quality than convolutional autoencoders. In addition, it provides access to dense geometry, reflectance, and illumination.

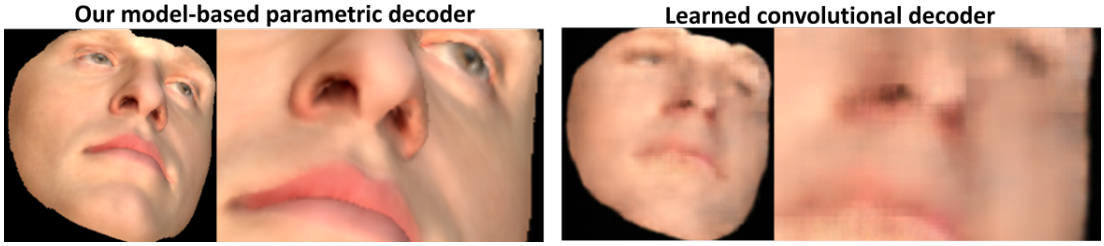

Figure 4.18: The model-based decoder provides higher fidelity than a learned convolutional decoder in terms of image quality.

faster (4ms vs. a few minutes) and requires no landmarks at test time. The Hausdorff distance error metric does not penalize misalignments in the tangent plane (surface sliding). To also quantitatively evaluate the reconstructions in terms of surface drift, a dense correspondence map between the employed test set and the mesh topology of MoFA is precomputed using a non-rigid registration approach. The correspondences are computed based on two almost neutral meshes with a slightly open mouth (to not erroneously bring the upper lip of one topology into correspondence with the lower lip of the other mesh). Based on this fixed set of correspondences, additional evaluation of the surface-to-surface error (including surface sliding) is performed on the same test set, see Table. 4.2. The results are comparable to the very recent coarse-level results of Tewari et al. (2018) (presented in Chapter 5) and Kim et al. (2018b). Our refined results, see Sec. 4.8, outperform these two other state-of-the-art learning-based techniques on the coarse level. The results of Garrido et al. (2016a) are still slightly better, but our approach runs orders of magnitude faster.

COMPARISON TO AUTOENCODERS AND LEARNED DECODERS The model-based decoder is compared with a convolutional autoencoder in Fig. 4.17. The autoencoder uses four 3×3 convolution layers (64, 96, 128, 256 channels), a fully connected layer (257 outputs, same as the number of model parameters), and four 4×4 deconvolution layers (128, 96, 64, 3 channels). The model-based approach obtains sharper reconstruction results and provides fine granular semantic parameters, allowing access to dense geometry, reflectance and illumination, see Fig. 4.17 (middle). Explicit disentanglement (Grant et al., 2016; Kulkarni et al., 2015) of a convolutional autoencoder requires labeled ground truth data. It is also compared to image formation based on a trained decoder. To this end, the decoder is trained (similar parameters as above) based on synthetic imagery generated by the parametric model to learn the parameter-to-image mapping. The model-based decoder obtains renderings of higher fidelity compared to the learned decoder, see Fig. 4.18.

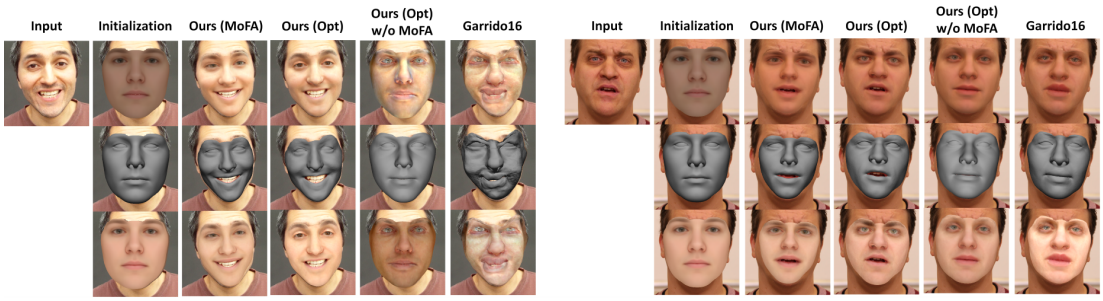

Figure 4.19: Qualitative comparison between MoFA and MoFA with analysis-by-synthesis optimization (Opt) without the landmark term. Opt improves the MoFA estimates while Garrido et al. (2016a), which starts from a neutral initialization (second column), often ends up in local minima in the absence of landmarks. Opt, when starting from a neutral initialization also fails to estimate plausible reconstructions.

4.8 OPTIMIZATION-BASED REFINEMENT

Similar to other data-driven techniques, neural networks have a limited capacity and might not generalize well to inputs outside the span of the employed training corpus. Finding the right balance between under- and over-fitting is a highly challenging problem on its own. Under-fitting leads to a loss of reconstruction quality and over-smoothed results, while over-fitting leads to bad generalization to unseen images. On the other hand, standard optimization-based approaches (without the guidance of discriminative detected landmarks) often get stuck in a bad local minimum, which leads to low reconstruction quality, as shown in Fig. 4.19. In this section, the combination of a coarse discriminative estimate with an optimization-based analysis-by-synthesis approach and a shading-based surface refinement step is demonstrated to significantly improve the quality of the obtained reconstructions. First, a local minimization of the energy E_{loss} is described in Eq. (4.7) based on the Gauss-Newton method, which leads to an improved reconstruction that remains within the span of the employed model (Sec. 4.3). Moreover, in order to explain fine-scale details on a wrinkle-level, (local) optimization of a modified energy function over per-vertex displacements is performed. These displacements are able to represent faces that are outside the (restricted) model-subspace.

4.8.1 *Analysis-by-synthesis Optimization*

Since the trained network has limited capacity, it has to trade-off the quality of individual reconstructions in order to work on a diverse range of images. Running an analysis-by-synthesis optimizer on the output of MoFA can significantly improve the results. The optimizer minimizes the energy E_{loss} in (4.7) as used to train the network.

Starting from the MoFA output as initialization, Gauss-Newton optimization is run. Since the Gauss-Newton method requires the energy to be represented as a sum of squares, the photometric term in (4.8) is implemented as explained in (4.11). Additionally, the optimizer is implemented in a data-parallel fashion on the GPU, as explained next.

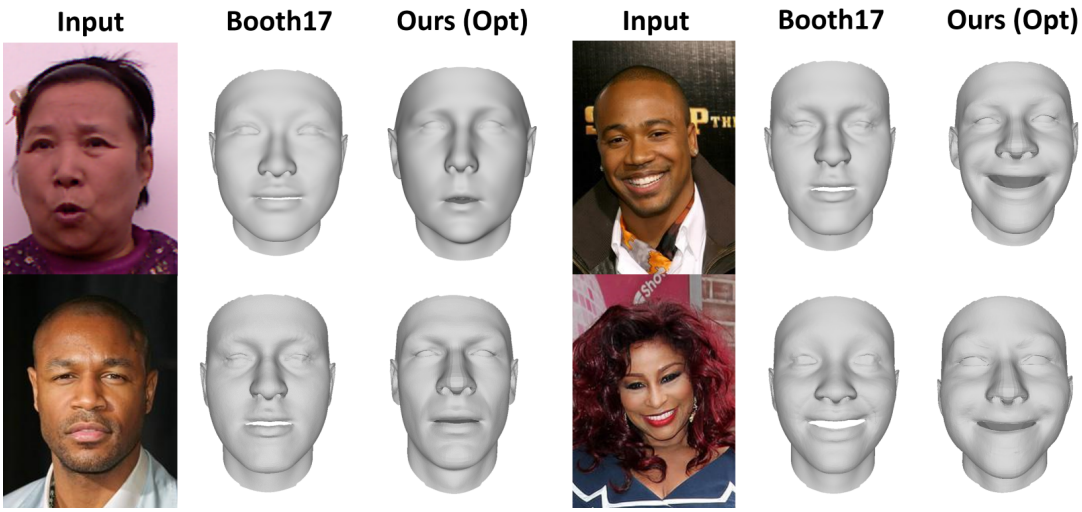

Figure 4.20: Comparison between Opt and the approach by Booth et al. (2017), which learns an in-the-wild texture model from images to improve the reconstruction of geometry. Opt obtains similar or better quality results only using the reflectance model of Blanz and Vetter (1999).

DATA-PARALLEL GPU IMPLEMENTATION The face reconstruction energy is in a general non-linear least-squares form:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}\, \mathbf{E}_{\text{loss}}(\mathbf{x}), \text{ where} \tag{4.15}$$

$$\mathbf{E}_{\text{loss}}(\mathbf{x}) = \sum_i \left(\mathbf{F}_i(\mathbf{x})\right)^2 \ . \tag{4.16}$$

Thus, the (local) optimum $\mathbf{x}^*$ is obtained using the Gauss-Newton algorithm. In each iteration step, the problem is linearized based on Taylor expansion to solve the resulting normal equations:

$$\mathbf{J}^T\mathbf{J}\mathbf{ffi} = \mathbf{J}^T\mathbf{F} \ . \tag{4.17}$$

$\mathbf{J}$ and $\mathbf{F}$ are the same as defined in Sec. 4.5 and are computed in the same manner. $\mathbf{ffi}$ is the optimal update of the parameters. A data-parallel implementation (nVidia, 2012) of dense matrix-matrix and matrix-vector multiplication is used to compute the system matrix $\mathbf{J}^T\mathbf{J}$ and right hand side $\mathbf{J}^T\mathbf{F}$ of Eq. (4.17), respectively. Afterwards, the resulting small linear system is copied to the CPU and solved via Cholesky factorization to compute the optimal update δ. This process is iterated for 5 Gauss-Newton steps. The runtime to obtain the final reconstructions (network inference + optimizer) is 110 ms for one image, orders of magnitude faster compared to a few mins per image for Garrido et al. (2016a).

RESULTS The combination of the discriminative approach with this analysis-by-synthesis fitting strategy (henceforth referred to as "Opt") leads to higher quality results, as shown in Figs. 4.19, 4.21, 4.22 and Table 4.2. Purely optimization-based approaches are highly sensitive to the initialization and often fall into local minima in the absence of the landmark alignment term. The parameter regression result of MoFA provides a good initialization that can reliably be refined by the local optimizer such that good reconstructions can be obtained even without landmarks, cf. Fig. 4.19. Note, all results obtained with the optimizer other than Fig. 4.19 use the MoFA network with

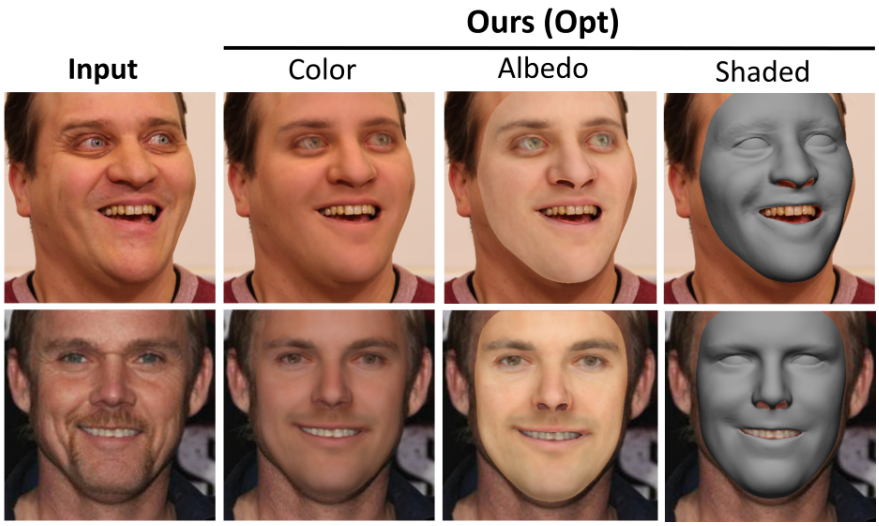

Figure 4.21: MoFA with analysis-by-synthesis optimization allows for high-quality geometry and appearance reconstructions.

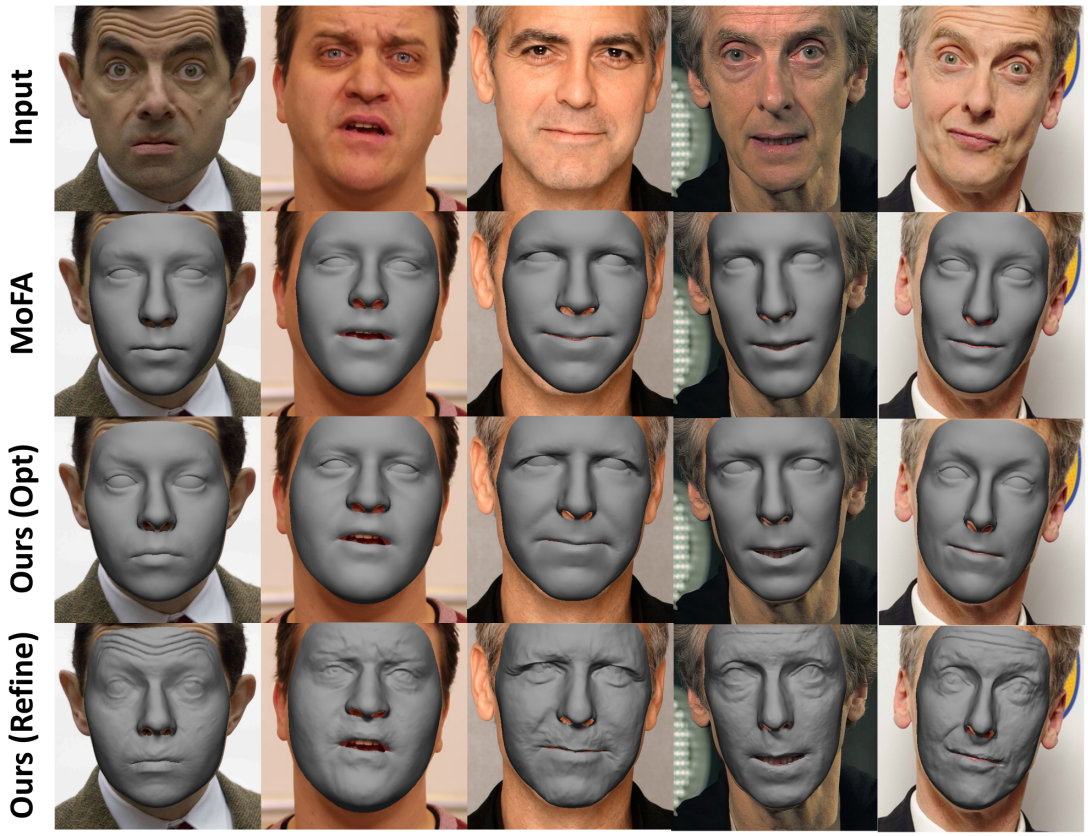

Figure 4.22: Qualitative comparison of MoFA with and without refinement. While MoFA provides good reconstructions, the analysis-by-synthesis optimization (Opt) significantly improves reconstruction quality. Shading-based-refinement (Refine) further adds high-frequency details on the surface, leading to high-fidelity reconstructions.

the surrogate loss and the landmark alignment term for higher quality results. The weights used for the optimizer are $w_{\text{photo}} = 0.44$, $w_{\text{reg}} = 0.01$, $w_{\boldsymbol{\beta}} = 0.11$, $w_{\boldsymbol{\delta}} = 0.01$. Using only 5 Gauss-Newton iterations leads to significant improvements over MoFA.

Figure 4.23: Comparison of our method with shading-based surface refinement (Refine), Richardson et al. (2017) and Sela et al. (2017). Richardson et al. (2017) only estimate the refined depth maps while Sela et al. (2017), need an expensive non-rigid template alignment step to compute the final reconstructions. The proposed approach obtain similar or higher quality reconstructions by directly optimizing for the surface details on the mesh.

Table 4.2 provides quantitative results comparing various methods, where it can be seen that Opt is able to reduce the MoFA reconstruction error further. Although Opt is not able to outperform the results achieved by Garrido et al. (2016a), note that Garrido et al. (2016a) runs for 50 iterations (with the landmark alignment term, starting from a neutral face), thus requiring significantly more time. The Opt approach is compared with the optimization-based approach of Booth et al. (2017) (Fig. 4.20) where the proposed approach obtains comparable or better results. It also provides individual estimates for the reflectance and illumination channels while Booth et al. (2017) only estimates the combined texture.

4.9 SHADING-BASED SURFACE REFINEMENT

The results presented so far are limited to the subspace spanned by the underlying low-dimensional affine model (Sec. 4.3). This limits the ability of the method to capture fine-scale wrinkle-level details. Hence, the output of Opt is further refined by allowing the mesh to go outside of this restricted low-dimensional deformation space. Consider the vertex positions of the low-dimensional coarse reconstruction as $\mathbf{V}^{\mathrm{C}} = \{\mathbf{v}_i^{\mathrm{C}} \in \mathbb{R}^3 | 1 \leq i \leq N\}$. Out-of-subspace deformations are modeled using per-vertex displacements $\mathbf{D} = \{\mathbf{d}_i \in \mathbb{R}^3 | 1 \leq i \leq N\}$, such that the final vertex positions $\mathbf{V}^{\mathrm{F}} = \{\mathbf{v}_i^{\mathrm{F}} = \mathbf{v}_i^{\mathrm{C}} + \mathbf{d}_i | 1 \leq i \leq N\}$ align well to the input image $\mathcal{I}$. The optimal displacements are determined as

$$\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}}\, E_{\mathrm{ref}}(\mathbf{D}) \ , \tag{4.18}$$

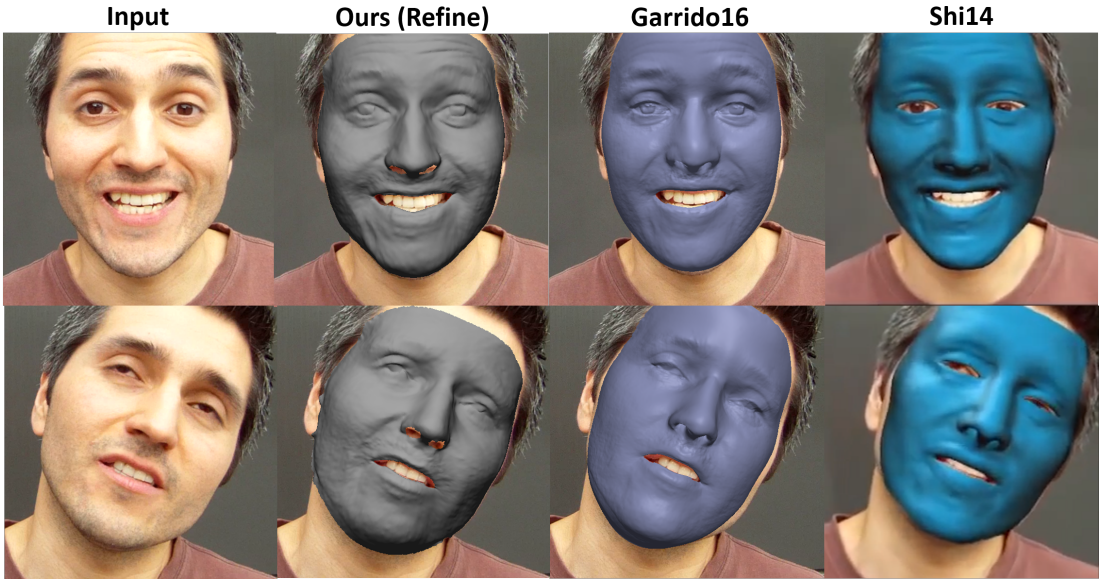

Figure 4.24: Comparison between the proposed method with shading-based surface refinement (Refine), Garrido et al. (2016a) and Shi et al. (2014). Refine obtains similar results, while being significantly faster.

where

$$E_{\text{ref}}(\mathbf{D}) = \underbrace{E_{\text{photo}}(\mathbf{D}) + w_{\text{grad}} E_{\text{grad}}(\mathbf{D})}_{\text{data term}} + \underbrace{w_{\text{reg}} E_{\text{reg}}(\mathbf{D})}_{\text{regularizer}} . \tag{4.19}$$

DENSE PHOTOMETRIC ALIGNMENT Similar to (4.8), a dense photometric alignment term is used

$$E_{\text{photo}}(\mathbf{D}) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left\| \mathcal{I}(\mathbf{u}_i(\mathbf{D})) - \mathbf{c}_i(\mathbf{D}) \right\|_2 , \tag{4.20}$$

where $\mathcal{V}$ is the set of visible vertices (vertex visibility is approximated by the set of front-facing vertices), and $\mathbf{u}_i(\mathbf{D})$ and $\mathbf{c}_i(\mathbf{D})$ are the screen space position and color of vertex i, respectively. They are computed analogously to Eq. (4.5). The photometric term is implemented in an iteratively reweighted fashion, as in Eq. (4.11).

GRADIENT ALIGNMENT TERM High-frequency shading details are also considered, similarly as proposed in Wu et al. (2011). More precisely, a gradient alignment term tries to match the color gradients between the input and a synthetic rendering of the model, as follows:

$$E_{\text{grad}}(\mathbf{D}) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} \left\| (\mathbf{c}_i(\mathbf{D}) - \mathbf{c}_j(\mathbf{D})) - (\mathcal{I}(\mathbf{u}_i(\mathbf{D})) - \mathcal{I}(\mathbf{u}_j(\mathbf{D}))) \right\|_2^2 , \tag{4.21}$$

where $\mathcal{N}_i$ is the one-ring neighborhood of vertex i. Finite differences efficiently approximate image gradients based on mesh gradients.

REGULARIZATION TERM Additionally, a Laplacian regularizer is used on the displacements, as follows:

$$E_{\text{reg}}(\mathcal{D}) = \frac{1}{N} \sum_{i \in \mathbf{V}} \left\| \sum_{j \in \mathcal{N}_i} (\mathbf{d}_i - \mathbf{d}_j) \right\|_2^2 . \tag{4.22}$$

Note that Eq. (4.22) enforces smoothness of the reconstructions and stability of the optimizer.

MESH TOPOLOGY Both MoFA and the analysis-by-synthesis optimization use the topology of Blanz and Vetter (1999), as described in Sec. 4.8. In order to ensure numerical stability for shading-based refinement, one has to take care of near-degenerate mesh faces in the topology of Blanz and Vetter (1999). To this end, the neutral face $\mathbf{V}_N^{T1}$ is remeshed from the topology $T1$ of Blanz and Vetter (1999) to a face $\mathbf{V}_N^{T2}$ represented by a more uniform topology $T2$. The transformation of vertices of topology $T1$ to vertices of topology $T2$ can be represented by the linear map $L\colon \mathbf{V}_N^{T1} \rightarrow \mathbf{V}_N^{T2}$. After analysis-by-synthesis optimization (Sec. 4.8), the results from topology $T1$ are transferred to the topology $T2$ using L, and then optimized over per-vertex displacements using the topology $T2$.

OPTIMIZATION Since the number of unknowns is much larger than for the problem in Sec. 4.8.1, gradient descent is used to optimize for the displacements. Similarly as before, $|\mathcal{V}|$ is approximated in the individual energy terms using N. The weights used in the energy term are $w_{\text{grad}} = 1.0$, $w_{\text{reg}} = 133.3$. The optimization runs for 250 iterations with a step-size of 0.008, which is sufficient to achieve convergence.

DATA-PARALLEL GPU IMPLEMENTATION The per-vertex displacement optimization is also implemented in a data-parallel fashion on the GPU. Since the Jacobian matrix here is much bigger and sparse, the approach from Sec. 4.8.1 is not used. Instead, to compute the gradients, one dedicated thread is launched for each element of $\mathbf{F}$, where thread i computes $\frac{\mathrm{d}(\mathbf{F}_i)^2}{\mathrm{d}\mathbf{D}}$. The gradients for each variable coming from different threads are integrated using global memory atomics. Using this optimized parallel implementation results in a processing time of 450 ms for one image. Thus, the overall time to obtain a high-quality reconstruction including fine-scale details is 450 + 110 = 560 ms.

RESULTS The refinement approach is initialized with the results from Opt. The proposed approach with refinement ("Refine") recovers high-frequency geometry details from images (Fig. 4.22). Comparisons to two high-quality face reconstruction approaches (Richardson et al., 2017; Sela et al., 2017) are shown in Figs. 4.23 and 4.24. Refine obtains details directly on the mesh in contrast to Richardson et al. (2017) that obtain only refined depth maps. Sela et al. (2017) reconstruct details on the mesh but at the cost of an expensive non-rigid template alignment step. The proposed approach obtains similar or higher quality while directly optimizing for the details on the mesh topology. The reflectance and illumination channels are additionally estimated. Similar results are obtained compared to Garrido et al. (2016a) and Shi et al. (2014). However, both Garrido et al. (2016a) and Shi et al. (2014) are orders of magnitude slower.

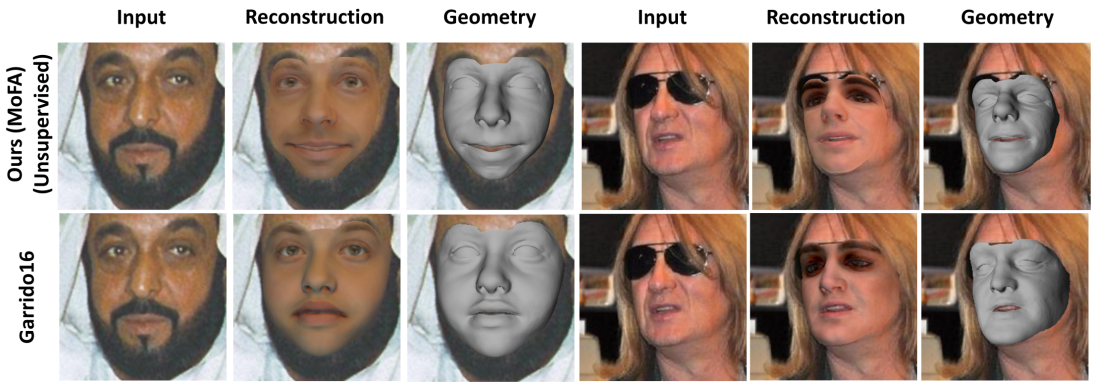

Figure 4.25: Limitations: Facial hair and occlusions are challenging to handle.

4.10 LIMITATIONS

This chapter demonstrated compelling monocular reconstructions using a novel model-based autoencoder that is trained in an unsupervised manner. Similar to other regression approaches, implausible reconstructions are possible with MoFA when the regressed parameters are outside the span of the training data. This can be alleviated by enlarging the training corpus, which is easy to achieve in our unsupervised setting. Since a face model is employed, MoFA reconstructions are limited to the modeled subspace. Similar to optimization-based approaches, strong occlusions, for example, by facial hair or external objects, cause our approach to fail, see Fig. 4.25. Even with the refinement strategies, the proposed approach can fail in such cases. Unsupervised occlusion-aware training is an interesting open research problem. Similar to related approaches, strong head rotations are challenging. Since the background is not modeled, our reconstructions can slightly shrink. Shrinking is discussed and addressed in Schönborn et al. (2015).

4.11 CONCLUSION

This chapter presented a deep convolutional model-based face autoencoder that can be trained in an unsupervised manner and learns meaningful semantic parameters. Semantic meaning in the code vector is enforced by a parametric model that encodes variation along with the pose, shape, expression, skin reflectance and illumination dimensions. The model-based decoder is fully differentiable and allows end-to-end learning of our network. This chapter additionally showed a stochastic vertex sampling strategy in the loss function for faster training, and analysis-by-synthesis optimization and shape-from-shading refinement methods for high-fidelity reconstruction.

The concepts introduced in this chapter will be used throughout the thesis. The integration of CNNs with physically-based rendering is an important concept for self-supervised learning. While the morphable model was pretrained and fixed in this chapter, the next chapter will introduce a method for refining this model in the self-supervised loop. This will bring us a step closer to learning morphable models entirely from 2D data.

5 SELF-SUPERVISED MULTI-LEVEL FACE MODEL LEARNING FOR MONOCULAR RECONSTRUCTION

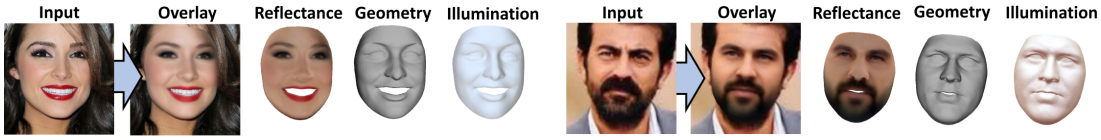

Figure 5.1: This chapter presentes a monocular reconstruction approach which estimates high-quality facial geometry, skin reflectance (including facial hair) and incident illumination at over 250 Hz. A trainable multi-level face representation is learned jointly with the feed forward inverse rendering network. End-to-end training is based on a self-supervised loss that requires no dense ground truth.

The reconstruction of dense 3D models of face geometry and appearance from a single image is highly challenging and ill-posed. To constrain the problem, the previous chapter relied on strong parametric face models learned from 3D scans. However, prior models restrict generalization of the true diversity in facial geometry, skin reflectance and illumination. To alleviate this problem, this chapter (published as Tewari et al. (2018)) presents the first approach that jointly learns 1) a regressor for face shape, expression, reflectance and illumination on the basis of 2) a concurrently learned parametric face model. The multi-level face model combines the advantage of 3D Morphable Models for regularization with the out-of-space generalization of a learned corrective space, see Fig. 5.1. The network is trained end-to-end on in-the-wild images without dense annotations by fusing a convolutional encoder with a differentiable expert-designed renderer and a self-supervised training loss, both defined at multiple detail levels. The proposed approach compares favorably to the state-of-the-art in terms of reconstruction quality, better generalizes to real world faces, and runs at over 250 Hz.

5.1 INTRODUCTION

Monocular face reconstruction has drawn an incredible amount of attention in computer vision and graphics in the last decades. The goal is to estimate a high-quality personalized model of a human face from just a single photograph. Such a model ideally comprises several interpretable semantic dimensions, for example, 3D facial shape and expressions as well as surface reflectance properties. Research in this area is motivated by the increasing availability of face images, for example, captured by webcams at home, as well as a wide range of important applications across several fields, such as facial motion capture, content creation for games and movies, virtual and augmented reality, and communication.

The reconstruction of faces from a single photograph is a highly challenging and ill-posed inverse problem, since the image formation process convolves multiple complex physical dimensions (geometry, reflectance, and illumination) into a single color mea-

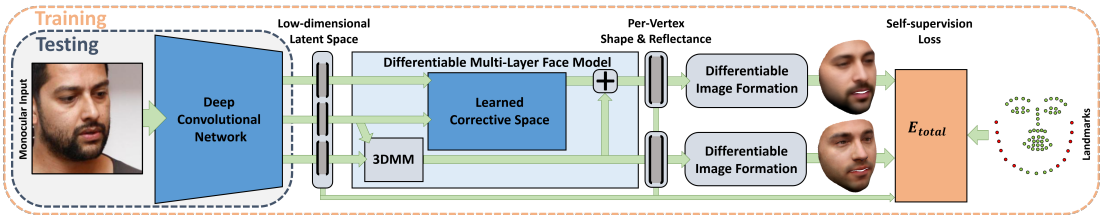

Figure 5.2: The proposed approach regresses a low-dimensional latent face representation at over 250 Hz. The feed forward CNN is jointly learned with a multi-level face model that goes beyond the low-dimensional subspace of current 3DMMs. Trainable layers are shown in blue and expert-designed layers in gray. Training is based on differentiable image formation in combination with a self-supervision loss (orange).

surement per pixel. To deal with this ill-posedness, researchers have made additional prior assumptions, such as constraining faces to lie in a low-dimensional subspace, for example, 3D Morphable Models (3DMM) (Blanz and Vetter, 1999) learned from scan databases of limited size. Many state-of-the-art optimization-based (Blanz et al., 2003, 2004; Garrido et al., 2016a; Saito et al., 2016; Thies et al., 2016b) and learning-based face reconstruction approaches (Cao et al., 2015; Richardson et al., 2016, 2017; Tewari et al., 2017; Tuan Tran et al., 2017) heavily rely on such priors. While these algorithms yield impressive results, they do not generalize well beyond the restricted low-dimensional subspace of the underlying model. Consequently, the reconstructed 3D face may lack important facial details, contain incorrect facial features and not align well to an image. For example, beards have shown to drastically deteriorate the reconstruction quality of algorithms that are trained on pure synthetic data (Richardson et al., 2016, 2017; Sela et al., 2017) or employ a 3DMM for regularization (Blanz and Vetter, 1999; Garrido et al., 2016a; Tewari et al., 2017; Thies et al., 2016b; Tuan Tran et al., 2017). Some approaches try to prevent these failures via heuristics, for example, a separate segmentation method to disambiguate disjunct skin and hair regions (Saito et al., 2016). Recent methods refine a fitted prior by adding fine-scale details, either based on shape-from-shading (Garrido et al., 2016a; Richardson et al., 2016) or pre-learned regressors (Cao et al., 2015; Richardson et al., 2017). However, these approaches rely on slow optimization or require a high-quality annotated training corpus. Besides, they do not build an improved subspace of medium-scale shape, reflectance, and expression, which is critical for generalization. Very recently, Sela et al. (2017) predicted a per-pixel depth map to deform and fill holes of a limited geometry subspace learned during training. While the results are impressive, the non-rigid registration runs offline. Furthermore, their method captures face geometry only and fails if the faces differ drastically from the training corpus, for example, in terms of skin reflectance, and facial hair. Ideally, one would like to build better priors that explain a rich variety of real-world faces with meaningful and interpretable parameters. Learning such models in the traditional way requires large amounts of densely labeled real world data, which is practically infeasible.

This chapter presents an entirely new end-to-end trainable method that jointly learns 1) an efficient regressor to estimate high-quality identity geometry, face expression, and colored skin reflectance, alongside 2) the parameterization of an improved multi-level face model that better generalizes and explains real world face diversity. The method can be trained end-to-end on sparsely labeled in-the-wild images and reconstructs

face and illumination from monocular RGB input at over 250 Hz. The approach takes advantage of a 3DMM for regularization and a learned corrective space for out-of-space generalization. To make end-to-end training on in-the-wild images feasible, a hybrid convolutional auto-encoder is proposed that combines a CNN encoder with a differentiable expert-designed rendering layer and a self-supervision loss, both defined at multiple levels of details. In addition, a novel contour constraint is proposed that generates a better face alignment. Unlike Chapter 4, the proposed auto-encoder learns an improved multi-level model that goes beyond a predefined low-dimensional parametric face prior. Experimental evaluations show that the proposed approach is more robust, generalizes better, and estimates geometry, reflectance, and lighting at higher quality.

5.2 METHOD OVERVIEW

The proposed face reconstruction approach estimates high-quality geometry, skin reflectance and incident illumination from a single image. A regressor is trained for parameters of a multi-level parametric face model, which is also trained concurrently, see Fig. 5.2.

PARAMETER REGRESSION At test time (Fig. 5.2, left), a low-dimensional, yet expressive and discriminative, latent space face representation is computed in under 4ms using a feed forward CNN, for example, AlexNet (Krizhevsky et al., 2012) or VGG-Face (Parkhi et al., 2015). The latent space is based on a novel multi-level face model (Sec. 5.3) that combines a coarse-scale 3DMM with trainable per-vertex geometry and skin reflectance correctives. This enables the approach to go beyond the restricted low-dimensional geometry and skin reflectance subspaces, commonly used by 3DMM-based methods for face fitting.

SELF-SUPERVISED TRAINING The feed forward network is jointly trained (Fig. 5.2, right) with the corrective space based on a novel CNN architecture that does not rely on a densely annotated training corpus of ground-truth geometry, skin reflectance and illumination. To this end, the multi-level model is combined with an expert-designed image formation layer (Sec. 5.4) to obtain a differentiable computer graphics module. To enable the joint estimation of the multi-level face model, this module renders both the coarse 3DMM model and the medium-scale model that includes the correctives. For training, self-supervised loss functions (Sec. 5.5) are employed to enable efficient end-to-end training of our architecture on a large corpus of in-the-wild face images without the need for densely annotated ground truth. The approach is evaluated qualitatively and quantitatively, and compared to state-of-the-art optimization- and learning-based face reconstruction techniques (see Sec. 5.6).

5.3 TRAINABLE MULTI-LEVEL FACE MODEL

At the core of the proposed approach is a novel multi-level face model that parameterizes facial geometry and skin reflectance. The model is based on a manifold template mesh with $N \sim 30$k vertices and per-vertex skin reflectance. The x-, y- and z-coordinates of all vertices $\mathbf{v}_i \in \mathcal{V}$ are stacked in a geometry vector $\mathbf{v}^{\mathrm{f}} \in \mathbb{R}^{3N}$. Sim-

ilarly, a vector of per-vertex skin reflectance is obtained as $\mathbf{r}^{\mathrm{f}} \in \mathbb{R}^{3N}$. Geometry and reflectance are parameterized as follows:

$$\mathbf{v}^{\mathrm{f}}(\mathbf{x}_g) = \mathbf{v}^{\mathrm{b}}(\boldsymbol{\alpha}) + \mathcal{F}_g(\boldsymbol{\delta}_g|\Theta_g) \in \mathbb{R}^{3N}(\text{geometry}), \tag{5.1}$$

$$\mathbf{r}^{\mathrm{f}}(\mathbf{x}_r) = \mathbf{r}^{\mathrm{b}}(\boldsymbol{\beta}) + \mathcal{F}_r(\boldsymbol{\delta}_r|\Theta_r) \in \mathbb{R}^{3N}(\text{reflectance}), \tag{5.2}$$

where $\mathbf{x}_g = (\boldsymbol{\alpha}, \boldsymbol{\delta}_g, \Theta_g)$ and $\mathbf{x}_r = (\boldsymbol{\beta}, \boldsymbol{\delta}_r, \Theta_r)$ are the geometry and reflectance parameters, respectively. At the base level is an affine face model that parameterizes the (coarse) facial geometry $\mathbf{v}^{\mathrm{b}}$ and (coarse) skin reflectance $\mathbf{r}^{\mathrm{b}}$ via a low-dimensional set of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. In addition, correctives are employed to add medium-scale geometry $\mathcal{F}_g$ and reflectance $\mathcal{F}_r$ deformations, parameterized by $(\boldsymbol{\delta}_g, \Theta_g)$ and $(\boldsymbol{\delta}_r, \Theta_r)$, respectively. A detailed explanation will follow in Sec. 5.3.2. A combination of the base level model with the corrective model yields the final level model, parameterizing $\mathbf{v}^{\mathrm{f}}$ and $\mathbf{r}^{\mathrm{f}}$. In the following, the different levels of the multi-level face model are described.

5.3.1 *Static Parametric Base Model*

The parametric face model employed on the base level expresses the space of plausible facial geometry and reflectance via two individual affine models:

$$\mathbf{v}^{\mathrm{b}}(\boldsymbol{\alpha}) = \mathbf{a}_g + \sum_{k=1}^{m_s+m_e} \boldsymbol{\alpha}_k \mathbf{b}_k^g \qquad (\text{geometry}), \tag{5.3}$$

$$\mathbf{r}^{\mathrm{b}}(\boldsymbol{\beta}) = \mathbf{a}_r + \sum_{k=1}^{m_r} \boldsymbol{\beta}_k \mathbf{b}_k^r \qquad (\text{reflectance}) . \tag{5.4}$$

Here, $\mathbf{a}_g \in \mathbb{R}^{3N}$ is the average facial geometry and $\mathbf{a}_r \in \mathbb{R}^{3N}$ the corresponding average reflectance. The subspace of reflectance variations is spanned by the vectors $\{\mathbf{b}_k^r\}_{k=1}^{m_r}$, created using PCA from a dataset of 200 high-quality face scans (100 male, 100 female) of Caucasians (Blanz and Vetter, 1999). The geometry subspace is split into m_s and m_e modes, representing shape and expression variations, respectively. The vectors spanning the subspace of shape variations $\{\mathbf{b}_k^g\}_{k=1}^{m_s}$ are constructed from the same data as the reflectance space (Blanz and Vetter, 1999). The subspace of expression variations is spanned by the vectors $\{\mathbf{b}_k^g\}_{k=m_s+1}^{m_s+m_e}$. These vectors were created using PCA of a subset of blendshapes from the datasets of Alexander et al. (2009) and Cao et al. (2013). Note that these blendshapes have been transferred to the used topology using deformation transfer (Sumner and Popović, 2004). The basis captures 99% of the variance of the used blendshapes. The approach uses $m_s = m_r = 80$ shape and reflectance vectors, and $m_e = 64$ expression vectors. The associated standard deviations $\boldsymbol{\sigma}_g$ and $\boldsymbol{\sigma}_r$ have been computed assuming a normally distributed population. The model parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{80+64} \times \mathbb{R}^{80}$ constitute a low-dimensional encoding of a particular face. Even though such a parametric model provides a powerful prior, its low dimensionality is a severe weakness as it can only represent coarse-scale geometry.

5.3.2 *Trainable Shape and Reflectance Corrections*

Having only a coarse-scale face representation is one of the major shortcomings of many optimization- and learning-based reconstruction techniques, such as Blanz et al. (2003), Blanz and Vetter (1999), Tewari et al. (2017), and Thies et al. (2016b). Due to its

low dimensionality, the base model described in Sec. 5.3.1 has a limited expressivity for modeling the facial shape and reflectance at high accuracy. A particular problem is skin albedo variation, since the employed model has an ethnic bias and lacks facial hair, for example, beards. The purpose of this work is to improve upon this by learning a trainable corrective model that can represent these out-of-space variations. Unlike other approaches that use a fixed pre-defined corrective basis Garrido et al. (2016a), both the generative model for correctives and the best corrective parameters are learned. Furthermore, no ground truth annotations are required for geometry, skin reflectance and incident illumination.

The corrective model is based on (potentially non-linear) mappings $\mathcal{F}_\bullet : \mathbb{R}^C \rightarrow \mathbb{R}^{3N}$ that map the C-dimensional corrective parameter space onto per-vertex corrections in shape or reflectance. The mapping $\mathcal{F}_\bullet(\delta_\bullet|\Theta_\bullet)$ is a function of $\delta_\bullet \in \mathbb{R}^C$ that is parameterized by $\Theta_\bullet$. The motivation for disambiguating between $\delta_\bullet$ and $\Theta_\bullet$ is that during training both $\delta_\bullet$ and $\Theta_\bullet$ are learned, while at test time $\Theta_\bullet$ is kept fixed and the corrective parameters $\delta_\bullet$ are directly regressed using the feed forward network. In the affine/linear case, one can interpret $\Theta_\bullet$ as a basis that spans a subspace of the variations, and $\delta_\bullet$ is the coefficient vector that reconstructs a given sample using the basis. However, in general $\mathcal{F}_\bullet$ is not assumed to be affine/linear. The key difference to the base level is that the correction level does not use a fixed pre-trained basis but learns a generative model, along with the coefficients, directly from the training data.

5.4 DIFFERENTIABLE IMAGE FORMATION MODEL

To train the novel multi-level face reconstruction approach end-to-end, a differentiable image formation model is required.

FULL PERSPECTIVE CAMERA The position and rotation of the virtual camera is parameterized based on a rigid transformation $\Phi(\mathbf{v}) = \mathbf{R}\mathbf{v} + \mathbf{t}$, which maps a model space 3D point $\mathbf{v}$ onto camera space $\hat{\mathbf{v}} = \Phi(\mathbf{v})$. Here, $\mathbf{R} \in \mathrm{SO}(3)$ is the camera rotation and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector. To render virtual images of the scene, a full perspective camera model is used to project the camera space point $\hat{\mathbf{v}}$ into a 2D point $\mathbf{p} = \Pi(\hat{\mathbf{v}}) \in \mathbb{R}^2$. The camera model contains the intrinsics and performs the perspective division.

ILLUMINATION MODEL Distant lighting is assumed and the incoming radiance is approximated using spherical harmonics (SH) basis functions $H_b : \mathbb{R}^3 \rightarrow \mathbb{R}$. The incoming radiance is assumed to only depend on the surface normal $\mathbf{n}$:

$$\tilde{B}(\mathbf{r}, \mathbf{n}, \boldsymbol{\gamma}) = \mathbf{r} \odot \sum_{b=1}^{B^2} \boldsymbol{\gamma}_b H_b(\mathbf{n}) \ . \tag{5.5}$$

Here, $\odot$ denotes the Hadamard product, $\mathbf{r}$ is the surface reflectance and B is the number of spherical harmonics bands. $\boldsymbol{\gamma}_b \in \mathbb{R}^3$ are coefficients to control the illumination. Since the incident radiance is sufficiently smooth, an average error below 1% (Ramamoorthi and Hanrahan, 2001a) can be achieved with only $B = 3$ bands independent of the illumination. This leads to $m_l = B^2 = 9$ variables per color channel.

IMAGE FORMATION The differentiable image formation layer takes as input the per-vertex shape and reflectance in model space. This can be the model of the base level $\mathbf{v}^{\mathrm{b}}$ and $\mathbf{r}^{\mathrm{b}}$ or of the final level $\mathbf{v}^{\mathrm{f}}$ and $\mathbf{r}^{\mathrm{f}}$ that include the learned correctives. Let $\mathbf{v}_i^\ell \in \mathbb{R}^3$ and $\mathbf{r}_i^\ell \in \mathbb{R}^3$ denote the position and the reflectance of the i-th vertex for the base level ($\ell = \mathrm{b}$) and the final level ($\ell = \mathrm{f}$). The rendering layer takes this information and forms a point-based rendering of the scene, as follows. First, it maps the points onto camera space, i.e., $\hat{\mathbf{v}}_i^\ell = \Phi(\mathbf{v}_i^\ell)$, and then computes the projected pixel positions of all vertices as

$$\mathbf{u}_i^\ell(\mathbf{x}) = \Pi(\hat{\mathbf{v}}_i^\ell) \ .$$

The shaded colors $\mathbf{c}_i^\ell$ at these pixel locations are computed based on the illumination model described before:

$$\mathbf{c}_i^\ell(\mathbf{x}) = \tilde{B}(\mathbf{r}_i^\ell, \hat{\mathbf{n}}_i^\ell, \boldsymbol{\gamma}) \ ,$$

where $\hat{\mathbf{n}}_i^\ell$ are the associated camera space normals to $\hat{\mathbf{v}}_i^\ell$. The image formation model is differentiable, which enables end-to-end training using back propagation. The free variables that the regressor learns to predict are: The model parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}_g, \boldsymbol{\delta}_r)$, the camera parameters $\mathbf{R}$, $\mathbf{t}$ and the illumination parameters $\boldsymbol{\gamma}$. In addition, the corrective shape and reflectance bases Θ_g, Θ_r are learned during training. This leads to the following vector of unknowns:

$$\mathbf{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}_g, \boldsymbol{\delta}_r, \mathbf{R}, \mathbf{t}, \boldsymbol{\gamma}, \Theta_g, \Theta_r) \in \mathbb{R}^{257+2C+|\Theta_g|+|\Theta_r|} \ .$$

5.5 SELF-SUPERVISED LEARNING

The face regression network is trained using a novel self-supervision loss that enables fitting the base model and learning per-vertex correctives end-to-end. The loss function consists of a data fitting and regularization term:

$$E_{\mathrm{total}}(\mathbf{x}) = E_{\mathrm{data}}(\mathbf{x}) + w_{\mathrm{reg}} E_{\mathrm{reg}}(\mathbf{x}) \ , \tag{5.6}$$

where E_{data} penalizes misalignments of the model to the input image and E_{reg} encodes prior assumptions about faces at the coarse and medium scale. Here, w_{reg} is a trade-off factor that controls the amount of regularization. The data fitting term is based on sparse and dense alignment constraints:

$$E_{\mathrm{data}}(\mathbf{x}) = E_{\mathrm{sparse}}(\mathbf{x}) + w_{\mathrm{photo}} E_{\mathrm{photo}}(\mathbf{x}) \ . \tag{5.7}$$

The regularization term represents prior assumptions on the base and corrective model:

$$E_{\mathrm{reg}}(\mathbf{x}) = E_{\mathrm{std}}(\mathbf{x}) + E_{\mathrm{smo}} + E_{\mathrm{ref}}(\mathbf{x}) + E_{\mathrm{glo}}(\mathbf{x}) + E_{\mathrm{sta}}(\mathbf{x}) \, . \tag{5.8}$$

In the following, the individual terms are explained in detail.

5.5.1 *Data Terms*

MULTI-LEVEL DENSE PHOTOMETRIC LOSS A dense multi-level photometric loss function is employed that measures the misalignment of the coarse and fine fit to the input. Let $\bar{\mathcal{V}}$ be the set of all visible vertices. The photometric term is then defined as:

$$E_{\mathrm{photo}}(\mathbf{x}) = \sum_{\ell \in \{\mathrm{b},\mathrm{f}\}} \frac{1}{N} \sum_{i \in \bar{\mathcal{V}}} \left\| \mathcal{I}\left(\mathbf{u}_i^\ell(\mathbf{x})\right) - \mathbf{c}_i^\ell(\mathbf{x}) \right\|_2 \ . \tag{5.9}$$

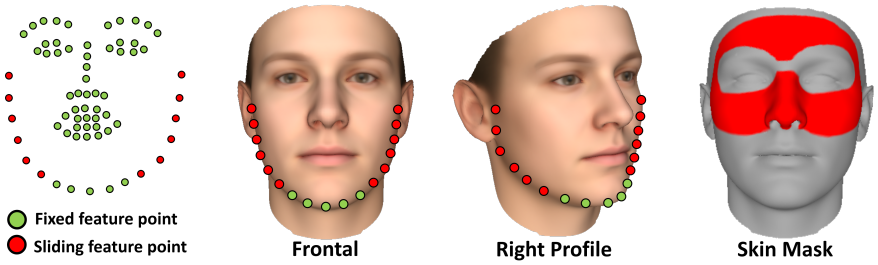

Figure 5.3: Fixed and sliding feature points are treated differetly. This leads to better contour alignment. Note how the outer contour depends on the rigid head pose (left). The skin mask (right) is employed in the global reflectance constancy constraint.

Here, $\mathbf{u}_i^\ell(\mathbf{x})$ is the screen space position, $\mathbf{c}_i^\ell(\mathbf{x})$ is the shaded color of the i-th vertex, and $\mathcal{I}$ is the current image during training. For robustness, the $\ell_{2,1}$-norm is employed, which measures the color distance using the ℓ_2-norm, while the summation over all pixel-wise ℓ_2-norms encourages sparsity as it corresponds to the ℓ_1-norm. Visibility is computed using backface culling, as explained in Chapter 2. This is an approximation, but works well, since faces are almost convex.

SPARSE FEATURE POINTS Faces contain many salient feature points. This is exploited by using a weak supervision in the form of automatically detected 66 facial landmarks $\mathbf{f} \in \mathcal{F} \subset \mathbb{R}^2$ (Saragih et al., 2011) and associated confidence $c_\mathbf{f} \in [0, 1]$ (1 confident). The set of facial landmarks falls in two categories: Fixed and sliding feature points. Fixed feature points, for example, eyes and nose, are associated with a fixed vertex on the template model, whereas sliding feature points, for example, the face contour, change their position on the template based on the rigid pose, see Fig. 5.3. This is explicitly modeled as follows:

$$E_{\text{sparse}}(\mathbf{x}) = \frac{1}{|\mathcal{F}|} \sum_{\mathbf{f} \in \mathcal{F}} c_\mathbf{f} \cdot \left\| \mathbf{f} - \mathbf{u}_{k_\mathbf{f}}^{\text{b}}(\mathbf{x}) \right\|_2^2 . \tag{5.10}$$

Here, $k_\mathbf{f}$ is the index of the target vertex. For fixed feature points, the index of the corresponding mesh vertex is hard-coded. The indexes for sliding feature points are computed via an alternation scheme: In each step of stochastic gradient descent, mesh vertex that is closest to the 3D line is computed, defined by the camera center and the back-projection of the detected 2D feature point. Based on the squared Euclidean distance $k_\mathbf{f}$ is set to the index of the closest vertex.

5.5.2 *Regularization Terms*

STATISTICAL REGULARIZATION Statistical regularization is enforced on the 3DMM model parameters of the base level to ensure plausible reconstructions. Based on the assumption that the model parameters follow a zero-mean Gaussian distribution, Tikhonov regularization is employed:

$$E_{\text{std}}(\mathbf{x}) = \sum_{k=1}^{m_s+m_e} \left(\frac{\alpha_k}{(\sigma_g)_k} \right)^2 + w_{\text{rstd}} \sum_{k=1}^{m_r} \left(\frac{\beta_k}{(\sigma_r)_k} \right)^2 . \tag{5.11}$$

This is a common constraint (Blanz and Vetter, 1999; Garrido et al., 2016a; Tewari et al., 2017; Thies et al., 2016b) that prevents the degeneration of the facial geometry and face reflectance in the ill-posed monocular reconstruction scenario.

CORRECTIVE SMOOTHNESS Local smoothness is also imposed by adding Laplacian regularization on the vertex displacements for the set of all vertices $\mathcal{V}$:

$$E_{\text{smo}}(\mathbf{x}) = \frac{w_{\text{smo}}}{N} \sum_{i \in \mathcal{V}} \left\| \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left((\mathcal{F}_g(\mathbf{x}))_i - (\mathcal{F}_g(\mathbf{x}))_j \right) \right\|_2^2 . \tag{5.12}$$

Here, $(\mathcal{F}_g(\mathbf{x}))_i = (\mathcal{F}_g(\boldsymbol{\delta}_g | \Theta_g))_i$ denotes the correction for the i-th vertex given the parameter $\mathbf{x}$, and $\mathcal{N}_i$ is the 1-ring neighborhood of the i-th vertex.

LOCAL REFLECTANCE SPARSITY In spirit of recent intrinsic decomposition approaches (Bonneel et al., 2014; Meka et al., 2016), sparsity is enforced to further regularize the reflectance of the full reconstruction:

$$E_{\text{ref}}(\mathbf{x}) = w_{\text{ref}} \frac{1}{N} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} w_{i,j} \cdot \left\| \mathbf{r}_i^{\text{f}}(\mathbf{x}) - \mathbf{r}_j^{\text{f}}(\mathbf{x}) \right\|_2^p . \tag{5.13}$$

Here, $w_{i,j} = \exp\left(-\alpha \cdot ||\mathcal{I}(u_i^{\text{f}}(\mathbf{x}^{\text{old}})) - \mathcal{I}(u_j^{\text{f}}(\mathbf{x}^{\text{old}}))||_2 \right)$ are constant weights that measure the chromaticity similarity between the colors in the input, where $\mathbf{x}^{\text{old}}$ are the parameters estimated in the previous iteration. Pixels with the same chromaticity are assumed to be more likely to have the same reflectance. The term $\|\cdot\|_2^p$ enforces sparsity on the combined reflectance estimate. The hyperparameters are fixed as $\alpha = 50$ and $p = 0.9$ in all the experiments.

GLOBAL REFLECTANCE CONSTANCY Skin reflectance constancy is enforced over a fixed set of vertices that covers only the skin region, see Fig. 5.3 (right):

$$E_{\text{glo}}(\mathbf{x}) = w_{\text{glo}} \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{G}_i} \left\| \mathbf{r}_i^{\text{f}}(\mathbf{x}) - \mathbf{r}_j^{\text{f}}(\mathbf{x}) \right\|_2^2 . \tag{5.14}$$

Here, $\mathcal{M}$ is the per-vertex skin mask and $\mathcal{G}_i$ stores 6 random samples of vertex indexes of the mask region. The idea is to enforce the whole skin region to have the same reflectance. For efficiency, reflectance similarity between random pairs of vertices in the skin region is used. Note that regions that may have facial hair were not included in the mask. In combination, local and global reflectance constancy efficiently removes shading from the reflectance channel.

STABILIZATION It is ensured that the corrected geometry stays close to the base reconstruction by enforcing small vertex displacements:

$$E_{\text{sta}}(\mathbf{x}) = w_{\text{sta}} \frac{1}{N} \sum_{i \in \mathcal{V}} \left\| (\mathcal{F}_g(\mathbf{x}))_i \right\|_2^2 . \tag{5.15}$$

5.5.3 *Training Details*

Training the face regressor and the corrective space jointly is challenging. Thus, for robust training, the network up to the base level is pretrained for 200k iterations

with a learning rate of 0.01. The weights $w_\bullet$ are empirically determined and fixed for all the experiments. For training the base level, the following weights are used: $w_{\text{photo}} = 1.9, w_{\text{reg}} = 0.00003, w_{\text{rstd}} = 0.002, w_{\text{smo}} = 0.0, w_{\text{ref}} = 0.0, w_{\text{glo}} = 0.0$ and $w_{\text{sta}} = 0.0$. Afterwards, the complete network is finetuned for 190k iterations end-to-end with a learning rate of 0.001 for the base level network, 0.005 for the geometry correctives network and 0.01 for the reflectance correctives network. For finetuning, the loss is instantiated based on the following weights: $w_{\text{photo}} = 0.2, w_{\text{reg}} = 0.003, w_{\text{rstd}} = 0.002, w_{\text{smo}} = 3.2 \cdot 10^4, w_{\text{ref}} = 13, w_{\text{glo}} = 80, w_{\text{sta}} = 0.08$. The method uses 500 correctives for both geometry and reflectance. Please note, the illumination estimate for rendering the base and final model is not shared between the two levels, but independently regressed. This is due to the fact that a different illumination estimate might be optimal for the coarse and final reconstruction due to the shape and skin reflectance correctives. During finetuning, all weights associated with the correctives receive a higher learning rate ($\times 100$) than the pretrained layers. This two stage strategy enables robust and efficient training of the architecture.

5.6 RESULTS

This section demonstrates joint end-to-end self-supervised training of the feed forward encoder and the novel multi-level face representation based on in-the-wild images without the need for densely annotated ground truth. The proposed approach regresses pose, shape, expression, reflectance and illumination at high-quality with over 250 Hz, see Fig. 5.4. A modified version of AlexNet (Krizhevsky et al., 2012) that outputs the parameters of the face model is used as the feed-forward encoder. Note that other feed forward architectures could be used. The approach is implemented using *Caffe* (Jia et al., 2014). Training is based on *AdaDelta* with a batch size of 5. The network is pretrained up to the base level for 200k iterations with a learning rate of 0.01. Afterwards, the complete network is finetuned for 190k iterations with a learning rate of 0.001 for the base level, 0.005 for the geometry and 0.01 for the reflectance correctives. All components of the network are implemented in CUDA (NVIDIA, 2008) for efficient training, which takes 16 hours. The same weights $w_\bullet$ are used in all experiments. In the following, the size, C of the corrective parameters are fixed to 500 for both geometry and reflectance. Different corrective spaces (linear and non-linear) are tested, see Fig. 5.6. A linear corrective basis gave the best results, which is used for all following experiments. The approach is trained on a corpus of in-the-wild face images, without densely annotated ground truth. Training is performed on a combination of four different datasets: CelebA (Liu et al., 2015), LFW (Huang et al., 2007), FaceWarehouse (Cao et al., 2013), and 300-VW (Chrysos et al., 2015; Shen et al., 2015; Tzimiropoulos, 2015). Sparse landmark annotations are obtained automatically (Saragih et al., 2011) and images are cropped to a tight face bounding box of 240×240 pixels using Haar Cascade Face Detection (Bradski, 2000). Images with bad detections are automatically removed based on landmark confidence. In total, 144k images are used, which are split into a training (142k images) and validation (2k images) set.

The final output ('final') is compared to the base low-dimensional 3DMM reconstruction ('base') obtained from the pretrained network to illustrate that the multi-level

Figure 5.4: The proposed approach allows for high-quality reconstruction of facial geometry, reflectance and incident illumination from just a single monocular color image. Note the reconstructed facial hair, for example, the beard, reconstructed make-up, and the eye lid closure, which are outside of the space of the used 3DMM.

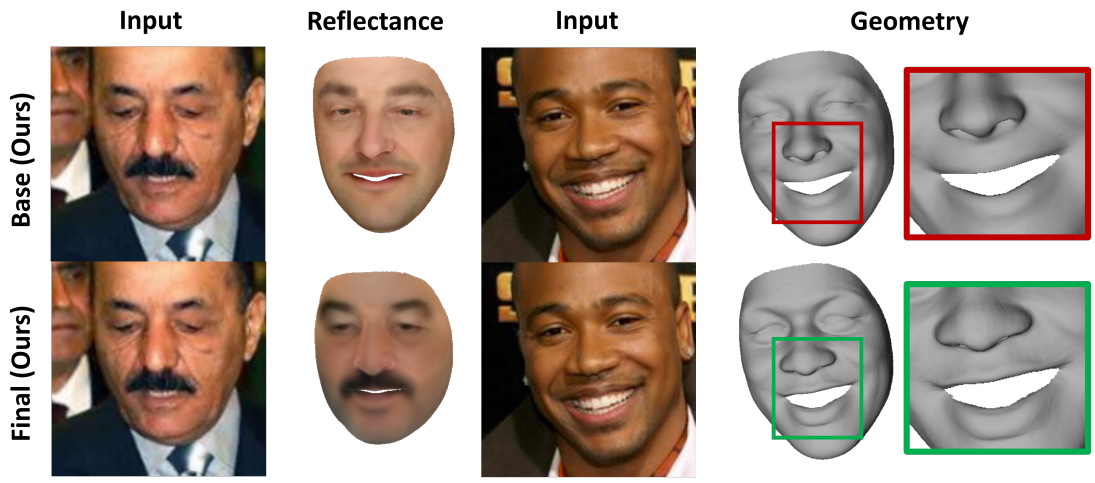

Figure 5.5: Jointly learning a multi-level model improves the geometry and reflectance compared to the 3DMM. Note the better aligning nose, lips and the reconstructed facial hair.

Figure 5.6: Comparison of linear and non-linear corrective spaces.

model recovers higher quality geometry and reflectance (Fig. 5.5). In the following, more results are shown, and the approach is compare to the state of the art.

5.6.1 *Comparisons to the State of the Art*

OPTIMIZATION-BASED TECHNIQUES The method is compared to the optimization-based high-quality reconstruction method of Garrido et al. (2016a), see Fig. 5.7. The proposed approach obtains similar geometry quality but better captures the person's characteristics due to the learned corrective space. Since the approach jointly learns a corrective reflectance space, it can leave the restricted subspace of the underlying 3DMM and thus produces more realistic appearance. Note, unlike Garrido et al. (2016a), the proposed approach does not require landmarks at test time and runs orders of magnitude faster (4ms vs. 120s per image). See Fig. 5.8 for comparisons with the approach of Booth et al. (2017). The proposed approach jointly learns a better shape and reflectance model, while their approach only builds an 'in-the-wild' texture model that contains shading. In contrast to the proposed approach, Booth et al. is based on optimization and requires initialization or landmarks.

LEARNING-BASED TECHNIQUES This paragraph provides comparisons to the high-quality learning-based reconstruction approaches of Tewari et al. (2017), the method presented in Chapter 4 (Fig. 5.9), Richardson et al. (2016, 2017) (Fig. 5.10) and Sela et al. (2017) (Fig. 5.10). These approaches obtain impressive results within the span of the used synthetic training corpus or the employed 3DMM model, but

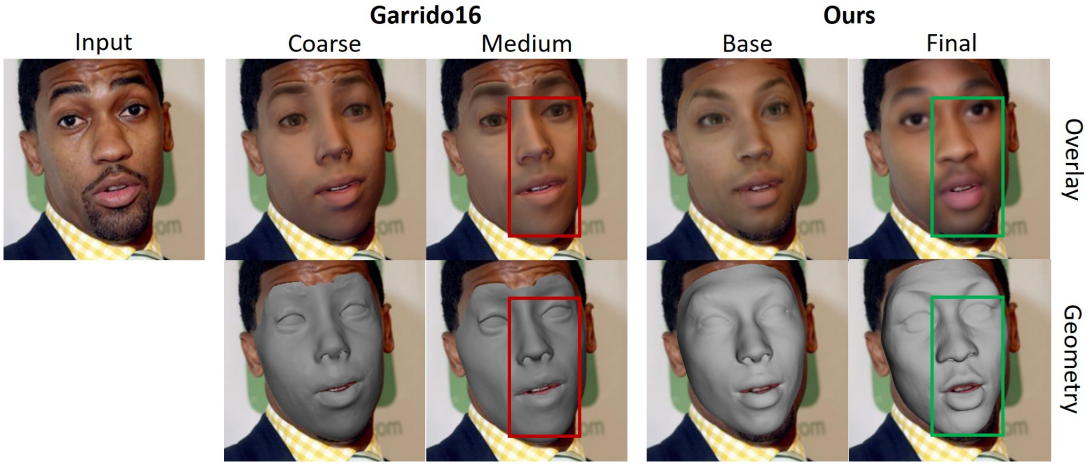

Figure 5.7: Comparison to Garrido et al. (2016a). The approach presented achieves higher quality reconstructions, since the jointly learned model generalizes better than a corrective space based on manifold harmonics.

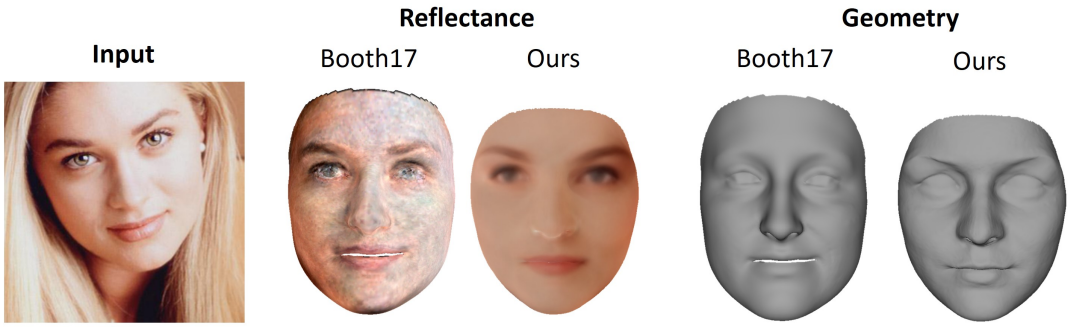

Figure 5.8: In contrast to the texture model of Booth et al. (2017) that contains shading, the proposed approach yields a reflectance model.

suffer from out-of-subspace shape and reflectance variations, for example, people with beards. The proposed approach is not only robust to facial hair and make-up, but also automatically learns to reconstruct such variations based on the jointly learned model. Reconstruction requires 4 ms, while Sela et al. (2017) requires slow off-line non-rigid registration to obtain a hole free reconstruction from the predicted depth map. In addition, a reconstruction of colored reflectance and illumination is jointly obtained. Due to the model learning, the proposed approach is able to leave the low-dimensional space of the 3DMM, which leads to a more realistic reconstruction of facial appearance and geometry.

5.6.2 *Quantitative Results*

Quantitative evaluations are also performed. For geometry, the FaceWarehouse (Cao et al., 2013) dataset is used, and 180 meshes (9 identities, 20 expressions each) are reconstructed. Various approaches are compared using the metric introduced in Chapter 4, after alignment (rigid transform plus isotropic scaling) to the provided ground truth. The proposed approach outperforms the learning-based techniques of Tewari et al. (2017) and Kim et al. (2018b), see Tab. 5.1. The results are close to the high-quality

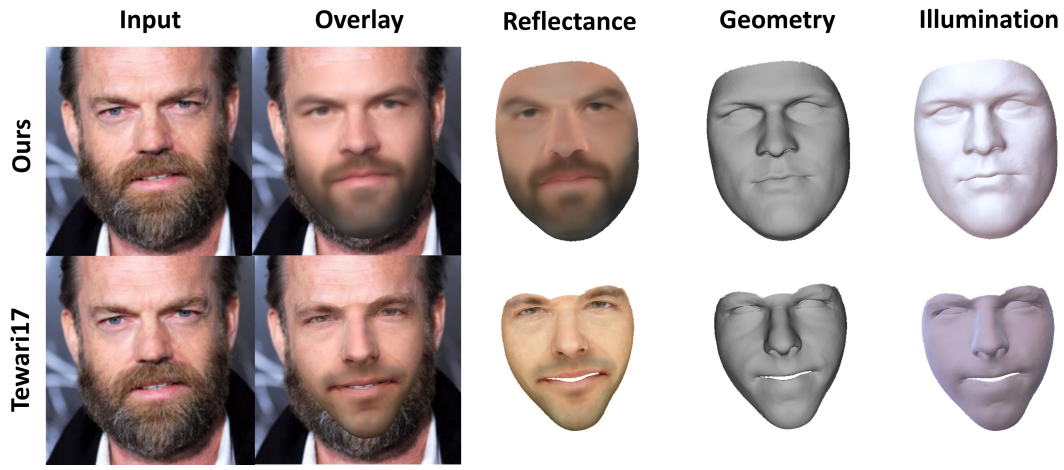

Figure 5.9: Comparison to Tewari et al. (2017), the method presented in Chapter 4. Higher quality (without surface shrinkage) is achieved due to the jointly trained model.

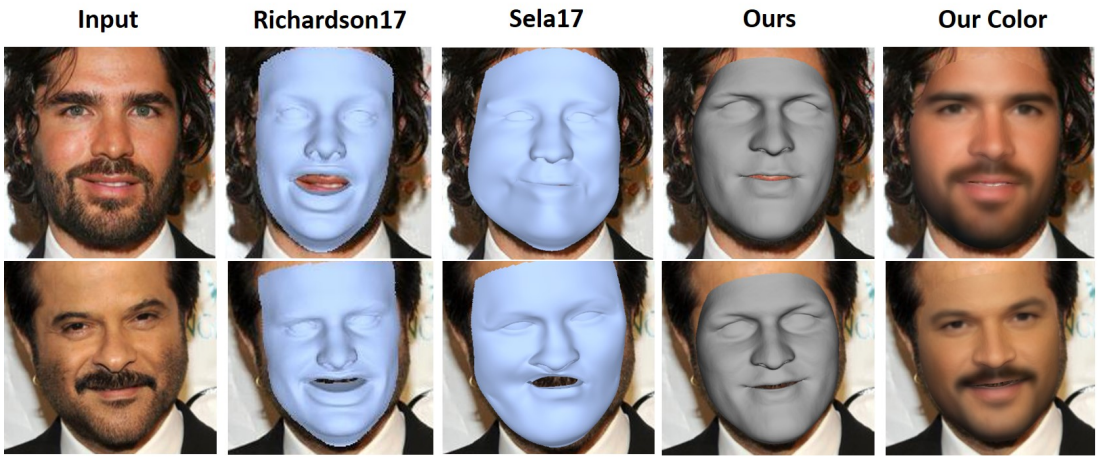

Figure 5.10: Comparison to Richardson et al. (2016, 2017) and Sela et al. (2017). They obtain impressive results within the span of the synthetic training corpus, but do not handle out-of-subspace variations, for example, beards. The proposed approach is robust to hair and make-up, since the model is jointly learned.

optimization approach of Garrido et al. (2016a), while being orders of magnitude faster (4ms vs. 120sec) and not requiring feature detection at test time, see Fig. 5.11 (top). Cao et al. (2013) contains mainly 'clean' faces without make-up or beards, since this causes problems even for high-quality offline 3D reconstruction approaches. The primary interest is in robustly handling the harder in-the-wild scenario, in which the proposed approach significantly outperforms previous approaches, see Figs. 5.7, 5.9, and 5.10. The approach is also evaluated on a video sequence (300 frames) with challenging expressions and a characteristic face, which is outside the span of the 3DMM. The ground truth has been obtained by Valgaerts et al. (2012). The results can be found in Tab. 5.2 and in Fig. 5.11 (bottom), where it can be seen that the proposed method outperforms other learning- and optimization-based approaches (Garrido et al., 2016a; Tewari et al., 2017). The photometric fitting error of the proposed approach is evaluated on the validation set, see Fig. 5.12. The final results (mean: 0.072,

Table 5.1: Geometric error on FaceWarehouse (Cao et al., 2013). The proposed approach outperforms the deep learning techniques of Tewari et al. (2017) and Kim et al. (2018b). It comes close to the high-quality approach of Garrido et al. (2016a), while being orders of magnitude faster and not requiring feature detection.

	Ours		Others		
	Learning		Learning		Optimization
	Fine	Coarse	Tewari et al. (2017)	Kim et al. (2018b)	Garrido et al. (2016a)
Mean	**1.84 mm**	2.03 mm	2.19 mm	2.11 mm	**1.59 mm**
SD	**0.38 mm**	0.52 mm	0.54 mm	0.46 mm	**0.30 mm**
Time	**4 ms**	**4 ms**	**4 ms**	**4 ms**	120 s

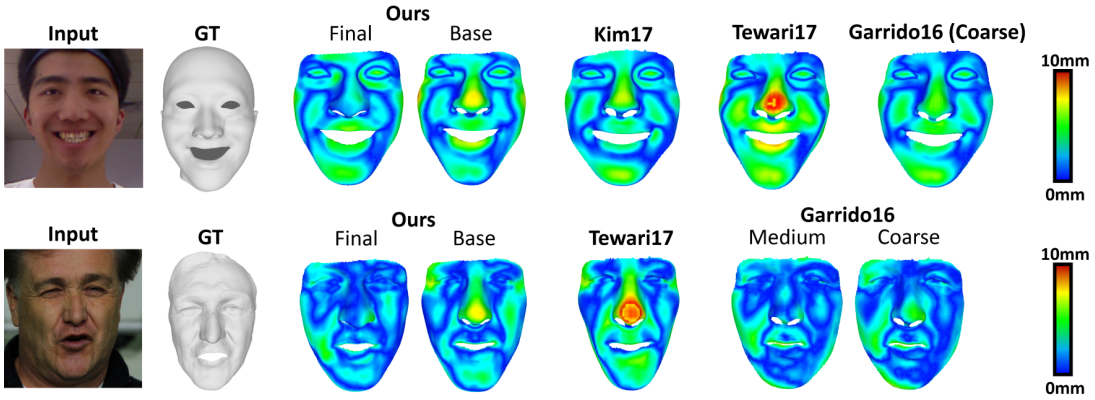

Figure 5.11: Higher quality is obtained compared to the previous learning-based approaches on the FaceWarehouse (Cao et al., 2013) and Volker (Valgaerts et al., 2012) datasets.

SD: 0.020) have significantly lower error (distance in RGB space, channels in $[0, 1]$) than the base level (mean: 0.092, SD: 0.025) due to the learned corrective basis.

5.7 LIMITATIONS

This chapter presented high-quality monocular reconstruction at over 250Hz, even in the presence of facial hair, or for challenging faces. Still, the approach has a few limitations, which can be addressed in future work: External occlusion, for example, by glasses, are baked into our correctives, see Fig. 5.13. Resolving this would require

Table 5.2: On the Volker sequence, the proposed approach outperforms the results of Garrido et al. (2016a), even if their fixed shape correctives are employed.

	Ours		Others		
	Learning		Learning	Optimization Garrido et al. (2016a)	
	Fine	Coarse	Tewari et al. (2017)	Medium	Coarse
Mean	**1.77 mm**	2.16 mm	2.94 mm	1.97 mm	1.96 mm
SD	**0.29 mm**	0.29 mm	0.28 mm	0.41 mm	0.35 mm

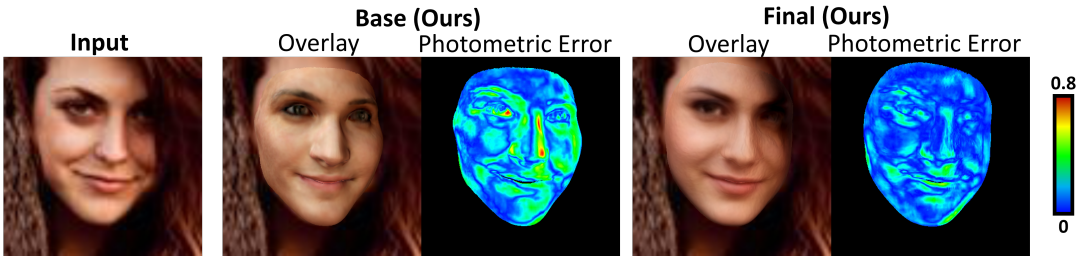

Figure 5.12: Euclidean photometric error in RGB space, each channel in $[0, 1]$. Final results significantly improve the fitting quality.

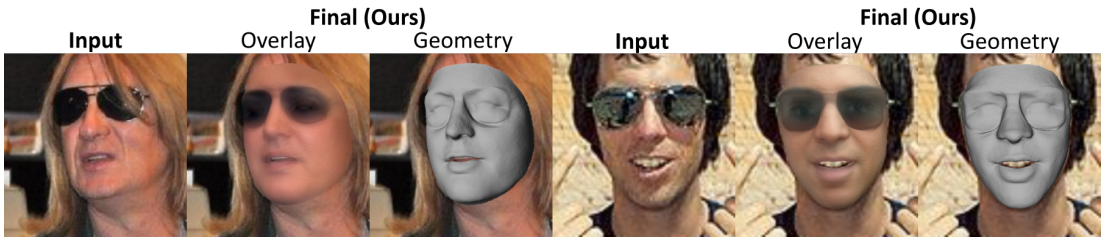

Figure 5.13: External occluders are baked into the correctives.

a semantic segmentation of the training corpus. The consistent reconstruction of occluded face regions is not guaranteed. Low-dimensionality of the corrective space is enforced for robust model learning. Thus, fine-scale surface detail can not be recovered. This as an orthogonal research direction, which has already produced impressive results (Richardson et al., 2016, 2017; Sela et al., 2017).

5.8 CONCLUSION

This chapter presented the first approach that jointly learns a face model and a parameter regressor for face shape, expression, appearance and illumination. It combines the advantages of 3DMM regularization with the out-of-space generalization of a learned corrective space. This overcomes the disadvantages of current approaches that rely on strong priors, increases generalization and robustness, and leads to high quality reconstructions at over 250Hz. While this work focused on face reconstruction, the approach is not restricted to faces only as it can be generalized to further object classes. As such, this was a first important step towards building 3D models from in-the-wild images. The next chapter will take a step forward in this direction in order to learn the identity components of the morphable model entirely from 2D images, without the coarse 3DMM.

6

FML: FACE MODEL LEARNING FROM VIDEOS

Figure 6.1: This chapter proposes multi-frame self-supervised training of a deep network based on in-the-wild video data for jointly learning a face model and 3D face reconstruction. The proposed approach successfully disentangles facial shape, appearance, expression, and scene illumination.

Chapter 4 presented a method to reconstruct 3D faces from images using morphable model priors. Chapter 5 allowed for refinement of the morphable model for higher quality reconstructions. This chapter, published as Tewari et al. (2019), takes this idea further and learns the face identity variations from 2D data without using any existing morphable model, see Fig. 6.1. The face model is learned using only corpora of *in-the-wild* video clips collected from the Internet. This virtually endless source of training data enables learning of a highly general 3D face model. In order to achieve this, a novel multi-frame consistency loss is proposed that ensures consistent shape and appearance across multiple frames of a subject's face, thus minimizing depth ambiguity. At test time an arbitrary number of frames can be used, so that both monocular as well as multi-frame reconstruction can be performed.

6.1 INTRODUCTION

The reconstruction of faces from visual data has a wide range of applications in vision and graphics, including face tracking, emotion recognition, and interactive image/video editing tasks relevant in multimedia. Facial images and videos are ubiquitous, as smart devices as well as consumer and professional cameras provide a continuous and virtually endless source thereof. When such data is captured without controlled scene location, lighting, or intrusive equipment (for example, egocentric cameras or markers on actors), one speaks of *"in-the-wild"* images. Usually *in-the-wild* data is of low resolution, noisy, or contains motion and focal blur, making the reconstruction problem much harder than in a controlled setup. 3D face reconstruction from *in-the-wild* monocular 2D image and video data (Zollhöfer et al., 2018) deals with disentangling facial shape identity (neutral geometry), skin appearance (or albedo) and expression, as well as estimating the scene lighting and camera parameters. Some of these attributes, for example, albedo and lighting, are not easily separable in monocular images. Besides, poor scene lighting, depth ambiguity, and occlusions due to facial hair, sunglasses and large head rotations complicates 3D face reconstruction.

In order to tackle the difficult monocular 3D face reconstruction problem, most existing methods rely on the availability of strong prior models that serve as regularizers for an otherwise ill-posed problem (Blanz et al., 2003; Ekman and Rosenberg, 1997;

Vlasic et al., 2005). Although such approaches achieve impressive facial shape and albedo reconstruction, they introduce an inherent bias due to the used face model. For instance, the 3D Morphable Model (3DMM) by Blanz et al. (2003) is based on a comparably small set of 3D laser scans of Caucasian actors, thus limiting generalization to general real-world identities and ethnicities. With the rise of CNN-based deep learning, various techniques have been proposed, which in addition to 3D reconstruction also perform face model learning from monocular images (Shu et al., 2017b; Tran and Liu, 2018a,b). Chapter 5 also presented such a method. However, these methods heavily rely on a pre-existing 3DMM to resolve the inherent depth ambiguities of the monocular reconstruction setting. Another line of work, where 3DMM-like face models are not required, are based on photo-collections (Kemelmacher-Shlizerman, 2013; Liang et al., 2016; Suwajanakorn et al., 2014). However, these methods need a very large number ($\approx$100) of facial images of the same subject, and thus they impose strong demands on the training corpus.

This chapter introduces an approach that learns a comprehensive face identity model using clips crawled from *in-the-wild* Internet videos (Chung et al., 2018). This face identity model comprises two components: One component to represent the geometry of the facial identity (modulo expressions), and another to represent the facial appearance in terms of the albedo. As there are only weak requirements on the training data (described later in Sec. 6.2.1), the proposed approach can employ a virtually endless amount of community data and thus obtain a model with better generalization; laser scanning a similarly large group of people for model building would be practically impossible. Unlike most previous approaches, the proposed method does not require a pre-existing shape identity and albedo model as initialization, but instead learns their variations from scratch. As such, the methodology is applicable in scenarios when no existing model is available, or if it is difficult to create such a model from 3D scans (for example, faces of babies).

From a technical point of view, one of the main contributions is a novel multi-frame consistency loss, which ensures that the face identity and albedo reconstruction is consistent across frames of the same subject. This way depth ambiguities present in many monocular approaches can be resolved to obtain a more accurate and robust model of facial geometry and albedo. Moreover, by imposing orthogonality between the learned face identity model and an existing blendshape expression model, the approach automatically disentangles facial expressions from identity-based geometry variations, without resorting to a large set of hand-crafted priors. In summary, the approach is based on the following technical contributions:

1. A deep neural network that learns a facial shape and appearance space from a big dataset of unconstrained images that contains multiple images of each subject, for example, multi-view sequences, or even monocular videos.
2. Explicit blendshape and identity separation by a projection onto the blendshapes' nullspace that enables a multi-frame consistency loss.
3. A novel multi-frame identity consistency loss based on a Siamese network (Vinyals et al., 2016), with the ability to handle monocular and multi-frame reconstruction.

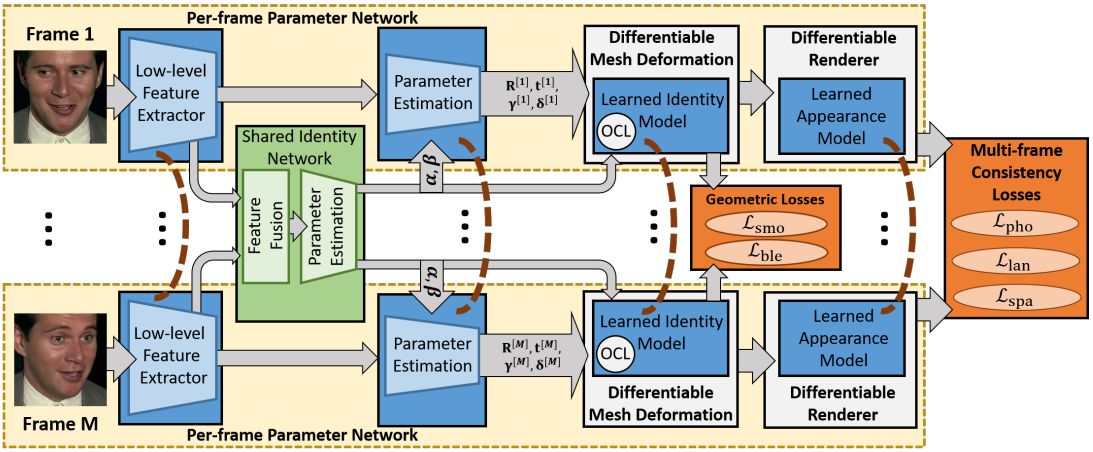

Figure 6.2: Pipeline overview. Given multi-frame input that shows a person under different facial expression, head pose, and illumination, the proposed approach first estimates these parameters per frame. In addition, it jointly obtains the shared identity parameters that control facial shape and appearance, while at the same time learning a graph-based geometry and a per-vertex appearance model. A differentiable mesh deformation layer is used in combination with a differentiable face renderer to implement a model-based face autoencoder.

6.2 FACE MODEL LEARNING

The face model learning approach solves two tasks: it jointly learns (i) a parametric face geometry and appearance model, and (ii) an estimator for facial shape, expression, albedo, rigid pose and incident illumination parameters. An overview of the approach is shown in Fig. 6.2.

TRAINING: The network is trained in a self-supervised fashion based on a training set of multi-frame images, i.e., multiple images of the same person sampled from a video clip, see Section 6.2.1. The network jointly learns an appearance and shape identity model (Section 6.2.2). It also estimates per-frame parameters for the rigid head pose, illumination, and expression parameters, as well as shape and appearance identity parameters that are shared among all frames. The network is trained using a differentiable renderer that incorporates a per-vertex appearance model and a graph-based shape deformation model (Section 6.2.3). A set of training losses are proposed that account for geometry smoothness, photo-consistency, sparse feature alignment and appearance sparsity, see Section 6.2.4.

TESTING: At test time, the network jointly reconstructs shape, expression, albedo, pose and illumination from an arbitrary number of face images of the same person. Hence, the same trained network is usable both for monocular and multi-frame face reconstruction.

6.2.1 *Dataset*

The approach is trained using the VoxCeleb2 multi-frame video dataset (Chung et al., 2018). This dataset contains over 140k videos of over 6000 celebrities crawled from Youtube. A total of $N = 404$k multi-frame images $\mathcal{F}_1, \ldots, \mathcal{F}_N$ are sampled

from this dataset. The ℓ-th multi-frame image $\mathcal{F}_\ell = \{F_\ell^{[f]}\}_{f=1}^{M}$ comprises $M = 4$ frames $F_\ell^{[1]}, \ldots, F_\ell^{[M]}$ of the same person ℓ extracted from the same video clip to avoid unwanted variations, for example, due to aging or accessories. The same person can appear multiple times in the dataset. Several sequential steps are performed to obtain these images. First, the face region is cropped based on automatically detected facial landmarks (Saragih et al., 2011). Afterwards, the pipeline discards images whose cropped region is smaller than a threshold (i.e., 200 pixels) and that have low landmark detection confidence, as provided by the landmark tracker (Saragih et al., 2011). The remaining crops are re-scaled to 240×240 pixels. When sampling the M (possibly non-consequent) frames in $\mathcal{F}_\ell$, sufficient diversity in head pose is ensured based on the head orientation obtained by the landmark tracker. The multi-frame dataset $\mathcal{F}_1, \ldots, \mathcal{F}_N$ is split into a training (383k images) and test set (21k images).

6.2.2 *Graph-based Face Representation*

A multi-level face representation is proposed that is based on both a coarse shape deformation graph and a high-resolution surface mesh, where each vertex has a color value that encodes the facial appearance. This representation enables the approach to learn a face model of geometry and appearance based on multi-frame consistency. In the following, the components are explained in detail.

LEARNABLE GRAPH-BASED IDENTITY MODEL: Rather than learning the identity model on the high-resolution mesh $\mathcal{V}$ with $|\mathcal{V}| = 60$k vertices, this task is simplified by considering a lower-dimensional parametrization inspired by deformation graphs (Sumner et al., 2007). The (coarse) deformation graph $\mathcal{G}$ is obtained by downsampling the mesh to $|\mathcal{G}| = 521$ nodes, see Fig. 6.3. The network now learns a deformation of $\mathcal{G}$ that is then transferred to the mesh $\mathcal{V}$ via linear blend skinning Jacobson et al., 2014. The vector $\mathbf{g} \in \mathbb{R}^{3|\mathcal{G}|}$ of the $|\mathcal{G}|$ stacked node positions of the 3D graph is defined as

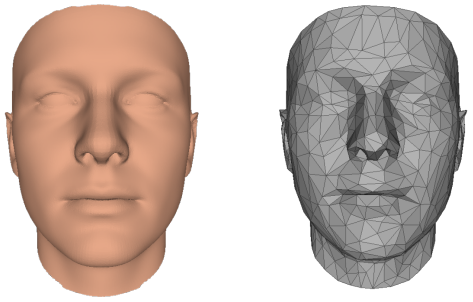

Figure 6.3: Neutral face shape and appearance (left), and the coarse deformation graph of the face mesh (right).

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{\Theta}_s \boldsymbol{\alpha}\,, \tag{6.1}$$

where $\bar{\mathbf{g}} \in \mathbb{R}^{3|\mathcal{G}|}$ denotes the mean graph node positions. $\bar{\mathbf{g}}$ is obtained by downsampling a face mesh with slightly open mouth (to avoid connecting the upper and lower lips). The columns of the learnable matrix $\mathbf{\Theta}_s \in \mathbb{R}^{3|\mathcal{G}|\times g}$ span the g-dimensional ($g = 500$) graph deformation subspace, and $\boldsymbol{\alpha} \in \mathbb{R}^g$ represents the graph deformation parameters.

The vertex positions $\mathbf{v} \in \mathbb{R}^{3|\mathcal{V}|}$ of the high-resolution mesh $\mathcal{V}$ that encode the shape identity are then given as

$$\mathbf{v}(\mathbf{\Theta}_s, \boldsymbol{\alpha}) = \bar{\mathbf{v}} + \mathbf{S}\mathbf{\Theta}_s \boldsymbol{\alpha}\,. \tag{6.2}$$

Here, $\bar{\mathbf{v}} \in \mathbb{R}^{3|\mathcal{V}|}$ is fixed to the neutral mean face shape as defined in the 3DMM (Blanz and Vetter, 1999). The skinning matrix $\mathbf{S} \in \mathbb{R}^{3|\mathcal{V}| \times 3|\mathcal{G}|}$ is obtained based on the mean shape $\bar{\mathbf{v}}$ and mean graph nodes $\bar{\mathbf{g}}$.

To sum up, the identity model is represented by a deformation graph $\mathcal{G}$, where the deformation parameter $\boldsymbol{\alpha}$ is regressed by the network while learning the deformation subspace basis $\boldsymbol{\Theta}_s$. This ill-posed learning problem is regularized by exploiting multi-frame consistency.

BLENDSHAPE EXPRESSION MODEL: A linear blendshape model is used for capturing the expressions. This model combines the facial expression models from Alexander et al. (2009) and Cao et al. (2013). The blendshape model is fixed, i.e. not learned. The expression deformations are directly applied to the high-res mesh. The vertex positions of the high-res mesh that account for shape identity as well as the facial expression are given by

$$\mathbf{v}(\boldsymbol{\Theta}_s, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \bar{\mathbf{v}} + \mathbf{S} \cdot \mathrm{OCL}(\boldsymbol{\Theta}_s)\boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\delta}\,, \tag{6.3}$$

where $\mathbf{B} \in \mathbb{R}^{3|\mathcal{V}| \times b}$ is the fixed blendshape basis, $\boldsymbol{\delta} \in \mathbb{R}^b$ is the vector of $b = 80$ blendshape parameters, and OCL is explained next.

SEPARATING SHAPE AND EXPRESSION: The approach ensures a separation of shape identity from facial expressions by imposing orthogonality between the learned shape identity basis and the fixed blendshape basis. To this end, the blendshape basis $\mathbf{B} \in \mathbb{R}^{3|\mathcal{V}| \times b}$ is first represented with respect to the deformation graph domain by solving $\mathbf{B} = \mathbf{S}\mathbf{B}_{\mathcal{G}}$ for the graph-domain blendshape basis $\mathbf{B}_{\mathcal{G}} \in \mathbb{R}^{3|\mathcal{G}| \times b_{\mathcal{G}}}$ in a least-squares sense. Here, $b_{\mathcal{G}} = 80$ is fixed. Then, the columns of $\mathbf{B}_{\mathcal{G}}$ are orthogonalized. The *Orthogonal Complement Layer (OCL)* is proposed to ensure that the learned $\mathrm{OCL}(\boldsymbol{\Theta}_s)$ fulfills the orthogonality constraint $\mathbf{B}_{\mathcal{G}}^T \mathrm{OCL}(\boldsymbol{\Theta}_s) = \mathbf{0}$. This layer is defined in terms of the projection of $\boldsymbol{\Theta}_s$ onto the orthogonal complement $\mathbf{B}_{\mathcal{G}}^{\perp}$ of $\mathbf{B}_{\mathcal{G}}$, i.e.,

$$\mathrm{OCL}(\boldsymbol{\Theta}_s) = \mathrm{proj}_{\mathbf{B}_{\mathcal{G}}^{\perp}}(\boldsymbol{\Theta}_s) = \boldsymbol{\Theta}_s - \mathrm{proj}_{\mathbf{B}_{\mathcal{G}}}(\boldsymbol{\Theta}_s) \tag{6.4}$$

$$= \boldsymbol{\Theta}_s - \mathbf{B}_{\mathcal{G}}(\mathbf{B}_{\mathcal{G}}^T\mathbf{B}_{\mathcal{G}})^{-1}\mathbf{B}_{\mathcal{G}}^T\boldsymbol{\Theta}_s\,. \tag{6.5}$$

The desired property $\mathbf{B}_{\mathcal{G}}^T \mathrm{OCL}(\boldsymbol{\Theta}_s) = \mathbf{0}$ follows directly.

LEARNABLE PER-VERTEX APPEARANCE MODEL: The facial appearance is encoded in the $3|\mathcal{V}|$-dimensional vector

$$\mathbf{r}(\boldsymbol{\beta}) = \bar{\mathbf{r}} + \boldsymbol{\Theta}_a\boldsymbol{\beta} \tag{6.6}$$

that stacks all $|\mathcal{V}|$ per-vertex diffuse reflectance colors represented as RGB triplets. The mean facial appearance $\bar{\mathbf{r}} \in \mathbb{R}^{3|\mathcal{V}|}$ and the appearance basis $\boldsymbol{\Theta}_a \in \mathbb{R}^{3|\mathcal{V}| \times |\boldsymbol{\beta}|}$ are learnable, while the facial appearance parameters $\boldsymbol{\beta}$ are regressed. Note that the mean appearance $\bar{\mathbf{r}}$ is initialized to a constant skin tone and the reflectance is defined directly on the high-res mesh $\mathcal{V}$.

6.2.3 *Differentiable Image Formation*

To enable end-to-end self-supervised training, a differentiable image formation model is employed that maps 3D model space coordinates $\mathbf{v} \in \mathbb{R}^3$ onto 2D screen space

coordinates $\mathbf{u} \in \mathbb{R}^2$. The mapping is implemented as $\mathbf{u} = \Pi(\Phi(\mathbf{v}))$, where Φ and Π denote the rigid head pose and camera projection, respectively. A differentiable illumination model is employed that transforms illumination parameters $\boldsymbol{\gamma}$ as well as per-vertex appearance $\mathbf{r}_i$ and normal $\mathbf{n}_i$ into shaded per-vertex color $\mathbf{c}_i(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma})$.

CAMERA MODEL: Without loss of generality, it is assumed that the camera space corresponds to world space. The head pose is modeled via a rigid mapping $\Phi(\mathbf{v}) = \mathbf{R}\mathbf{v} + \mathbf{t}$, defined by the global rotation $\mathbf{R} \in SO(3)$ and the translation $\mathbf{t} \in \mathbb{R}^3$. After mapping a vertex from model space $\mathbf{v}$ onto camera space $\hat{\mathbf{v}} = \Phi(\mathbf{v})$, the full perspective camera model $\Pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ projects the points $\hat{\mathbf{v}}$ into screen space $\mathbf{u} = \Pi(\hat{\mathbf{v}}) \in \mathbb{R}^2$.

ILLUMINATION MODEL: Under the assumption of distant smooth illumination and purely *Lambertian* surface properties, Spherical Harmonics (SH) (Ramamoorthi and Hanrahan, 2001a) is employed to represent the incident radiance at a vertex $\mathbf{v}_i$ with normal $\mathbf{n}_i$ and appearance $\mathbf{r}_i$ as

$$\mathbf{c}_i(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma}) = \mathbf{r}_i \cdot \sum_{b=1}^{B^2} \boldsymbol{\gamma}_b \cdot H_b(\mathbf{n}_i) \ . \tag{6.7}$$

The illumination parameters $\boldsymbol{\gamma} \in \mathbb{R}^{27}$ stack $B^2 = 9$ weights per color channel. Each $\boldsymbol{\gamma}_b \in \mathbb{R}^3$ controls the illumination w.r.t. the red, green and blue channel.

6.2.4 *Multi-frame Consistent Face Model Learning*

A novel network is proposed for consistent multi-frame face model learning. It consists of M Siamese towers that simultaneously process M frames of the multi-frame image in different streams, see Fig. 6.2. Each tower consists of an encoder that estimates frame-specific parameters and identity feature maps. Note that the jointly learned geometric identity $\boldsymbol{\Theta}_s$ and appearance model $(\boldsymbol{\Theta}_a, \bar{\mathbf{r}})$, which are common to all faces, are shared across streams.

REGRESSED PARAMETERS: The network is trained in a self-supervised manner based on the multi-frame images $\{\mathcal{F}_\ell\}_{\ell=1}^N$. For each frame $F_\ell^{[f]}, \forall f = 1 : M$ of the multi-frame image $\mathcal{F}_\ell$, the frame-specific parameters regressed by a Siamese tower (see *Parameter Estimation* in Fig. 6.2) are stacked in a vector $\mathbf{p}^{[f]} = (\mathbf{R}^{[f]}, \mathbf{t}^{[f]}, \boldsymbol{\gamma}^{[f]}, \boldsymbol{\delta}^{[f]})$ that parametrizes rigid pose, illumination and expression. The frame-independent person-specific identity parameters $\hat{\mathbf{p}} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ for the multi-frame image $\mathcal{F}_\ell$ are pooled from all the towers. All regressed frame-independent and frame-specific parameters of $\mathcal{F}_\ell$ are denoted as $\mathbf{p} = (\hat{\mathbf{p}}, \mathbf{p}^{[1]}, \ldots, \mathbf{p}^{[M]})$.

PER-FRAME PARAMETER ESTIMATION NETWORK: A convolutional network is employed to extract low-level features. A series of convolutions, ReLU, and fully connected layers are then applied to regress the per-frame parameters $\mathbf{p}^{[f]}$.

MULTI-FRAME IDENTITY ESTIMATION NETWORK: As explained in Section 6.2.1, each frame of the multi-frame input exhibits the same face identity under different head poses and expression. This information is exploited and a single identity estimation network (see Fig. 6.2) is used to impose the estimation of common identity

parameters $\hat{\mathbf{p}}$ (shape $\boldsymbol{\alpha}$, appearance $\boldsymbol{\beta}$) for all M frames. This way, a hard constraint on $\hat{\mathbf{p}}$ is modeled by design. More precisely, given the frame-specific low-level features obtained by the Siamese networks two additional convolution layers are applied to extract medium-level features. The resulting M medium-level feature maps are fused into a single multi-frame feature map via average pooling. Note that the average pooling operation allows us to handle a variable number of inputs. As such, monocular or multi-view reconstruction can be performed at test time, as demonstrated in Sec. 6.3. This pooled feature map is then fed to an identity parameter estimation network that is based on convolution layers, ReLU, and fully connected layers.

6.2.5 *Loss Functions*

Let $\mathbf{x} = (\mathbf{p}, \boldsymbol{\Theta})$ denote the regressed parameters $\mathbf{p}$ as well as the learnable network weights $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_s, \boldsymbol{\Theta}_a, \bar{\mathbf{r}})$. Note, $\mathbf{x}$ is fully learned during training, whereas the network infers only $\mathbf{p}$ at test time. Here, $\mathbf{p}$ is parameterized by the trainable weights of the network. To measure the reconstruction quality during mini-batch gradient descent, the following loss function is employed:

$$\mathcal{L}(\mathbf{x}) = \lambda_{\text{pho}} \cdot \mathcal{L}_{\text{pho}}(\mathbf{x}) + \lambda_{\text{lan}} \cdot \mathcal{L}_{\text{lan}}(\mathbf{x}) + \tag{6.8}$$

$$\lambda_{\text{smo}} \cdot \mathcal{L}_{\text{smo}}(\mathbf{x}) + \lambda_{\text{spa}} \cdot \mathcal{L}_{\text{spa}}(\mathbf{x}) + \lambda_{\text{ble}} \cdot \mathcal{L}_{\text{ble}}(\mathbf{x}) \ , \tag{6.9}$$

which is based on two data terms (6.8) and three regularization terms (6.9). The weights $\lambda_\bullet$ are determined empirically and kept fixed in all experiments as $\lambda_{\text{pho}} = 1.6/|\hat{\mathcal{V}}|, \lambda_{\text{lan}} = 4.7, \lambda_{\text{smo}} = 0.001, \lambda_{\text{spa}} = 1e{-}7, \lambda_{\text{ble}} = 1e{-}8$.

MULTI-FRAME PHOTOMETRIC CONSISTENCY: One of the key contributions of this chapter is to enforce multi-frame consistency of the shared identity parameters $\hat{\mathbf{p}}$. This can be thought of as solving model-based non-rigid structure-from-motion (NRSfM) on each of the multi-frame inputs during training. This is done by imposing the following photometric consistency loss with respect to the frame $F^{[f]}$:

$$\mathcal{L}_{\text{pho}}(\mathbf{x}) = \sum_{f=1}^{M} \sum_{i=1}^{|\hat{\mathcal{V}}|} \left|\left| F^{[f]}(\mathbf{u}_i(\mathbf{p}^{[f]}, \hat{\mathbf{p}})) - \mathbf{c}_i(\mathbf{p}^{[f]}, \hat{\mathbf{p}}) \right|\right|_2^2 \ .$$

Here, with abuse of notation, $\mathbf{u}_i$ is used to denote the projection of the i-th vertex into screen space, $\mathbf{c}_i$ is its rendered color, and $\hat{\mathcal{V}}$ is the set of all visible vertices, as determined by back-face culling in the forward pass. Note that the identity related parameters $\hat{\mathbf{p}}$ are shared across all frames in $\mathcal{F}$. This enables a better disentanglement of illumination and appearance, since only the illumination and head pose are allowed to change across the frames.

MULTI-FRAME LANDMARK CONSISTENCY: To better constrain the problem, a sparse 2D landmark alignment is also employed. This is based on a set of 66 automatically detected 2D feature points $\mathbf{s}_i^{[f]} \in \mathbb{R}^2$ (Saragih et al., 2011) in each frame $F^{[f]}$. Each feature point $\mathbf{s}_i^{[f]}$ comes with a confidence $\mathrm{c}_i^{[f]}$. The following loss is used:

$$\mathcal{L}_{\text{lan}}(\mathbf{x}) = \sum_{f=1}^{M} \sum_{i=1}^{66} \mathrm{c}_i^{[f]} \cdot \left|\left| \mathbf{s}_i^{[f]} - \mathbf{u}_{\mathbf{s}_i}(\mathbf{p}^{[f]}, \hat{\mathbf{p}}) \right|\right|_2^2 \ .$$

Here, $\mathbf{u}_{\mathbf{s}_i} \in \mathbb{R}^2$ is the 2D position of the i-th mesh feature point in screen space. Sliding correspondences are used, akin to Chapter 5. Note, the position of the mesh landmarks depends both on the predicted per-frame parameters $\mathbf{p}^{[f]}$ and the shared identity parameters $\hat{\mathbf{p}}$.

GEOMETRY SMOOTHNESS ON GRAPH-LEVEL: A linearized membrane energy (Botsch and Sorkine, 2008) is employed to define a first-order geometric smoothness prior on the displacements $\mathbf{t}_i(\hat{\mathbf{p}}) = \mathbf{g}_i(\hat{\mathbf{p}}) - \bar{\mathbf{g}}_i$ of the deformation graph nodes

$$\mathcal{L}_{\text{smo}}(\mathbf{x}) = \sum_{i=1}^{|\mathcal{G}|} \sum_{j \in \mathcal{N}_i} \left|\left|\mathbf{t}_i(\hat{\mathbf{p}}) - \mathbf{t}_j(\hat{\mathbf{p}})\right|\right|_2^2 , \tag{6.10}$$

where $\mathcal{N}_i$ is the set of nodes that have a skinned vertex in common with the i-th node. Note, the graph parameterizes the geometric identity, i.e., it only depends on the shared identity parameters $\hat{\mathbf{p}}$. This term enforces smooth deformations of the parametric shape and leads to higher quality reconstruction results.

APPEARANCE SPARSITY: In the learned face model, skin appearance is parameterized on a per-vertex basis. To further constrain the underlying intrinsic decomposition problem, a local per-vertex spatial reflectance sparsity prior is employed as in Bonneel et al. (2014) and Meka et al. (2016), defined as follows

$$\mathcal{L}_{\text{spa}}(\mathbf{x}) = \sum_{i=1}^{|\mathcal{V}|} \sum_{j \in \mathcal{N}_i} w_{ij} \cdot \left|\left|\mathbf{r}_i(\hat{\mathbf{p}}) - \mathbf{r}_j(\hat{\mathbf{p}})\right|\right|_2^p . \tag{6.11}$$

The per-edge weights w_{ij} model the similarity of neighboring vertices in terms of chroma and are defined as

$$w_{ij} = \exp\{\left[-\eta \cdot ||\mathbf{h}_i(\hat{\mathbf{p}}_{\text{old}}) - \mathbf{h}_j(\hat{\mathbf{p}}_{\text{old}})||_2\right]\} .$$

Here, $\mathbf{h}_i$ is the chroma of $\mathbf{c}_i$ and $\hat{\mathbf{p}}_{\text{old}}$ denotes the parameters predicted in the last forward pass. Hyperparameters are fixed as $\eta = 80$ and $p = 0.9$.

EXPRESSION REGULARIZATION: To prevent over-fitting and enable a better learning of the identity basis, the magnitude of the expression parameters $\boldsymbol{\delta}$ is regularized:

$$\mathcal{L}_{\text{ble}}(\mathbf{x}) = \sum_{f=1}^{M} \sum_{u=1}^{|\boldsymbol{\delta}^{[f]}|} \left(\frac{\boldsymbol{\delta}_u^{[f]}}{\sigma_{\boldsymbol{\delta} u}}\right)^2 . \tag{6.12}$$

Here, $\boldsymbol{\delta}_u^{[f]}$ is the u-th expression parameter of frame f, and $\sigma_{\boldsymbol{\delta} u}$ is the corresponding standard deviation computed based on Principal Component Analysis (PCA).

6.3 RESULTS

Fig. 6.4 shows qualitative results reconstructing geometry, reflectance and scene illumination from monocular images. As the model is trained on a large corpus of multi-frame images, it generalizes well to different ethnicities, even in the presence of facial hair and makeup. The networks are implemented and trained in TensorFlow (Abadi et al., 2015). The expression model is first pretrained and then the full

Figure 6.4: The proposed approach produces high-quality monocular reconstructions of facial geometry, reflectance and illumination by learning an optimal model from in-the-wild data. This enables reconstruction of facial hair and makeup.

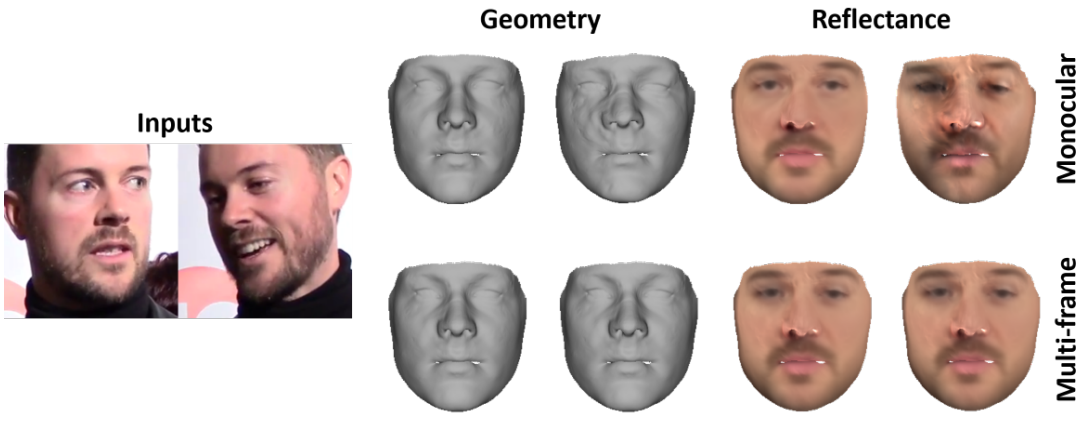

Figure 6.5: Monocular vs. multi-frame reconstruction. For clarity, all results are shown with a frontal pose and neutral expression. Multi-frame reconstruction improves consistency and quality especially in regions which are occluded in one of the images.

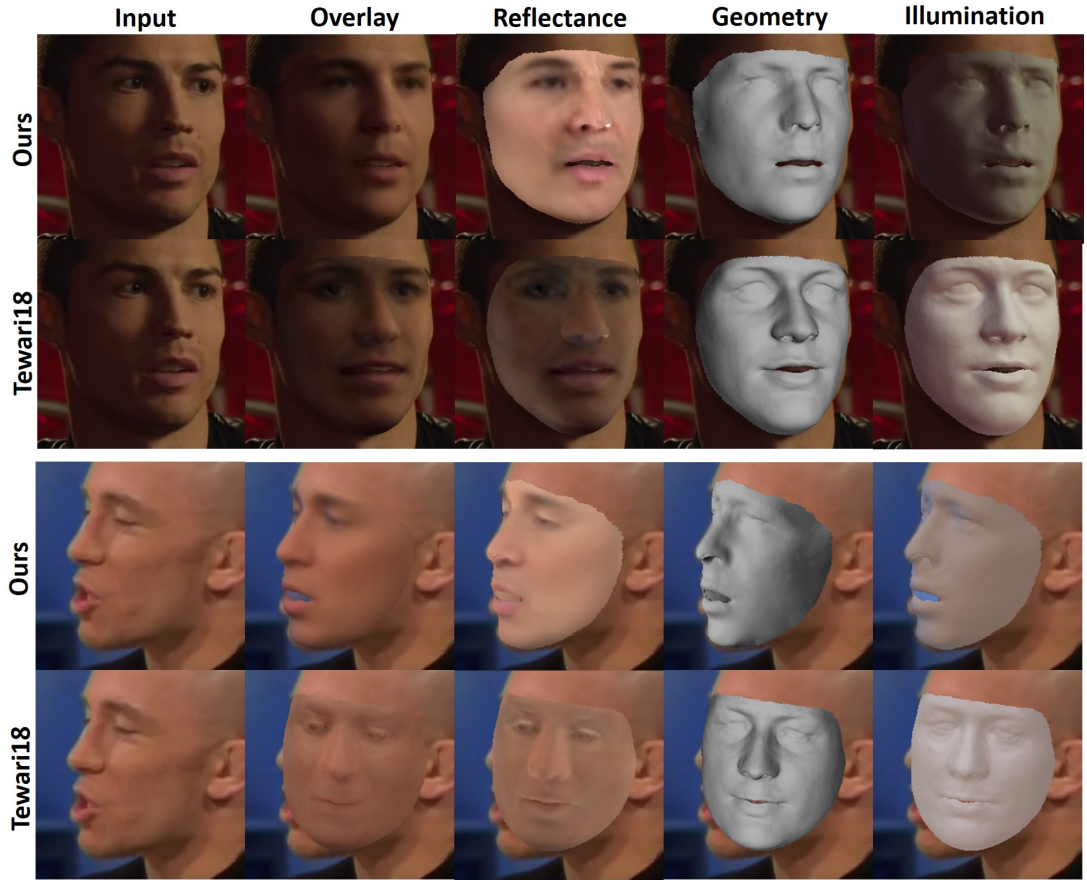

Figure 6.6: Comparison to Tewari et al. (2018), the method presented in Chapter 5. Multi-frame based training improves illumination estimation. The proposed approach also outperforms that of Tewari et al. (2018) under large poses.

network is trained end-to-end. After convergence, the network is fine-tuned using a larger learning rate for reflectance. This training strategy improves the capture of facial hair, makeup and eyelids, and thus the model's generalization. The method can also be applied to multi-frame reconstruction at test time. Fig. 6.5 shows that feeding two images simultaneously improves the consistency and quality of the obtained 3D reconstructions when compared to the monocular case. Please note that the approach can successfully separate identity and reflectance due to the novel Orthogonal Complement Layer (OCL). For the experiments shown in the following sections, the network is trained on $M = 4$ multi-frame images and only one input image is used at test time, unless stated otherwise. The networks take around 30 hours to train. Inference takes only 5.2 ms on a Titan Xp.

6.3.1 *Comparisons to Monocular Approaches*

State-of-the-art monocular reconstruction approaches that rely on an existing face model (Tewari et al., 2017) (presented in Chapter 4), or synthetically generated data (Richardson et al., 2017; Sela et al., 2017) during training do not generalize well to faces outside the span of the model. As such, they can not handle facial hair, makeup, and unmodeled expressions, see Fig. 6.7. Since the models in this chapter are

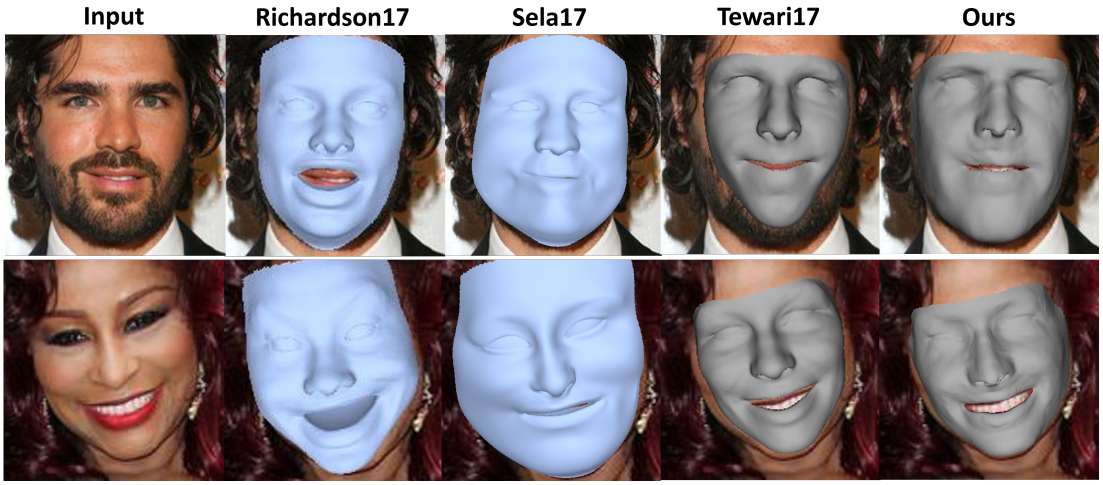

Figure 6.7: Comparison to Richardson et al. (2017), Sela et al. (2017), and Tewari et al. (2017). These approaches are constrained by the (synthetic) training corpus and/or underlying 3D face model. The optimal learned model of this chapter produces more accurate results, since it is learned from a large corpus of real images.

Table 6.1: Geometric reconstruction error on the BU-3DFE dataset (Yin et al., 2006). The proposed approach produces higher quality results than the current state of the art. The approach of Tewari et al. (2017) does not generalize to the ±45 degree head poses contained in this dataset.

	Ours					Tewari et al. (2018) Fine	Tewari et al. (2018) Coarse	Tewari et al. (2017)
Train	M = 1	M = 2	M = 4	M = 2	M = 4			
Test	M = 1	M = 1	M = 1	M = 2	M = 2			
Mean	1.92 mm	1.86 mm	1.79 mm	1.85 mm	**1.78 mm**	1.83 mm	1.81 mm	3.22 mm
SD	0.48 mm	0.47 mm	0.45 mm	0.50 mm	0.45 mm	0.39 mm	0.47 mm	0.77 mm

Table 6.2: Geometric error on FaceWarehouse (Cao et al., 2013). The proposed approach competes with Tewari et al. (2018) and Tewari et al. (2020d), and outperforms Tewari et al. (2017) and Kim et al. (2018b). Note, in contrast to these approaches, the proposed approach does not require a precomputed face model during training, but learns it from scratch. It comes close to the off-line high-quality approach of Garrido et al. (2016a), while being orders of magnitude faster and not requiring feature detection.

	Ours	Others					
	Learning	Learning				Optimization	Hybrid
		Tewari et al. (2018) Fine	Tewari et al. (2018) Coarse	Tewari et al. (2017)	Kim et al. (2018b)	Garrido et al. (2016a)	Tewari et al. (2020d)
Mean	**2.01 mm**	1.84 mm	2.03 mm	2.19 mm	2.11 mm	**1.59 mm**	1.87 mm
SD	0.41 mm	0.38 mm	0.52 mm	0.54 mm	0.46 mm	0.30 mm	0.42 mm
Time	5.2 ms	**4 ms**	**4 ms**	**4 ms**	**4 ms**	120 s	110 ms

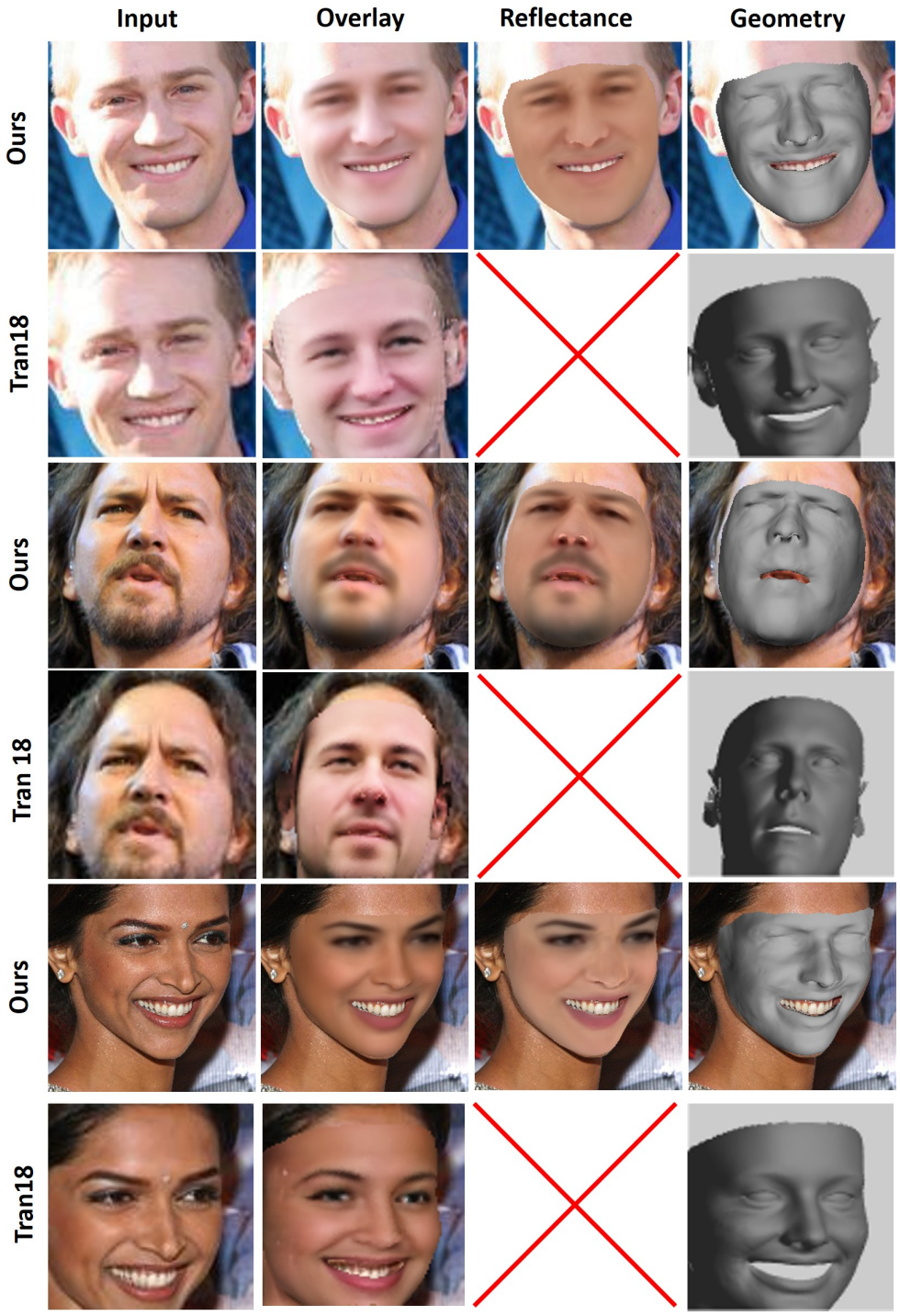

Figure 6.8: In contrast to Tran and Liu (2018a), the proposed approach estimates better geometry and separates reflectance from illumination. Note, the approach of Tran and Liu, 2018a does not disentangle reflectance and shading.

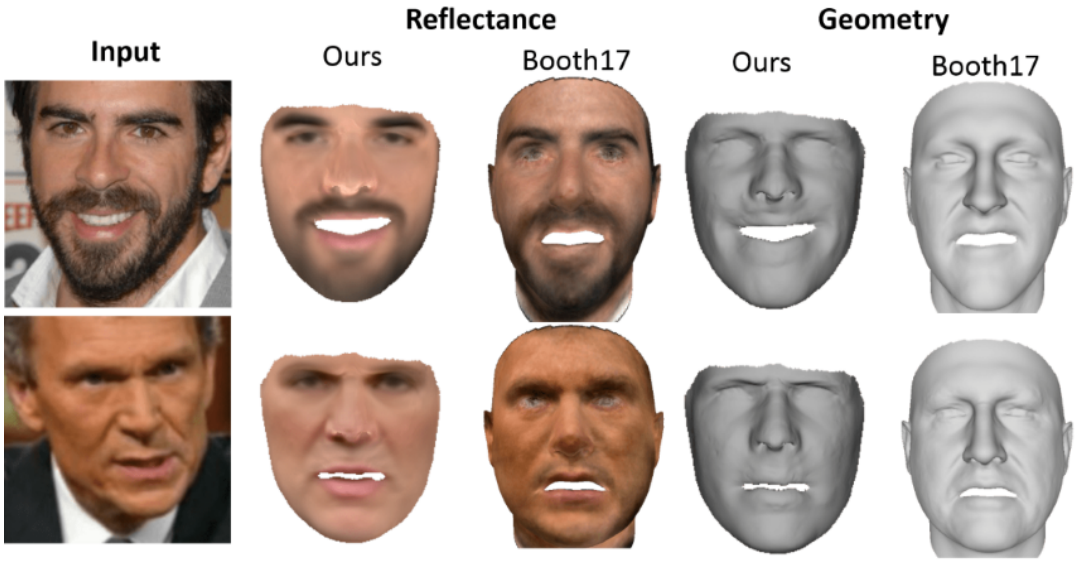

Figure 6.9: In contrast to the texture model of Booth et al. (2017) that contains shading, the proposed approach estimates a reflectance model.

trained on in-the-wild videos, these variations are captured leading to better generalization in such challenging cases. The approach is also compared to the refinement based approaches of Tewari et al. (2018) (presented in Chapter 5, and Tran and Liu (2018a). Tran and Liu (2018a) (see Fig 6.8) refine a 3DMM (Blanz and Vetter, 1999) based on in-the-wild data. The proposed approach produces better geometry without requiring a 3DMM and, contrary to Tran and Liu (2018a), it also separates albedo from illumination. The approach of Tewari et al. (2018), presented in Chapter 5 (see Fig 6.6), requires a 3DMM (Blanz and Vetter, 1999) as input and only learns shape and reflectance correctives. Since they learn from monocular data, their correctives are prone to artifacts, especially when occlusions or extreme head poses exist. In contrast, the proposed approach learns a complete model from scratch based on multi-frame supervision, thus improving robustness and reconstruction quality. Comparisons to Booth et al., 2017 can be seen in Fig. 6.9. Booth et al., 2017 only learn a texture model. In contrast, the proposed approach learns a model that separates albedo from illumination. Besides, their method needs a 3DMM (Blanz and Vetter, 1999) as initialization, while the proposed approach starts from a single constantly colored mesh and learns all variation modes (geometry and reflectance) from scratch.

6.3.2 *Quantitative Results*

The reconstructions are quantitatively evaluated on a subset of the BU-3DFE dataset (Yin et al., 2006), see Tab. 6.1. This dataset contains images and corresponding ground truth geometry of multiple people performing a variety of expressions. It includes two different viewpoints. The importance of multi-frame training is evaluated in the case of monocular reconstruction using per-vertex root mean squared error based on a pre-computed dense correspondence map. The lowest error is achieved with multi-view supervision during training, in comparison to monocular input data. Multi-view supervision can better resolve depth ambiguity and thus learn a more accurate

model. In addition, the multi-view supervision also leads to a better disentanglement of reflectance and shading. The advantage of multi-frame input at test time is also evaluated. When both images corresponding to a shape are given, consistently better results are obtained. Further, the estimates are better than the state-of-the-art approach of Tewari et al. (2018), presented in Chapter 5. Since Tewari et al. (2018) refine an existing 3DMM only using monocular images during training, it cannot resolve depth ambiguity well. Thus, it does not improve the performance compared to their coarse model on the ±45 degree poses of BU-3DFE (Yin et al., 2006). Similar to previous work, monocular reconstruction is evaluated on 180 meshes of FaceWarehouse (Cao et al., 2013), see Tab. 6.2. The approach performs similar to the 3DMM-based state of the art. Note that a precomputed 3DMM is not used, but a model is learned from scratch during training, unlike all other approaches in this comparison. For this test, a model learned starting from an Asian mean face is employed, as FaceWarehouse mainly contains Asians.

6.4 LIMITATIONS

Figure 6.10: Limitations of the proposed approach. From top to bottom: Extreme illumination conditions, severe occlusions by accessories, and thick facial hair.

The proposed approach still has a few limitations that can be addressed in follow-up work, see Fig. 6.10. Overall, the approach can deal with large head poses quite well. Still, reconstructing extreme poses is a hard task in itself that challenges all face reconstruction techniques. Occlusions, for example, by accessories or thick facial hair might adversely impact the reconstruction quality of our approach. Facial hair, such as beards are modeled in the reflectance channel, and thus are not reconstructed in a physically correct manner. Even though the multi-frame supervision approach can obtain quite clean reflectance estimates that are free of shading, there is still a

remaining global scale ambiguity between illumination and reflectance. As such, the global skin tone can not be reliably disentangled from the general ambient brightness of the illumination. Strong and colorful directional illumination outside the norm might also harm the estimation of 3D faces. Specular reflections and cast shadows are currently not modeled by the differentiable renderer, and thus they might be baked into the reflectance channel. Non-standard facial shapes challenge the approach.

6.5 CONCLUSION & DISCUSSION

This chapter proposed a self-supervised approach for joint multi-frame learning of a face model and a 3D face reconstruction network. The model is learned from scratch based on a large corpus of in-the-wild video clips without available ground truth. This chapter showed for the first time that 2D data could be used to learn morphable models. While the approach required the use of an expression model, recent work (B R et al., 2021b) has shown that all modes of the morphable model can be learned from videos and images.

The methods presented in Chapters. 4- 6 show entirely new ways of reconstructing 3D faces from images, and jointly learning 3D morphable models from images and videos. However, there are several limitations of these approaches. They do not explain the full head, hair and torso, and in addition cannot capture the high-frequency details due to their low dimensionality. Several approximations are made in the methods, for example, faces are assumed to be diffuse which is not sufficient for photorealistic rendering. However, these methods do offer very meaningful semantic control over the reconstructions. We can independently change the identity, expressions, albedo, and illumination in the scene. In the next chapter, we will demonstrate how to methodically integrate the rendering and reconstruction concepts from the first three chapters with generative adversarial networks (GANs). GANs can synthesize very high quality portrait images, including hair, torso, and with complex global illumination effects such as sub-surface scattering. However, they cannot be semantically controlled. The integration of the reconstruction and rendering pipeline presented in these chapters with GANs will result in a method which can synthesize photorealistic images with semantic control.

7

STYLERIG: RIGGING STYLEGAN FOR 3D CONTROL OVER PORTRAIT IMAGES

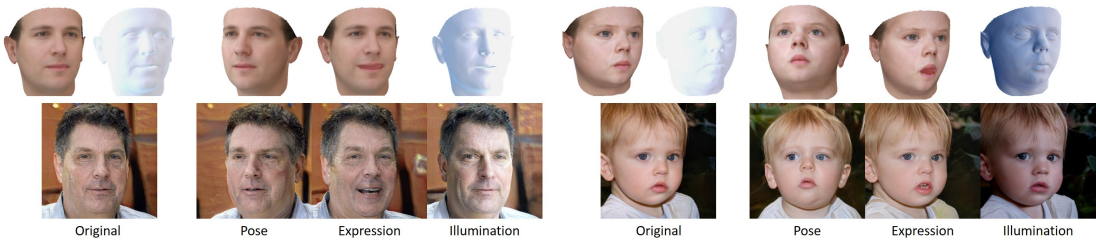

Figure 7.1: StyleRig allows for a face rig-like control over StyleGAN generated portrait images, by translating semantic edits on 3D face meshes to the input space of StyleGAN.

StyleGAN (Karras et al., 2019a) generates photorealistic portrait images of faces with eyes, teeth, hair and context (neck, shoulders, background), but lacks a rig-like control. The face rig of a character is its 3D representation which includes semantic control, such as head pose, expressions, and scene illumination. Three-dimensional morphable face models (3DMMs) (Egger et al., 2020) on the other hand offer control over the semantic parameters, but lack photorealism when rendered and only model the face interior, not other parts of a portrait image (hair, mouth interior, background). This chapter (published as Tewari et al. (2020b)) presents the first method to provide a face rig-like control over a pretrained and fixed StyleGAN via a 3DMM. A new rigging network, *RigNet* is trained between the 3DMM's semantic parameters and StyleGAN's input. The network is trained in a self-supervised manner, without the need for manual annotations. At test time, the method generates portrait images with the photorealism of StyleGAN and provides explicit control over the 3D semantic parameters of the face, see Fig. 7.1.

7.1 INTRODUCTION

Photorealistic synthesis of portrait face images finds many applications in several fields including special effects, extended reality, virtual worlds, and next-generation communication. During the content creation process for such applications, artist control over the *face rig's* semantic parameters, such as geometric identity, expressions, reflectance, or scene illumination is desired. The computer vision and graphics communities have a rich history of modeling face rigs (Li et al., 2017; Richardson et al., 2017; Sanyal et al., 2019; Tewari et al., 2019). These models provide artist-friendly control (often called a face rig), while navigating the various parameters of a morphable face model (3DMM) (Blanz et al., 2003; Blanz and Vetter, 1999). Such methods are often limited by the lack of training data, and more importantly, lack of photorealism in the final rendering.

Through 3D face scanning techniques high-quality face geometry datasets can be obtained (Booth et al., 2016; Li et al., 2017). However, models derived from these datasets are bound by the diversity of faces scanned and may limit the generalization over the rich set of human faces' semantic parameterization. Further, deep learning-based models trained on *in-the-wild* data (Tewari et al., 2019, 2018; Tran et al., 2019) also often rely on data-driven priors and other forms of regularization obtained from scan-based datasets. With respect to photorealism, perceptual losses recently showed an improvement of face modeling quality (Deng et al., 2019; Tran et al., 2019) over existing methods. However, they still do not engender photorealistic face renders. Mouth interiors, hair, or eyes, let alone image background are often not modeled by such approaches. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have lately achieved photorealism (Isola et al., 2017; Karras et al., 2018), especially for faces. Karras et al. (2018) show that through a progressive growth of a GAN's generator and discriminator, one can better stabilize and speed up training. When trained on the CelebA-HQ (Karras et al., 2018) dataset, this yields a remarkable level of photorealism for faces. Their approach also shows how photorealistic face images of non-existent people can be sampled from the learned GAN distribution. Building on Karras et al. (2018), StyleGAN (Karras et al., 2019a) uses ideas from the style transfer literature (Gatys et al., 2016; Selim et al., 2016) and proposes an architecture capable of disentangling various face attributes. Promising results of control over various attributes, including coarse (hair, geometry), medium (expressions, facial hair) and fine (color distribution, freckles) attributes were shown. However, these controllable attributes are not semantically well defined, and contain several similar yet entangled semantic attributes. For example, both coarse and medium level attributes contain face identity information. In addition, the coarse levels contain several entangled attributes such as face identity and head pose.

This chapter presents a novel solution to *rig* StyleGAN using a semantic parameter space for faces. The presented approach brings the best of both worlds: the controllable parametric nature of existing morphable face models (Sanyal et al., 2019; Tewari et al., 2019), and the high photorealism of generative face models (Karras et al., 2018, 2019a). A fixed and pretrained StyleGAN is employed without the need for more data for training. The focus is to provide computer graphics style rig-like control over the various semantic parameters. The novel training procedure is based on a self-supervised two-way cycle consistency loss that is empowered by the combination of a face reconstruction network with a differentiable renderer. This allows for measuring the photometric rerendering error in the image domain and leads to high quality results. The chapter demonstrates compelling results of the method, including interactive control of StyleGAN generated imagery as well as image synthesis conditioned on well-defined semantic parameters.

7.2 OVERVIEW

StyleGAN (Karras et al., 2019a) can be seen as a function that maps a latent code $\mathbf{w} \in \mathbb{R}^l$ to a realistic portrait image $\mathbf{I_w} = \mathit{StyleGAN}(\mathbf{w}) \in \mathbb{R}^{3\times w\times h}$ of a human face. While the generated images are of very high quality and at a high resolution ($w = h = 1024$), there is no semantic control over the generated output, such as the head pose, expression, or illumination. StyleRig allows us to obtain a rig-like control

over StyleGAN-generated facial imagery in terms of semantic and interpretable control parameters (Sec. 7.7). The following sections explain the semantic control space (Sec. 7.3), training data (Sec. 7.4), network architecture (Sec. 7.5) and loss function (Sec. 7.6).

7.3 SEMANTIC RIG PARAMETERS

The proposed approach uses a parametric face model to achieve an explicit rig-like control of StyleGAN-generated imagery based on a set of semantic control parameters. The control parameters are a subset of $\mathbf{p} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{R}, \mathbf{t}) \in \mathbb{R}^f$, which describes the facial shape $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, skin reflectance $\boldsymbol{\beta} \in \mathbb{R}^{80}$, facial expression $\boldsymbol{\delta} \in \mathbb{R}^{64}$, scene illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$, head rotation $\mathbf{R} \in SO(3)$, and translation $\mathbf{t} \in \mathbb{R}^3$, with the dimensionality of $\mathbf{p}$ being $f = 257$. The control space for the facial shape $\boldsymbol{\alpha}$ and skin reflectance $\boldsymbol{\beta}$ is defined using two low-dimensional affine models that have been computed via Principal Component Analysis (PCA) based on 200 (100 male, 100 female) scans of human faces (Blanz and Vetter, 1999). The output of this model is represented by a triangle mesh with $53k$ vertices and per-vertex color information. The control space for the expression $\boldsymbol{\delta}$ is given in terms of an additional affine model that captures the expression dependent displacement of the vertices. This model is obtained by applying PCA to a set of blendshapes (Alexander et al., 2009; Cao et al., 2013) which have been transferred to the topology of the shape and reflectance models. The affine models for shape, appearance, and expression cover more than 99% of the variance in the original datasets. Illumination $\boldsymbol{\gamma}$ is modeled based on three bands of spherical harmonics per color channel leading to an additional 27 parameters.

7.4 TRAINING CORPUS

Besides the parametric face model, the proposed approach requires a set of face images $\mathbf{I}_\mathbf{w}$ and their corresponding latent codes $\mathbf{w}$ as training data. To this end, $N = 200$k latent codes $\mathbf{w} \in \mathbb{R}^l$ are sampled and the corresponding photorealistic face images $\mathbf{I}_\mathbf{w} = \mathit{StyleGAN}(\mathbf{w})$ are generated using a pretrained StyleGAN network. The $l = 18 \times 512$ dimensional $\mathcal{W}+$ latent space is used, which has been shown to be more disentangled than the $\mathcal{W}$ space (Abdal et al., 2019; Karras et al., 2019a). Here, 18 latent vectors of size 512 are used at different resolutions. Each training sample is generated by combining up to 5 separately sampled latent vectors, similar to the mixing regularizer in Karras et al. (2019a). This allows the networks to reason independently about the latent vectors at different resolutions. Given these $(\mathbf{w}, \mathbf{I}_\mathbf{w})$ pairs, the approach can be trained in a self-supervised manner without requiring any additional image data or manual annotations.

7.5 NETWORK ARCHITECTURE

Given a latent code $\mathbf{w} \in \mathbb{R}^l$ that corresponds to an image $\mathbf{I}_\mathbf{w}$, and a vector $\mathbf{p} \in \mathbb{R}^f$ of semantic control parameters, the goal is to learn a function that outputs a modified latent code $\hat{\mathbf{w}} = \mathit{RigNet}(\mathbf{w}, \mathbf{p})$. The modified latent code $\hat{\mathbf{w}}$ should map to a modified face image $\mathbf{I}_{\hat{\mathbf{w}}} = \mathit{StyleGAN}(\hat{\mathbf{w}})$ that obeys the control parameters $\mathbf{p}$. One example would be changing the rotation of the face in an image such that it matches a given

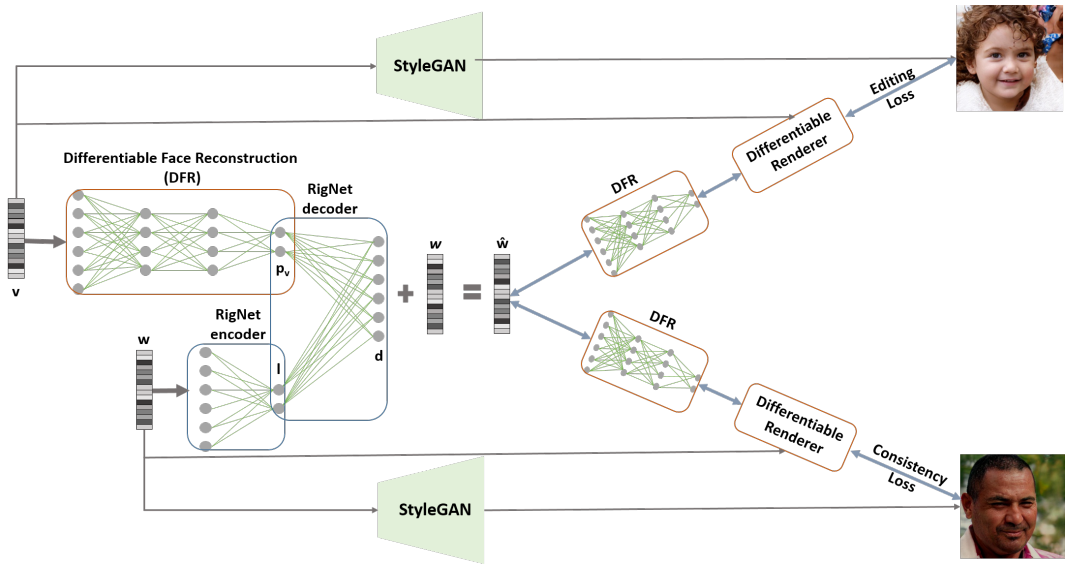

Figure 7.2: StyleRig enables a rig-like control over StyleGAN-generated facial imagery based on a learned rigger network (RigNet). To this end, a self-supervised training approach is employed based on a differentiable face reconstruction (DFR) and a neural face renderer (StyleGAN). The DFR and StyleGAN networks are pretrained and their weights are fixed, only RigNet is trainable. The consistency and edit losses are defined in the image domain using a differentiable renderer.

Figure 7.3: Differentiable Face Reconstruction. Visualized are (image, reconstruction) pairs. The network, however, only gets the latent vector corresponding to the images as input.

target rotation, while maintaining the facial identity, expression, and scene illumination (see Sec. 7.7 for examples). Separate RigNet networks are trained for the different modes of control i.e., pose, expressions and illumination. RigNet is implemented based on a linear two-layer perceptron (MLP). This chapter proposes a self-supervised training of RigNet based on two-way cycle consistency losses and a differentiable face reconstruction (DFR) network. Fig. 7.2 shows an overview of the architecture. The network combines several components that fulfill specific tasks.

DIFFERENTIABLE FACE RECONSTRUCTION One key component is a pretrained differentiable face reconstruction (DFR) network. This parameter regressor is a function $\mathcal{F} : \mathbb{R}^l \rightarrow \mathbb{R}^f$ that maps a latent code $\mathbf{w}$ to a vector of semantic control parameters $\mathbf{p_w} = \mathcal{F}(\mathbf{w})$. In practice, $\mathcal{F}$ is modeled using a three layer MLP with ELU activations after every intermediate layer, and is trained in a self-supervised manner. This requires a differentiable render layer $\mathcal{R} : \mathbb{R}^f \rightarrow \mathbb{R}^{3\times w\times h}$ that takes a face parameter vector $\mathbf{p}$ as

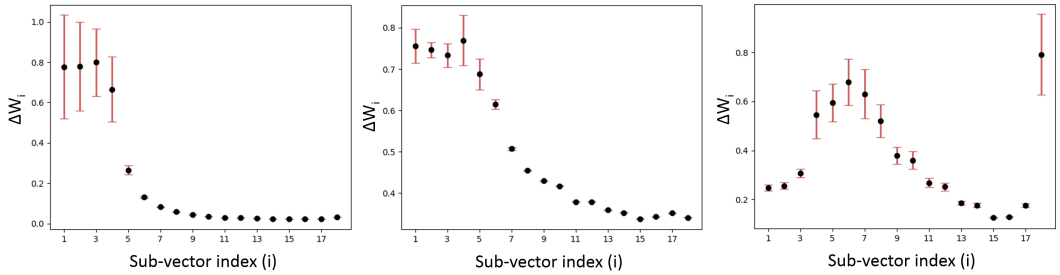

Figure 7.4: Change of latent vectors at different resolutions. Coarse vectors are responsible for rotation (left), medium for expressions (middle), medium and fine for illumination (right).

input, converts it into a 3D mesh and generates a synthetic rendering $\mathbf{S_w} = \mathcal{R}(\mathbf{p_w})$ of the face[1]. $\mathcal{F}$ is then trained using a rerendering loss:

$$\mathcal{L}_{\text{render}}(\mathbf{I_w}, \mathbf{p}) = \mathcal{L}_{\text{photo}}(\mathbf{I_w}, \mathbf{p}) + \lambda_{\text{land}}\mathcal{L}_{\text{land}}(\mathbf{I_w}, \mathbf{p}) \ . \tag{7.1}$$

The first term is a dense photometric alignment loss:

$$\mathcal{L}_{\text{photo}}(\mathbf{I_w}, \mathbf{p}) = \left\| \mathbf{M} \odot (\mathbf{I_w} - \mathcal{R}(\mathbf{p})) \right\|_2^2 \ .$$

Here, $\mathbf{M}$ is a binary mask with all pixels where the face mesh is rendered set to 1 and $\odot$ is element-wise multiplication. A sparse landmark loss is also used

$$\mathcal{L}_{\text{land}}(\mathbf{I_w}, \mathbf{p}) = \left\| \mathbf{L}_{\mathbf{I_w}} - \mathbf{L_M} \right\|_2^2 \ ,$$

where $\mathbf{L}_{\mathbf{I_w}} \in \mathbb{R}^{66\times 2}$ are 66 automatically computed landmarks (Saragih et al., 2011) on the image $\mathbf{I_w}$, and $\mathbf{L_M}$ are the corresponding landmark positions on the rendered reconstructed face. The landmark vertices on the mesh are manually annoted before training. λ_{land} is a fixed weight used to balance the loss terms. In addition, statistical regularization is also employed on the parameters of the face model, as done in MoFA (Chapter 4. After training, the weights of $\mathcal{F}$ are fixed. Fig. 7.3 shows some results of the reconstructions obtained by DFR.

RIGNET ENCODER The encoder takes the latent vector $\mathbf{w}$ as input and linearly transforms it into a lower dimensional vector $\mathbf{l}$ of size 18×32. Each sub-vector $\mathbf{w}_i$ of $\mathbf{w}$ of size 512 is independently transformed into a sub-vector $\mathbf{l}_i$ of size 32, for all $i \in \{0, \dots, 17\}$.

RIGNET DECODER The decoder tranforms $\mathbf{l}$ and the input control parameters $\mathbf{p}$ into the output $\hat{\mathbf{w}}$. Similar to the encoder, independent linear decoders are used for each $\mathbf{l}_i$. Each layer first concatenates $\mathbf{l}_i$ and $\mathbf{p}$, and transforms it into $\mathbf{d}_i$, for all $i \in \{0, \dots, 17\}$. The final output is computed as $\hat{\mathbf{w}} = \mathbf{d} + \mathbf{w}$.

7.6 SELF-SUPERVISED TRAINING

The goal is to train RigNet such that a subset of parameters can be injected into a given latent code $\mathbf{w}$. For example, one might want to inject a new head pose, while

1 Point-based rendering of the mesh vertices is used.

maintaining the facial identity, expression, and illumination in the original image synthesized from $\mathbf{w}$. The following loss function is employed for training:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{edit}} + \mathcal{L}_{\text{consist}} \ . \tag{7.2}$$

It consists of a reconstruction loss $\mathcal{L}_{\text{rec}}$, an editing loss $\mathcal{L}_{\text{edit}}$, and a consistency loss $\mathcal{L}_{\text{consist}}$. Since there is no ground truth for the desired modifications (the training corpus only contains one image per person), self-supervision is employed based on cycle-consistent editing and consistency losses. $\mathcal{L}_{\text{total}}$ is optimized based on AdaDelta (Zeiler, 2012) with a learning rate of 0.01. In the following, more details are provided.

RECONSTRUCTION LOSS RigNet is designed such that it reproduces the latent codes in the training corpus. Formally, the goal is to ensure $\mathit{RigNet}(\mathbf{w}, \mathcal{F}(\mathbf{w})) = \mathbf{w}$. This is enforced with the following ℓ_2-loss:

$$\mathcal{L}_{\text{rec}} = \left\| \mathit{RigNet}(\mathbf{w}, \mathcal{F}(\mathbf{w})) - \mathbf{w} \right\|_2^2 \ .$$

This constraint anchors the learned mapping at the right location in the latent space. Without this constraint, learning the mapping is underconstrained, which leads to a degradation in the image quality (see Sec. 7.7). Since $\mathcal{F}$ is pretrained and not updated, the semantics of the control space are enforced.

CYCLE-CONSISTENT PER-PIXEL EDITING LOSS Given two latent codes, $\mathbf{w}$ and $\mathbf{v}$ with corresponding images $\mathbf{I}_\mathbf{w}$ and $\mathbf{I}_\mathbf{v}$, the semantic parameters of $\mathbf{v}$ are transferred to $\mathbf{w}$ during training. First, the target parameter vector $\mathbf{p}_\mathbf{v} = \mathcal{F}(\mathbf{v})$ is extracted using the differentiable face reconstruction network. Next, a subset of the parameters of $\mathbf{p}_\mathbf{v}$ (the ones which will be modified) is injected into the latent code $\mathbf{w}$ to yield a new latent code $\hat{\mathbf{w}} = \mathit{RigNet}(\mathbf{w}, \mathbf{p}_\mathbf{v})$, so that $\mathbf{I}_{\hat{\mathbf{w}}} = \mathit{StyleGAN}(\hat{\mathbf{w}})$ (ideally) corresponds to the image $\mathbf{I}_\mathbf{w}$, modified according to the subset of the parameters of $\mathbf{p}_\mathbf{v}$. For example, $\hat{\mathbf{w}}$ might retain the facial identity, expression and scene illumination of $\mathbf{w}$, but should perform the head rotation specified in $\mathbf{p}_\mathbf{v}$.

Since there is no access to the ground truth for such a modification, i.e., the image $\mathbf{I}_{\hat{\mathbf{w}}}$ is unknown, supervision based on a cycle-consistent editing loss is employed. The editing loss enforces that the latent code $\hat{\mathbf{w}}$ contains the modified parameters. This is enforced by mapping from the latent to the parameter space $\hat{\mathbf{p}} = \mathcal{F}(\hat{\mathbf{w}})$. The regressed parameters $\hat{\mathbf{p}}$ should have the same rotation as $\mathbf{p}_\mathbf{v}$. This could be directly measured in the parameter space but this has been shown to not be very effective (Tewari et al., 2017).

Instead, a rerendering loss is employed similar to the one used for differentiable face reconstruction. The original target parameter vector $\mathbf{p}_\mathbf{v}$ is taken and its rotation parameters are replaced with the regressed rotation from $\hat{\mathbf{p}}$, resulting in $\mathbf{p}_{\text{edit}}$. This can now be compared to $\mathbf{I}_\mathbf{v}$ using the rerendering loss (see Eq. 7.1):

$$\mathcal{L}_{\text{edit}} = \mathcal{L}_{\text{render}}(\mathbf{I}_\mathbf{v}, \mathbf{p}_{\text{edit}}) \ .$$

No regularization terms are used here. Such a loss function ensures that the rotation component of $\mathbf{p}_{\text{edit}}$ aligns with $\mathbf{I}_\mathbf{v}$, which is the desired output. The component of $\mathbf{p}_\mathbf{v}$ which is replaced from $\hat{\mathbf{p}}$ depends on the property being changed. It could either be the pose, expression, or illumination parameters.

CYCLE-CONSISTENT PER-PIXEL CONSISTENCY LOSS In addition to the editing loss, consistency of the parameters that should not be changed by the performed edit operation is enforced. The regressed parameters $\hat{\mathbf{p}}$ should have the same unmodified parameters as $\mathbf{p}_\mathbf{w}$. Similarly as above, this is imposed in terms of a rerendering loss. The original parameter vector $\mathbf{p}_\mathbf{w}$ is taken and all parameters that should not be modified by the regressed ones are replaced from $\hat{\mathbf{p}}$, resulting in $\mathbf{p}_{\text{consist}}$. In the case of modifying rotation values, the parameters that should not be changed are expression, illumination as well as identity parameters (shape and skin reflectance). This leads to the loss function:

$$\mathcal{L}_{\text{consist}} = \mathcal{L}_{\text{render}}(\mathbf{I}_\mathbf{w}, \mathbf{p}_{\text{consist}}) \ .$$

SIAMESE TRAINING Since there are two sampled latent codes $\mathbf{w}$ and $\mathbf{v}$ during training, the same operations are also performed in a reverse order, i.e., in addition to injecting $\mathbf{p}_\mathbf{v}$ into $\mathbf{w}$, $\mathbf{p}_\mathbf{w}$ is also injected into $\mathbf{v}$. To this end, a Siamese network is used with two towers that have shared weights. This results in a two-way cycle consistency loss.

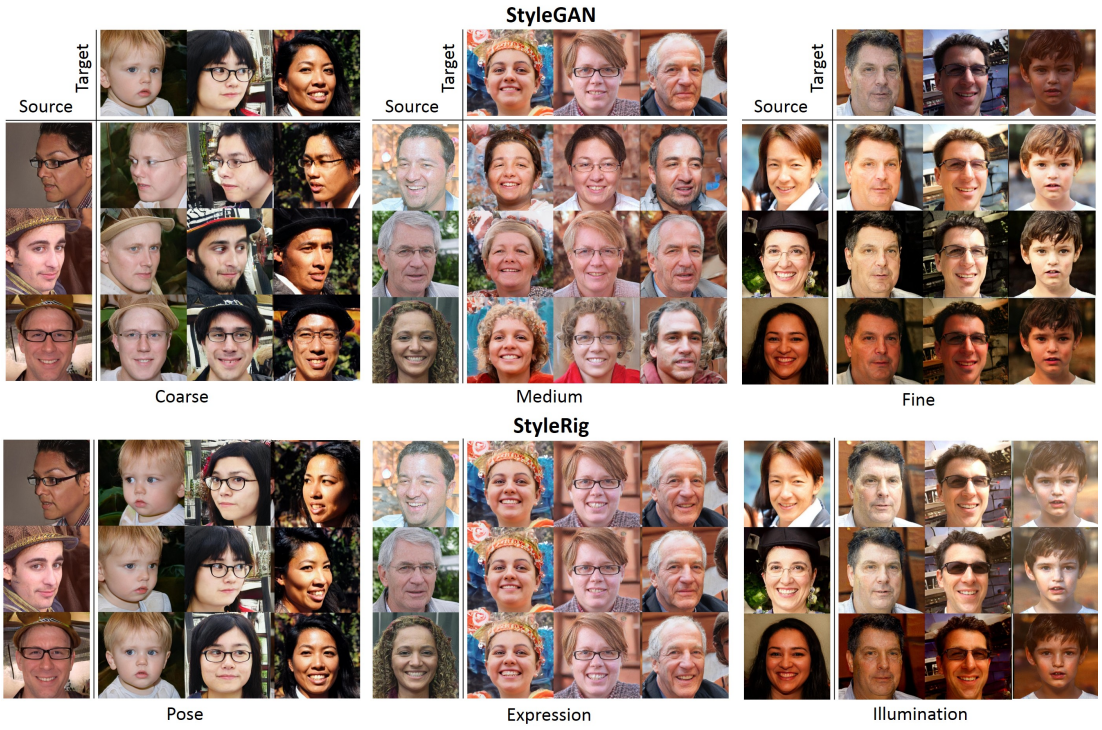

Figure 7.5: Mixing between source and target images generated by StyleGAN. For StyleGAN, the latent vectors of the source samples (rows) are copied to the target vectors (columns). StyleRig allows us to mix semantically meaningful parameters, i.e., head pose, expressions and scene illumination. These parameters can be copied over from the source to target images.

7.7 RESULTS

At test time, StyleRig allows control over the pose, expression, and illumination parameters of StyleGAN generated images. The efficacy of the approach is demon-

strated with three applications: Style Mixing (7.7.2), Interactive Rig Control (7.7.3) and Conditional Image Generation (7.7.4).

7.7.1 *Training Details*

The hyperparameters are empirically determined as $\lambda_{land} = 17.5$ for pose editing, $\lambda_{land} = 100.0$ for expression editing and $\lambda_{land} = 7.8$ for illumination editing networks. The same hyperparameters are used for both the editing and consistency losses. For networks with simultaneous control, the loss functions for the different parameters are weighed differently. Rotation losses are weighted by 1.0, expression by 1000.0 and illumination by 0.001. As before, the weights for both the editing and the consistency losses are equal.

7.7.2 *Style Mixing*

Karras et al. (2019a) show StyleGAN vectors at different scales that correspond to different *styles*. To demonstrate *style mixing*, latent vectors at certain resolutions are copied from a source to a target image, and new images are generated. As shown in Fig. 7.5, coarse styles contain information about the pose as well as identity, medium styles include information about expressions, hair structure, and illumination, while fine styles include the color scheme of the source. The proposed approach allows for a similar application of mixing, but with significantly more complete control over the semantic parameters. To generate images with a target identity, the source parameters of the face rig are transferred to the target latent, resulting in images with different head poses, expressions and illumination. This rig-like control is not possible via the mixing strategy of Karras et al. (2019a) which entangles multiple semantic dimensions in the mixed results. Fig. 7.4 analyzes how the latent vectors of StyleGAN are transformed by StyleRig. The figure shows the average change and variance (the change is measured as ℓ_2 distance) of StyleGAN latent vectors at all resolutions, computed over 2500 mixing results. As expected, coarse latent code vectors are mainly responsible for rotation. Expression is controlled both by coarse and medium level latent codes. The light direction is mostly controlled by the medium resolution vectors. However, the fine latent vector also plays an important role in the control of the global color scheme of the images. Rather than having to specify which vectors need to change and by how much, StyleRig recovers this mapping in a self-supervised manner. Fig. 7.5 shows that the approach can also preserve scene context like background, hair styles and accessories better.

7.7.3 *Interactive Rig Control*

Since the parameters of the 3DMM can also be controlled independently, StyleRig allows for explicit semantic control of StyleGAN generated images. A user interface is developed where a user can interact with a face mesh by interactively changing its pose, expression, and scene illumination parameters. These updated parameters are then fed into RigNet to generate new images at interactive frame rates ($\sim$ 5 fps). Fig. 7.1 shows the results for various controls over StyleGAN images: pose, expression,

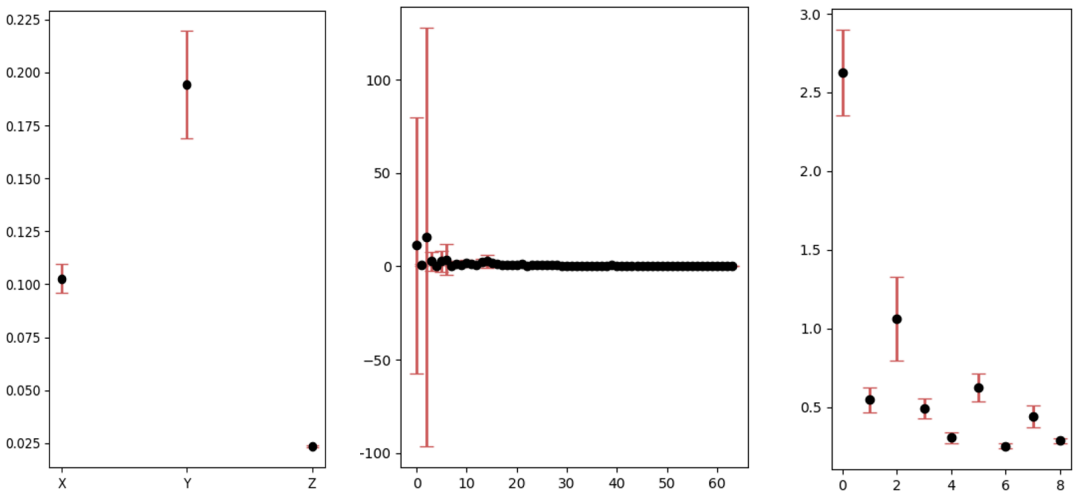

Figure 7.6: Distribution of face model parameters in the training data. x-axis shows the face model parameters for rotation, expression and illumination from left-right. y-axis shows the mean and variance of the parameters computed over $20k$ training samples.

and illumination edits. The control rig carries out the edits in a smooth interactive manner.

ANALYSIS OF STYLERIG The interactive editor allows for easy *inspection* of the trained networks. While the network does a good job at most controls, some expressivity of the 3D parametric face model is lost. That is, RigNet cannot transfer all modes of parametric control to similar changes in the StyleGAN generated images. For example, in-plane rotation of the face mesh is ignored. Similarly, many expressions of the face mesh do not translate well into the resultant generated images. These problems can be attributed to the bias in the images StyleGAN has been trained on. To analyze these modes, the distribution of face model parameters in the training data, generated from StyleGAN, is visualized in Fig. 7.6. In-plane rotations (rotation around the Z-axis) are not present in the data, due to the preprocessing pipeline of StyleGAN. In addition, most generated images consist of either neutral or smiling/laughing faces. These expressions can be captured using up to three blendshapes. Even though the face rig contains 64 vectors, the proposed method cannot control them well because of the biases in the distribution of the training data. Similarly, the lighting conditions are also limited in the dataset. There are larger variations in the global color and azimuth dimensions, as compared to the other dimensions. The proposed approach provides an intuitive and interactive user interface which allows us to inspect not only StyleRig, but also the biases present in StyleGAN.

7.7.4 *Conditional Image Generation*

Explicit and implicit control of a pretrained generative model allows us to turn it into a conditional one. The pose, expression, or illumination inputs to RigNet can be simply fixed in order to generate images which correspond to the specified parameters, see

Figure 7.7: Explicit control over the 3D parameters allows us to turn StyleGAN into a conditional generative model.

Fig. 7.7. This is a straightforward way to convert an unconditional generative model into a conditional model, and can produce high-resolution photorealistic results. It is also very efficient, as it takes less than 24 hours to train StyleRig, while training a conditional generative model from scratch should take at least as much time as StyleGAN, which takes more than 41 days to train (both numbers are for an Nvidia Volta GPU).

7.7.5 *Comparisons to Baseline Approaches*

In the following, StyleRig is compared with several baseline approaches.

"STEERING" THE LATENT VECTOR Inspired by Jahanian et al. (2019), a network architecture is designed which tries to *steer* the StyleGAN latent vector based on the change in parameters. This network architecture does not use the latent vector $\mathbf{w}$ as an input, and thus does not require an encoder. The inputs to the network are the delta in the face model parameters, with the output being the delta in the latent vector. In the settings of StyleRig, such an architecture does not lead to desirable results with the network not being able to deform the geometry of the faces, see Fig. 7.8. Thus, the semantic deltas in latent space should also be conditional on the the latent vectors, in addition to the target parameters.

DIFFERENT LOSS FUNCTIONS As explained in Eq. 4.7, the loss function consists of three terms. For the first baseline, the reconstruction loss is switched off. This can lead to the output latent vectors drifting from the space of StyleGAN latent codes, thus resulting in non-face images. Next, the consistency loss is switched off. This loss term enforces the consistency of all face model parameters, other than the one being changed. Without this term, changing one dimension, for example the illumination, also changes others such as the head pose. The final model ensures the desired edits with consistent identity and scene information. Note that switching off the editing loss is not a good baseline, as it would not add any control over the generator.

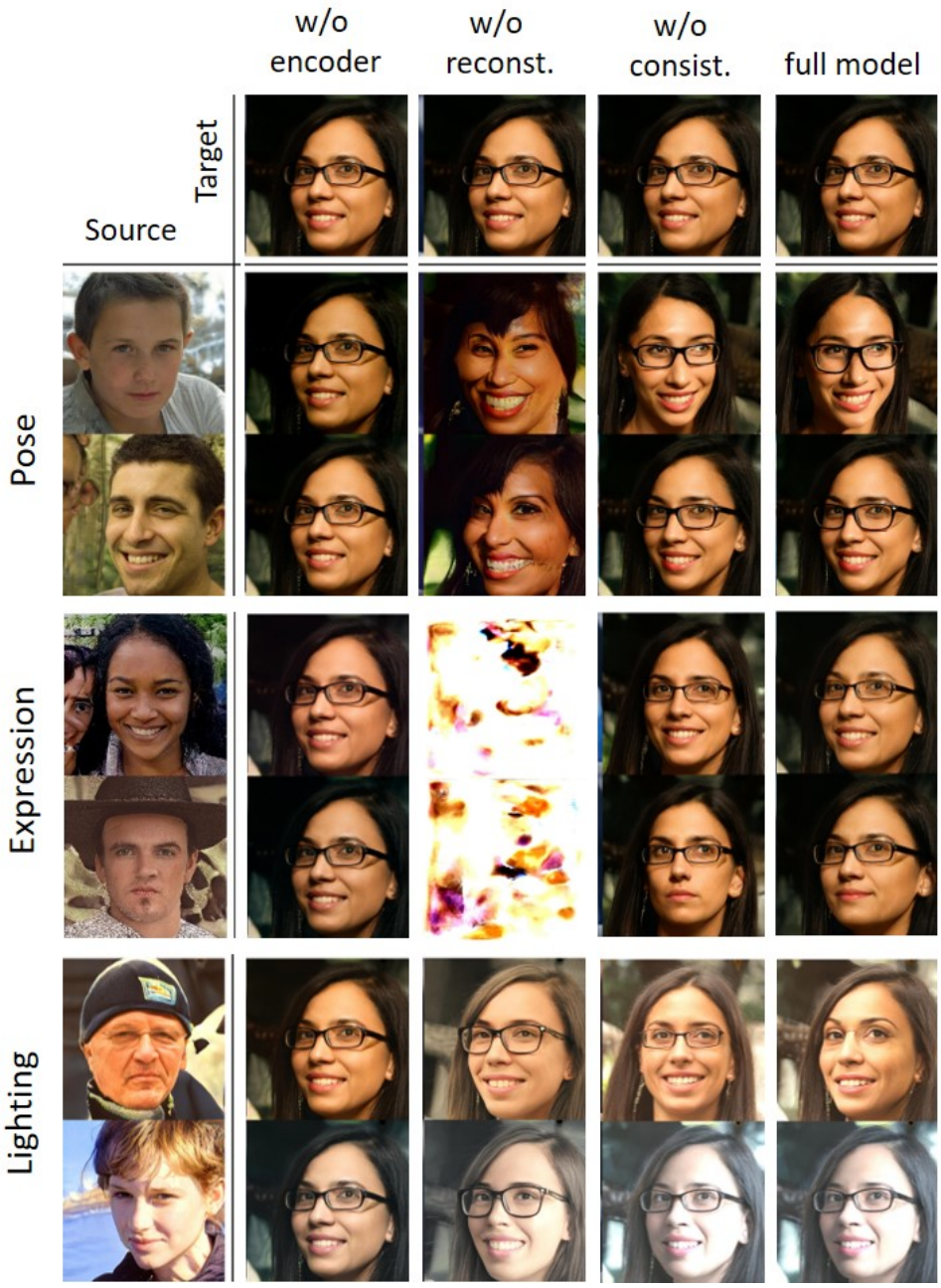

Figure 7.8: Baseline comparisons. The full approach obtains the highest quality results.

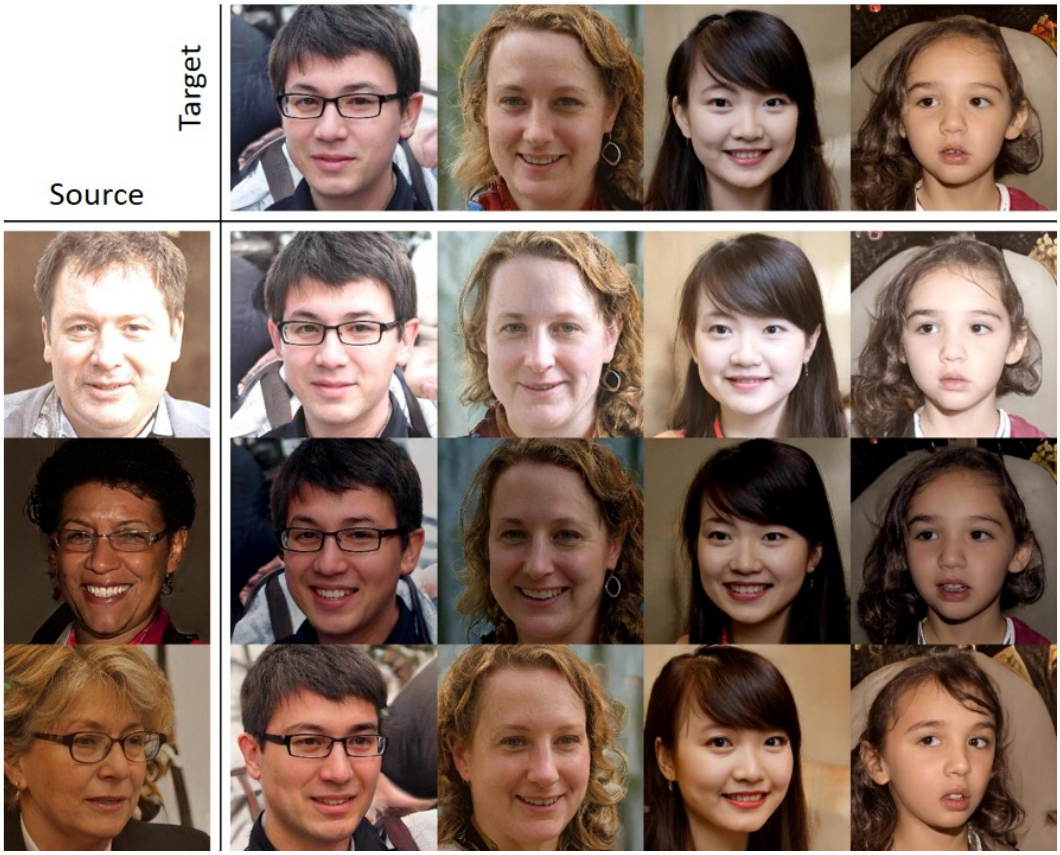

Figure 7.9: RigNet can also control pose, expression, and illumination parameters simultaneously. These parameters are transferred from source to target images, while the identity in the target images is preserved.

7.7.6 *Simultaneous Parameter Control*

In addition to controlling different parameters independently, they can also be controlled simultaneously. To this end, RigNet is trained such that it receives target pose, expression, and illumination parameters as input. For every $(\mathbf{w}, \mathbf{v})$ training code vector pair, three training samples are sampled. Here, one out of the three parameters (pose, expression or illumination) is changed in each sample. The loss function defined in Eq. 4.7 is used for each such sample. Thus, RigNet learns to edit each dimension of the control space independently, while also being able to combine the edits using the same network. Fig. 7.9 shows mixing results where pose, expression and illumination parameters are transferred from the source to target images.

7.8 LIMITATIONS

While this chapter demonstrated high-quality semantic control of StyleGAN-generated facial imagery, the approach is still subject to a few limitations that can be addressed in follow-up work. In the analysis sections, the chapter already discussed that StyleRig is not able to exploit the full expressivity of the parametric face model, see Fig. 7.10. This provides a nice insight into the inner workings of StlyeGAN and allows us to

Figure 7.10: Limitations: Transformations not present in the training data cannot be produced. Thus, the proposed method cannot handle in-plane rotation and asymmetrical expressions.

introspect the biases it learned. In the future, this might lead the ways to designing better generative models. The approach is also limited by the quality of the employed differentiable face reconstruction network. Currently, this model does not allow us to reconstruct fine-scale details, thus we cannot explicitly control them. Finally, there is no explicit constraint that tries to preserve parts of the scene that are not explained by the parameteric face model, for example, the background or hair style. Therefore, these parts cannot be controlled and might change when editing the parameters.

7.9 CONCLUSION

This chapter proposed StyleRig, a novel approach that provides a face rig-like control over a pretrained and fixed StyleGAN network. The network is trained in a self-supervised manner and does not require any additional images or manual annotations. At test time, the method generates images of faces with the photorealism of StyleGAN, while providing explicit control over a set of semantic control parameters. The combination of computer graphics control with deep generative models enables many exciting editing applications, and provides insights into the inner workings of the generative model.

While this method can only edit images synthesized by StyleGAN, an image editing application would require editing an existing image. This is not trivial with the proposed method as existing real images will have to be projected onto the latent space of StyleGAN. The next chapter will present a method which can faithfully project real images onto the StyleGAN latent space, ensuring good reconstruction and editing of the image.

8 PIE: PORTRAIT IMAGE EMBEDDING FOR SEMANTIC CONTROL

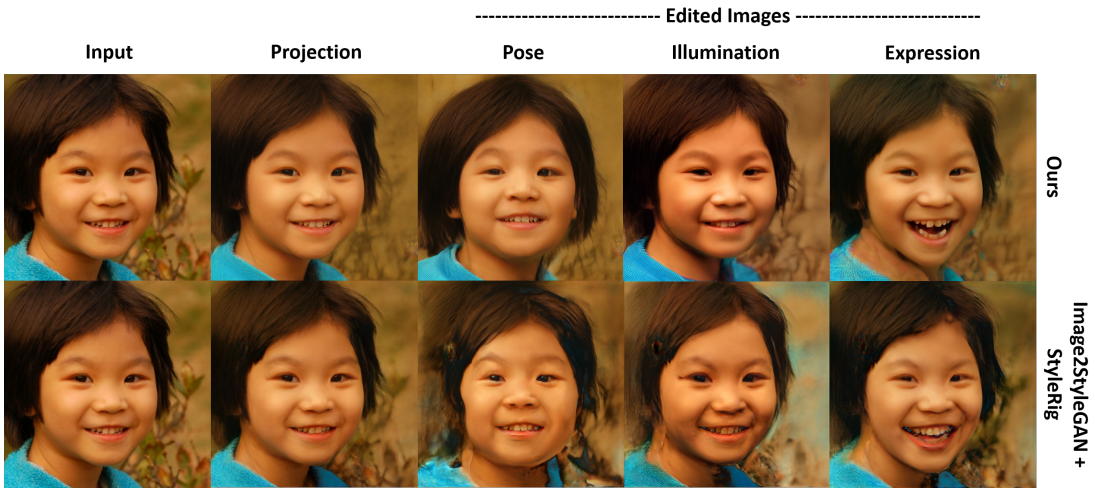

Figure 8.1: This chapter proposes an approach for embedding portrait images in the latent space of StyleGAN (Karras et al., 2019a) (visualized as "Projection") which allows for intuitive photo-real semantic editing of the head pose, facial expression, and scene illumination using StyleRig (Tewari et al., 2020b), presented in Chapter 7. Our optimization-based approach allows us to achieve higher quality editing results compared to the existing embedding method Image2StyleGAN (Abdal et al., 2019). Image from Shen et al. (2016).

Editing of portrait images is a very popular and important research topic with a large variety of applications. For ease of use, control should be provided via a semantically meaningful parameterization that is akin to computer animation controls. The vast majority of existing techniques do not provide such intuitive and fine-grained control, or only enable coarse editing of a single isolated control parameter. The method presented in the previous chapter allows for high-quality semantically controlled editing, however only on synthetically created StyleGAN images. This chapter, published as Tewari et al. (2020a), presents the first approach for embedding real portrait images in the latent space of StyleGAN, which allows for intuitive editing of the head pose, facial expression, and scene illumination in the image, see Fig. 8.1. Semantic editing in parameter space is achieved based on StyleRig, a pretrained neural network that maps the control space of a 3D morphable face model to the latent space of the GAN (Chapter 7). A novel hierarchical non-linear optimization problem is designed to obtain the embedding. An identity preservation energy term allows spatially coherent edits while maintaining facial integrity. The presented approach (PIE) runs at interactive frame rates and thus allows the user to explore the space of possible edits.

8.1 INTRODUCTION

Portrait images, showing mainly the face and upper body of people, are among the most common and important photographic depictions. We look at them to emotion-

ally connect with friends and family, we use them to best present ourselves in job applications and on social media, they remind us of memorable events with friends, and photographs of faces are omnipresent in advertising. Nowadays, tools to computationally edit and post-process photographs are widely available and heavily used. Professional and hobby photographers use them to bring out the best of portrait and social media photos, as well as of professional imagery used in advertising. Photos are often post-processed with the purpose to change the mood and lighting, to create a specific artistic look and feel, or to correct image defects or composition errors that only become apparent after the photo has been taken. Today, commercial software[1] or recent research software (Gatys et al., 2016; Luan et al., 2017) offers a variety of ways to edit the color or tonal characteristics of photos. Some tools even enable the change of visual style of photos to match certain color schemes (Luan et al., 2017; Shih et al., 2014), or to match a desired painterly and non-photo-realistic style (Gatys et al., 2016; Selim et al., 2016). In many cases, however, edits to a portrait are needed that require more complex and high-level modifications, for example, modifying head posture, smile or scene illumination after the capture. Enabling such edits from a single photograph is an extremely challenging and underconstrained problem. This is because editing methods need to compute reliable estimates of 3D geometry of the person and lighting in the scene. Moreover, they need to photo-realistically synthesize modified images of the person and background in a perspectively correct parallax-respecting manner, while inpainting disoccluding regions.

For ease of use, editing methods should use semantically meaningful parameterizations, which for the rest of the paper means the following: Head pose, face expression and scene lighting should be expressed as clearly disentangled and intuitive variables akin to computer animation controls, such as coordinates and angles, blendshape weights, or environment map parameterizations. Existing methods to edit human portrait imagery at best achieve parts of these goals. Some model-based methods to realistically edit human expression (Thies et al., 2016a, 2019) and head pose (Kim et al., 2018a) fundamentally require video input and do not work on single images. Other editing approaches are image-based and cannot be controlled by intuitive parametric controls (Averbuch-Elor et al., 2017; Geng et al., 2018; Siarohin et al., 2019; Wang et al., 2019b; Zakharov et al., 2019), only enable editing of a single semantic parameter dimension, for example, scene illumination (Meka et al., 2019; Sun et al., 2019; Zhou et al., 2019), or do not photo-realistically synthesize some important features such as hair (Nagano et al., 2018).

Recently, generative adversarial neural networks, such as StyleGAN (Karras et al., 2019a), were trained on community face image collections to learn a manifold of face images. They can be sampled to generate impressive photo-realistic face portraits, even of people not existing in reality. However, their learned parameterization entangles important face attributes (most notably identity, head pose, facial expression, and illumination), which thus cannot be independently and meaningfully controlled in the output. As a first step towards semantically meaningful editing, StyleRig (Tewari et al., 2020b), the method presented in Chapter 7, described a neural network that maps the parameters of a 3D morphable face model (3DMM) (Blanz and Vetter, 1999) to a pretrained StyleGAN for face images. However, while the results show disentangled

1 For example: www.adobe.com/Photoshop

control of face images synthesized by a GAN, they do not allow for editing real portrait photos.

On the other hand, some approaches have tried to embed real images in the StyleGAN latent space. Abdal et al. (2019, 2020a) demonstrate high-quality embedding results, which are used to perform edits such as style or expression transfer between two images, latent space interpolation for morphing, or image inpainting. However, when these embeddings are used to edit the input images using StyleRig (Tewari et al., 2020b), the visual quality is not preserved and the results often have artifacts. High-quality parametric control of expression, pose or illumination on real images has not yet been shown to be feasible.

This chapter therefore present the first method for embedding real portrait images in the StyleGAN latent space which allows for photo-realistic editing that combines all the following features: It enables photo-real semantic editing of all these properties — head pose, facial expression, and scene illumination, given only a single in-the-wild portrait photo as input, see Fig. 8.1. Edits are coherent in the entire scene and not limited to certain face areas. Edits maintain perspectively correct parallax, photo-real occlusions and disocclusions, and illumination on the entire person, without warping artifacts in the unmodeled scene parts, such as hair. The embedding is estimated based on a novel non-linear optimization problem formulation. Semantic editing in parameter space is then achieved based on the pretrained neural network of StyleRig, presented in the previous chapter, which maps the control space of a 3D morphable face model to the latent space of StyleGAN. These semantic edits are accessible through a simple user interface similar to established face animation control. This chapter makes the following contributions:

- This chapter proposes a hierarchical optimization approach that embeds a portrait image in the latent space of StyleGAN while ensuring high-fidelity as well as editability.
- Moreover, in addition to editability of the head pose, facial expression and scene illumination, this chapter proposes an energy function that enforces preservation of the facial identity.

8.2 RIGGING STYLEGAN-GENERATED IMAGES

StyleGAN (Karras et al., 2019a) can synthesize human faces at an unprecedented level of photorealism. However, their edits are defined in terms of three main facial levels (coarse, medium and fine), with no semantic meaning attached to them. As already covered in Chapter 7, StyleRig (Tewari et al., 2020b) attaches a semantic control for a StyleGAN embedding, allowing edits for head pose, illumination and expressions. The control is defined through a 3D Morphable Face Model (3DMM) (Blanz and Vetter, 1999). We recap StyleRig in the next section.

8.2.1 *StyleRig in more detail*

Faces are represented by a 3DMM model with $m = 257$ parameters

$$\theta = (\phi, \rho, \alpha, \delta, \beta, \gamma) \in \mathbb{R}^{257}. \tag{8.1}$$

Here, $(\phi, \rho) \in \mathbb{R}^6$ are the rotation and translation parameters of the head pose, where rotation is defined using Euler angles. The vector $\alpha \in \mathbb{R}^{80}$ represents the geometry of the facial identity, while $\beta \in \mathbb{R}^{64}$ are the expression parameters. Skin reflectance is defined by $\delta \in \mathbb{R}^{80}$ and the scene illumination by $\gamma \in \mathbb{R}^{27}$. The basis vectors of the geometry and reflectance models are learned from 200 facial 3D scans (Blanz and Vetter, 1999). The expression model is learned from FaceWarehouse (Cao et al., 2013) and the Digital Emily project (Alexander et al., 2009). Principal Components Analysis (PCA) is used to compress the original over-complete blendshapes to a subspace of 64 parameters. Faces are assumed to be Lambertian, where illumination is modeled using second-order spherical harmonics (SH) (Ramamoorthi and Hanrahan, 2001b).

StyleRig (Tewari et al., 2020b) allows one to semantically edit synthetic StyleGAN images. To this end, StyleRig trains a neural network, called *RigNet*, which can be understood as a function $\text{rignet}(\cdot, \cdot)$ that maps a pair of StyleGAN code $\mathbf{v}$ and subset of 3DMM parameters θ^τ to a new StyleGAN code $\hat{\mathbf{v}}$, i.e. $\hat{\mathbf{v}} = \text{rignet}(\mathbf{v}, \theta^\tau)$. In practice, the 3DMM parameters are first transformed before being used in the network. With that, $\mathbf{I}_{\hat{\mathbf{v}}}$ shows the face of $\mathbf{I}_\mathbf{v}$ modified according to θ^τ (i.e. with edited head pose, scene lighting, or facial expression), where $\mathbf{I}_\mathbf{v}$ is the StyleGAN image generated using the latent code $\mathbf{v}$. Thus, editing a synthetic image $\mathbf{I}_\mathbf{v}$ amounts to modifying the component τ in the parameter θ, and then obtaining the edited image as $\mathbf{I}_{\hat{\mathbf{v}}} = \mathbf{I}(\text{rignet}(\mathbf{v}, \theta^\tau))$. Multiple RigNet models are trained, each to deal with just one mode of control (pose, expression, lighting). Although RigNet allows for editing of facial images, it has the major shortcoming that only *synthetic* images can be manipulated, rather than real images. This is in contrast to the method presented in this chapter, where semantic editing of *real* images can be performed. Different from the original RigNet design where a differentiable face reconstruction network regresses the 3DMM parameters from a StyleGAN code, we use a model-based face autoencoder (Tewari et al., 2017) which takes an image as an input. This change is necessary, as we initially do not have the StyleGAN code for the real image we want to edit.

8.3 SEMANTIC EDITING OF REAL IMAGES

The key of the approach for semantic editing of real images is to embed the given image in the StyleGAN latent space (Karras et al., 2019a), where we pay particular attention to finding a latent encoding that is *suitable for editing the image*. This is crucial, since the parameter space of the StyleGAN architecture is generally under-constrained. For example, it has been shown that a StyleGAN trained for human faces is able to synthesize images that show completely different content with high fidelity, such as images of cat faces (Abdal et al., 2019) The goal is to compute embeddings which can be edited using 3DMM parameters using StyleRig.

PROBLEM STATEMENT The image that we want to make editable will be referred to as as $\mathbf{I}$ (without any subscripts or arguments), which we assume to be a given input. Moreover, the StyleGAN code that will make image $\mathbf{I}$ editable will be referred to as $\mathbf{w}$, which is the desired output of our approach. As such, an energy function $E(\mathbf{w})$ will be introduced, which is minimized by solving a numerical optimization problem. This energy function accounts for the high fidelity of the synthesized image based on $\mathbf{w}$ (explained in Sec. 8.3.1), for editing-suitability (described in Sec. 8.3.2), as well

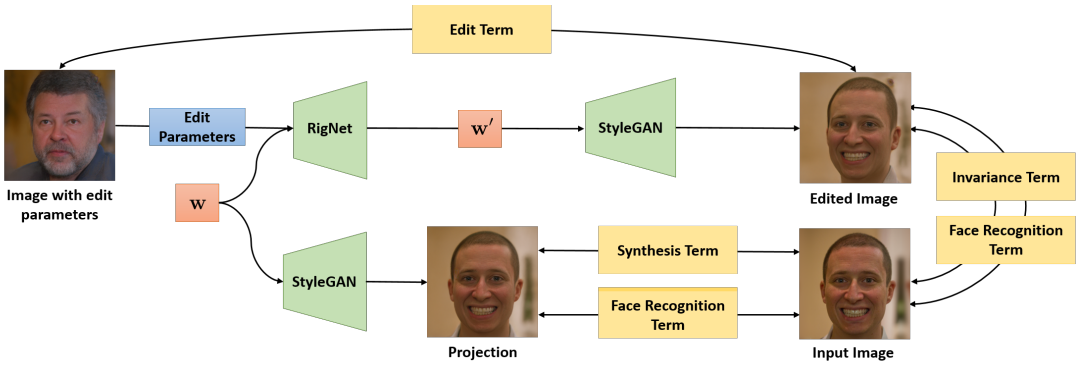

Figure 8.2: Given a portrait input image, a StyleGAN embedding is optimized for which allows to faithfully reproduce the image (synthesis and facial recognition terms), editing the image based on semantic parameters such as head pose, expressions and scene illumination (edit and invariance terms), as well as preserving the facial identity during editing (facial recognition term). A novel hierarchical non-linear optimization strategy is used to compute the result. StyleGAN generated images (image with edit parameters) are used to extract the edit parameters during optimization. At "test time", i.e. for performing portrait image editing, the image with edit parameters is not needed. Note that the identity term is not visualized here. Images from Shih et al. (2014).

as for consistent face identity before and after the edit (Sec. 8.3.3). The approach is based on non-linear optimization techniques, and does not perform any learning of network weights, which in turn means that any ground truth data of edited facial images is not required. Several existing neural networks are used to define the energy term, where all networks are pretrained and remain fixed throughout the optimization. Some technical notations will be introduced now, which will allow for an additional layer of abstraction and thereby facilitate a more comprehensive description of the main concepts.

NOTATION Throughout this paper $\mathbf{w}$ will be exclusively used to refer to the (unknown) desired StyleGAN embedding, and $\mathbf{v}$ (potentially with subscripts) will be used to refer to general StyleGAN embeddings. Note that the StyleGAN embeddings $\mathbf{w}$ and $\mathbf{v}$ can have two different forms, where each form has a different dimensionality, which we will describe in detail in Sec. 8.3.4. StyleGAN can be understood as a function $\text{stylegan}(\cdot)$ that maps a given latent code to a portrait image. To simplify notation, function notation $\mathbf{I}(\mathbf{v}) := \text{stylegan}(\mathbf{v})$ is used in order to emphasize that the StyleGAN embedding $\mathbf{v}$ is used to generate the image $\mathbf{I}(\mathbf{v})$. Analogously, $\mathbf{I}(\cdot)$ is overloaded, so that it can also take a 3DMM parameter θ as input. As such, $\mathbf{I}(\theta)$ refers to an image rendered using the face model that is parameterized by θ (Sec. 8.2.1), where differentiable rendering is employed (Tewari et al., 2017). Note that this rendered image is only defined on foreground face pixels as opposed to StyleGAN images.

The variable $\tau \in \{\phi, , \gamma\}$ is used to indicate the user-defined facial semantic variable that is to be edited, which in our case can be the head pose ϕ, facial expression β, or illumination γ. Similarly, the complement notation $\overline{\tau} \subset \{\phi, \rho, \alpha, \delta, \beta, \gamma\}$ is used to indicate all other variables, i.e., the ones that shall not be modified. With that, the notation θ^{τ} (or $\theta^{\overline{\tau}}$) is used to refer to the extraction of the τ-component (or $\overline{\tau}$-components) of θ. Since facial editing is implemented by modifying the τ-component of the 3DMM parameter θ, $\theta' = [\theta_1^{\overline{\tau}}, \theta_2^{\tau}]$ is used to indicate that the respective τ-

Table 8.1: Summary of notation.

Symbol	Meaning
$\mathbf{w}$	StyleGAN embedding that we want to find
$\mathbf{v}$	other StyleGAN embedding(s)
θ	3DMM parameter
τ	component that is to be edited ($\tau \in \{\phi, \beta, \gamma\}$)
$\mathbf{I}$	input image that we want to edit
$\mathbf{I}(\mathbf{v})$	StyleGAN-synthesized image
$\mathbf{I}(\theta)$	image of 3DMM rendering
θ^τ	extraction of τ-component of θ
$[\theta_1^{\overline{\tau}}, \theta_2^\tau]$	combine $\overline{\tau}$-components in θ_1 with τ-component in θ_2
$\theta(\mathbf{v}), \theta_\mathbf{v}$	3D reconstruction of 3DMM parameters from $\mathbf{I}(\mathbf{v})$
$\theta(\mathbf{I}'), \theta_{\mathbf{I}'}$	3D reconstruction of 3DMM parameters from $\mathbf{I}'$

component of θ_1 is replaced by the corresponding component in θ_2. For example, for $\tau = \beta$,

$$\theta_1 = (\phi_1, \rho_1, \alpha_1, \delta_{1,1}, \gamma_1)\,, \quad \text{and} \tag{8.2}$$

$$\theta_2 = (\phi_2, \rho_2, \alpha_2, \delta_{2,2}, \gamma_2)\,, \quad \text{we have} \tag{8.3}$$

$$[\theta_1^{\overline{\tau}}, \theta_2^\tau] = (\phi_1, \rho_1, \alpha_1, \delta_{1,2}, \gamma_1)\,. \tag{8.4}$$

Moreover, the notation $\theta(\mathbf{v})$ is used to extract the 3DMM parameters from the StyleGAN embedding $\mathbf{v}$. In order to compute this, the embedding $\mathbf{v}$ is first used to synthesize the image $\mathbf{I}(\mathbf{v})$ (using StyleGAN), followed by performing a 3D reconstruction based on the pretrained *Model-based Face Autoencoder* (MoFA) network (Tewari et al., 2017), presented in Chapter 4. Hence, for MoFA$(\cdot)$ being the function that performs 3D reconstruction for a given image by estimating the 3DMM parameters, we define

$$\theta(\mathbf{v}) = \text{MoFA}(\mathbf{I}(\mathbf{v}))\,. \tag{8.5}$$

For any image $\mathbf{I}'$, the short-hand notation $\theta(\mathbf{I}') = \text{MoFA}(\mathbf{I}')$ is used. Similarly as above, $\theta^\tau(\mathbf{v})$ and $\theta^\tau(\mathbf{I}')$ are used to extract only the τ-component from the 3DMM parameters. Whenever arguments of $\theta(\cdot)$ or $\mathbf{I}(\cdot)$ are fixed, i.e., the arguments are not a variable, the short-hand notations $\theta_\mathbf{v} = \theta(\mathbf{v})$, $\theta_{\mathbf{I}'} = \theta(\mathbf{I}')$, or $\mathbf{I}_\mathbf{v} = \mathbf{I}(\mathbf{v})$ are used. The most important parts of the notations are summarized in Table 8.1.

OBJECTIVE FUNCTION The optimization problem solves for $\mathbf{w}$ by minimizing the energy function

$$E(\mathbf{w}) = E_\text{syn}(\mathbf{w}) + E_\text{id}(\mathbf{w}) + E_\text{edit}(\mathbf{w}) + E_\text{inv}(\mathbf{w}) + E_\text{recog}(\mathbf{w})\,. \tag{8.6}$$

E_syn is a synthesis term enforcing the StyleGAN-synthesized image $\mathbf{I}(\mathbf{w})$ to be close to $\mathbf{I}$ (Sec. 8.3.1). E_id, E_edit, E_inv are face modification terms (Sec. 8.3.2) enforcing edits to take place on the modified facial semantics while at the same time ensuring unmodified facial semantics to remain un-edited. $E_\text{recog}(\mathbf{w})$ is a face recognition term that will be introduced in Sec. 8.3.3. A conceptual illustration of the energy function and the overall pipeline is shown in Fig. 8.2. Next, we will discuss each term in more detail.

8.3.1 *High-Fidelity Image Synthesis*

Similarly to Image2StyleGAN (Abdal et al., 2019), the following energy term is used that accounts for the StyleGAN-synthesized image $\mathbf{I}(\mathbf{w})$ being close to $\mathbf{I}$:

$$E_{\text{syn}}(\mathbf{w}) = \lambda_{\ell_2}\|\mathbf{I} - \mathbf{I}(\mathbf{w})\|_2^2 + \lambda_{\text{p}}\|\mathbf{\Phi}(\mathbf{I}) - \mathbf{\Phi}(\mathbf{I}(\mathbf{w})\|_2^2 \,. \tag{8.7}$$

The first term in the energy E_{syn} penalizes the discrepancy between $\mathbf{I}$ and the synthesized image in terms of the (squared) ℓ_2-norm, whereas the second term penalizes discrepancies based on the *perceptual loss* (Johnson et al., 2016). The perceptual loss is estimated on images downsampled by a factor of 4, based on ℓ_2-losses over VGG-16 layers `conv1_1`, `conv1_2`, `conv3_2` and `conv4_2` (Simonyan and Zisserman, 2015). The notation $\mathbf{\Phi}(\cdot)$ refers to the function that downsamples a given input image and extracts features. The scalars λ_{ℓ_2} and λ_{p} are the relative weights of both terms.

In principle, the energy E_{syn} in (8.7) could be minimized in order to obtain the StyleGAN code $\mathbf{w}$, as done in Abdal et al. (2019), and editing operations could be performed on $\mathbf{w}$. A so-obtained code vector $\mathbf{w}$ allows the use of StyleGAN to obtain a highly accurate synthetic version of the input face, which is even capable of reconstructing backgrounds with high accuracy. However, such a $\mathbf{w}$ is sub-optimal for performing *semantic face editing*, as we later demonstrate in Fig. 8.6.

8.3.2 *Face Image Editing*

The synthesis term is augmented with an editing energy that is based on the StyleRig framework (Tewari et al., 2020b), which allows for obtaining more accurate semantic editing while preserving the non-edited attributes. Here, the StyleGAN embedding $\mathbf{w}$ that is to be determined should have the following three properties in order to be suitable for semantic editing:

IDENTITY PROPERTY The identity property is phrased in terms of the ℓ_2-norm of the difference of StyleGAN embeddings and is given by

$$E_{\text{id}}(\mathbf{w}) = \lambda_{\text{id}}\|\mathbf{w} - \text{rignet}(\mathbf{w}, \theta^{\tau}(\mathbf{w}))\|_2^2 \,. \tag{8.8}$$

As such, whenever the RigNet is used to modify $\mathbf{w}$ with $\theta^{\tau}(\mathbf{w})$, i.e., a component of the 3DMM parameter extracted from $\mathbf{w}$, the embedding $\mathbf{w}$ should not be modified.

EDIT PROPERTY In order to get around the obstacle of defining a suitable metric for 3DMM parameter vectors, whose components may be of significantly different scale, and the relative relevance of the individual components is not easily determined, we phrase the edit property in image space, as in StyleRig (Tewari et al., 2020b). As such, a facial edit is implicitly specified in image space via the StyleGAN embedding $\mathbf{v}$, where the τ-component of the respective 3DMM parameters of $\mathbf{v}$, i.e. $\theta^{\tau}_{\mathbf{v}}$, specifies the edit operation. The image-space version of the edit property reads

$$\forall\, \mathbf{v}: \quad \mathbf{I}_{\mathbf{v}} = \mathbf{I}([\theta^{\overline{\tau}}_{\mathbf{v}}, \theta^{\tau}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))]) \,. \tag{8.9}$$

Note that this true equality cannot hold in practice, since the two images are from different domains (real image and a mesh rendering). We are interested in minimimzing the difference between these terms. This equation is best fulfilled whenever

the τ-component of the edited 3DMM parameters $\theta^\tau(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v}))$ is equal to $\theta^\tau_\mathbf{v}$, i.e. the edit has been successfully applied. Since computationally we cannot evaluate all choices of $\mathbf{v}$, we sample StyleGAN embeddings $\mathbf{v}$ as done in Chapter 7, and then use the expected value as loss. For integrating this property into the optimization framework, a combination of a photometric term and a landmark term is used, which is defined as

$$\ell(\mathbf{I}', \theta) = \lambda_\text{ph} \|\mathbf{I}' - \mathbf{I}(\theta)\|^2_\text{☺} + \lambda_\text{lm} \|\mathcal{L}_{\mathbf{I}'} - \mathcal{L}(\theta)\|^2_F \,. \tag{8.10}$$

The norm $\|\cdot\|_\text{☺}$ computes the ℓ_2-norm of all *foreground* pixels (the facial part of the image), whereas $\|\cdot\|_F$ is the Frobenius norm. The matrix of 2D facial landmarks (based on Saragih et al. (2011)) extracted from the image $\mathbf{I}$ is denoted by $\mathcal{L}_{\mathbf{I}'} \in \mathbb{R}^{66\times 2}$, and $\mathcal{L}(\theta) \in \mathbb{R}^{66\times 2}$ refers to the corresponding landmarks of the 3DMM after they have been projected onto the image plane. With that, the edit property energy reads

$$E_\text{edit}(\mathbf{w}) = \lambda_\text{e}\, \mathbb{E}_\mathbf{v}[\ell(\mathbf{I}_\mathbf{v}, [\theta^{\overline{\tau}}_\mathbf{v}, \theta^\tau(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v}))])] \,. \tag{8.11}$$

INVARIANCE PROPERTY Similarly as the edit property, the invariance property is also phrased in image space as

$$\forall\ \mathbf{v}: \quad \mathbf{I} = \mathbf{I}([\theta^{\overline{\tau}}(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v})), \theta^\tau_\mathbf{I}]) \,. \tag{8.12}$$

While the edit property imposes that the τ-component of the edited 3DMM parameter $\theta^\tau(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v}))$ is modified as desired, the invariance property takes care of all $\overline{\tau}$. It is fulfilled whenever it holds that $\theta^{\overline{\tau}}(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v})) = \theta^{\overline{\tau}}_\mathbf{I}$, i.e. the components $\overline{\tau}$ that are not to be edited are maintained from the input image $\mathbf{I}$.

Analogously to the edit property, the respective energy is based on the combination of a photometric term and a landmark term as implemented by $\ell(\cdot)$, so that

$$E_\text{inv}(\mathbf{w}) = \lambda_\text{inv}\, \mathbb{E}_\mathbf{v}[\ell(\mathbf{I}, [\theta^{\overline{\tau}}(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v})), \theta^\tau_\mathbf{I}])] \,. \tag{8.13}$$

8.3.3 *Face Recognition Consistency*

In addition to the synthesis and editing terms, two face recognition consistency terms are incorporated to preserve the facial integrity while editing. On the one hand, it is desirable that the synthesized image $\mathbf{I}(\mathbf{w})$ is recognized to depict the same person as shown in the given input image $\mathbf{I}$. On the other hand, the edited image, $\text{stylegan}(\text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v}))$ should also depict the same person as shown in the input $\mathbf{I}$.

In order to do so, VGG-Face (Parkhi et al., 2015) is used to extract *face recognition features*, where the notation $\mathbf{\Psi}(\cdot)$ is used to refer to the function that extracts such features from a given input image. The recognition loss is defined as

$$\ell_\text{recog}(\mathbf{I}', \mathbf{v}) = \|\mathbf{\Psi}(\mathbf{I}') - \mathbf{\Psi}(\mathbf{I}(\mathbf{v}))\|^2_F \,, \tag{8.14}$$

which is then used to phrase the recognition energy term as

$$E_\text{recog}(\mathbf{w}) = \lambda_{\text{r}_\mathbf{w}}\, \ell_\text{recog}(\mathbf{I}, \mathbf{w}) + \lambda_{\text{r}_{\hat{\mathbf{w}}}}\, \mathbb{E}_\mathbf{v}[\ell_\text{recog}(\mathbf{I}, \text{rignet}(\mathbf{w}, \theta^\tau_\mathbf{v}))] \,. \tag{8.15}$$

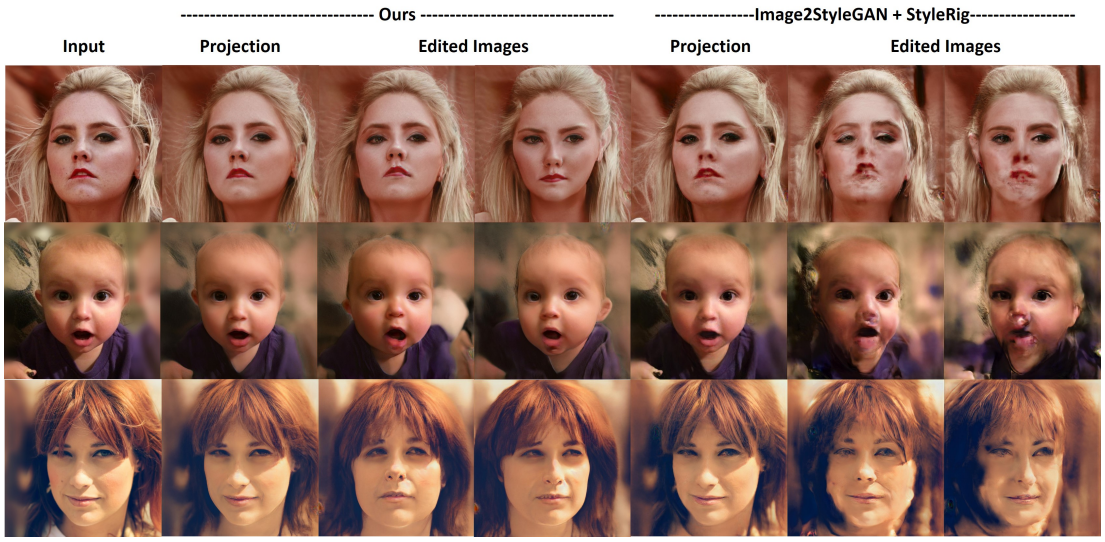

Figure 8.3: Pose Editing. The proposed approach can handle a large variety of head pose modifications including out-of-plane rotations in a realistic manner. Image2StyleGAN (Abdal et al., 2019) embeddings often lead to artifacts when edited using StyleRig. Images from Shen et al. (2016).

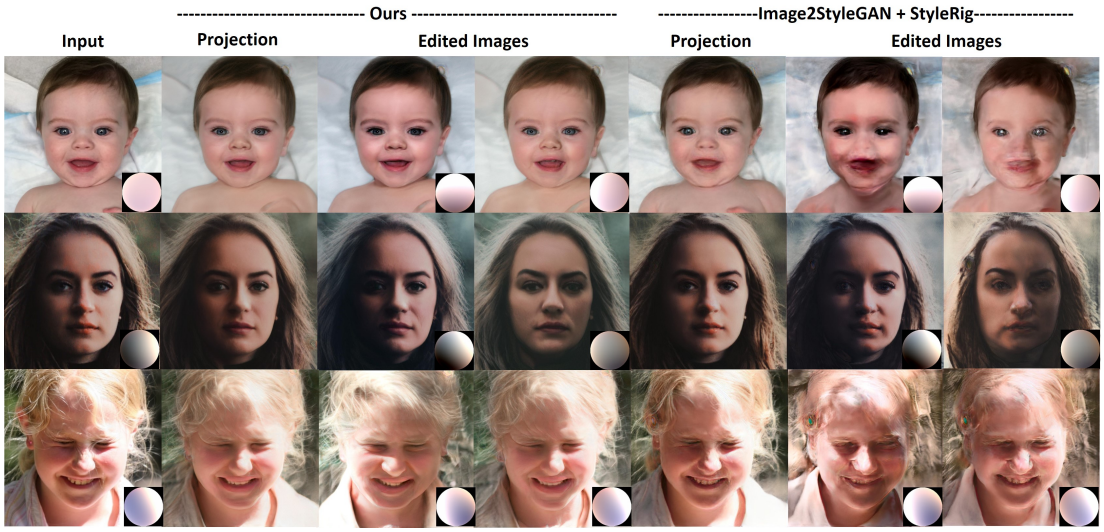

Figure 8.4: Illumination Editing. The proposed approach can realistically relight portrait images. Each edited image corresponds to changing a different Spherical Harmonics coefficient, while all other coefficients are kept fixed. The environment maps are visualized in the inset. Image2StyleGAN (Abdal et al., 2019) embeddings often lead to artifacts when edited using StyleRig. Images from Shen et al. (2016).

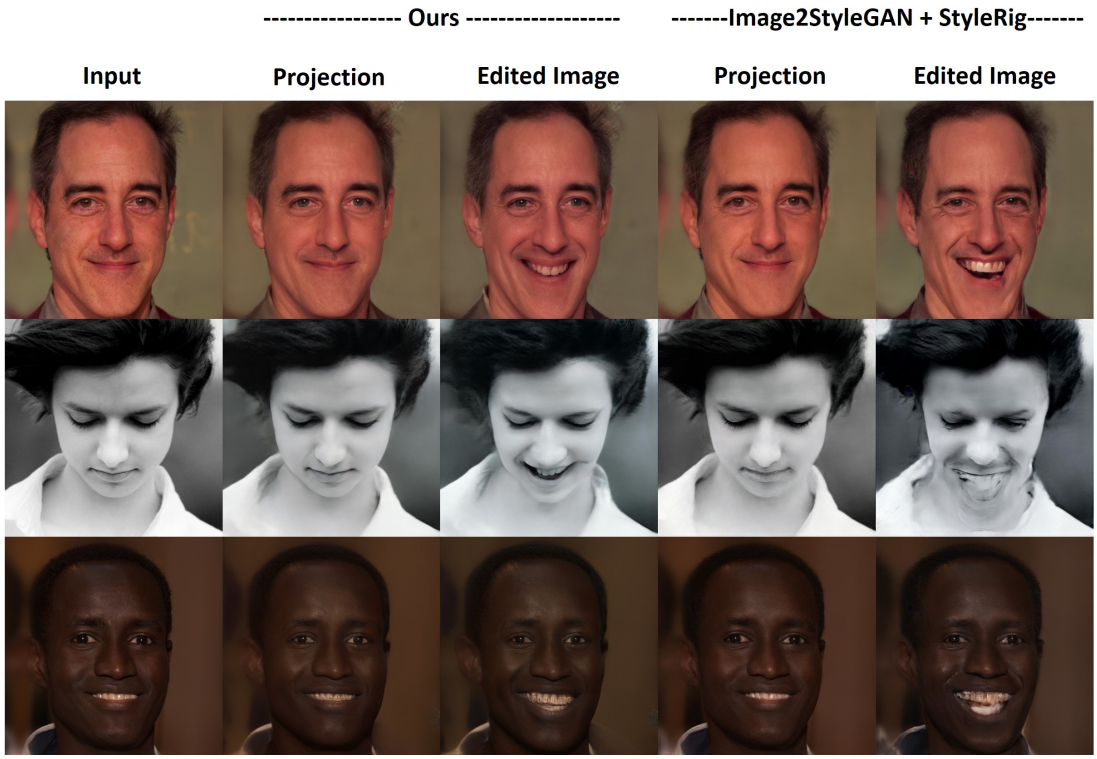

Figure 8.5: Expression Editing. The proposed approach can also be used to edit the facial expressions in a portrait image in a realistic manner. We obtain more plausible results, compared to Image2StyleGAN (Abdal et al., 2019) embeddings. Images from Shen et al. (2016) and Shih et al. (2014).

8.3.4 *Optimization*

The energy function $E(\cdot)$ in (8.6) depends on a range of highly non-linear functions, such as stylegan$(\cdot)$, MoFA$(\cdot)$, $\Phi(\cdot)$ and $\Psi(\cdot)$, which are implemented in terms of (pretrained) neural networks. The energy minimization is implemented within TensorFlow (Abadi et al., 2015) using ADADELTA optimization (Zeiler, 2012). In each iteration a different $\mathbf{v}$ is stochastically sampled. The optimization uses a hierarchical approach described next.

HIERARCHICAL OPTIMIZATION StyleGAN is based on a hierarchy of latent spaces, where a stage-one embedding Z with $|Z| = 512$ is randomly sampled first. This is then fed into a mapping network that produces W as output, where $|W| = 512$. Subsequently, W is extended to W^+, where $|W^+| = 18 \times 512$, and used as input to 18 network layers. It has been shown that W^+ is the most expressive space for fitting to real images (Abdal et al., 2019). However, a direct optimization over this space leads to lower-quality editing results with severe artifacts. This is because the optimized variable can be far from the prior distribution of StyleGAN. To address this, the proposed approach first optimizes for the embedding in the W-space, meaning that in the first stage of the optimization the variable $\mathbf{w}$ is understood as an embedding in the W-space. Optimization in W-space is run for 2000 iterations. The result is then transferred to W^+-space and the variable $\mathbf{w}$ is initialized respectively. The optimization is continued in the W^+-space for another 1000 iterations. Optimizing in this hierarchical way allows for representing the coarse version of the image in the W-space, which is less expressive and thereby closer to the prior distribution. Finetuning on the W^+ space then allows for fittting the fine-scale details, while preserving editing quality.

8.4 RESULTS

In the following, high-quality results of the method are demonstrated, its different components are analyzed, and the method is compared to several state-of-the-art approaches for portrait image editing. The proposed approach will be referred to as PIE, an abbreviation of Portrait Image Embedding.

IMPLEMENTATION DETAILS The following empirically determined weights are used for the energy terms: $\lambda_{\ell_2} = 10^{-6}$, $\lambda_{\mathrm{p}} = 10^{-6}$, $\lambda_{\mathrm{id}} = 1.0$, $\lambda_{\mathrm{ph}} = 0.001$, $\lambda_{\mathrm{lm}} = 0.2$, $\lambda_{\mathrm{e}} = 10.0$, $\lambda_{\mathrm{inv}} = 10.0$, $\lambda_{\mathrm{r_w}} = 0.1$, $\lambda_{\mathrm{r_{\hat{w}}}} = 0.1$. A starting step size of 50 is used when optimizing over embeddings in W space, and 10 in W^+ space. The step size is then exponentially decayed by a factor of 0.1 every 2000 steps. Optimization takes approximately 10 minutes for 3000 iterations per image on an NVIDIA V100 GPU. Once the embedding is obtained, the portrait image can be edited at an interactive speed.

FEEDBACK A simple feedback loop allows for more accurate editing results. The parameters used as input to RigNet are updated in order to correct for the editing inaccuracies in the output. Given target 3DMM parameters θ, the embedding for the edited image, rignet$(\mathbf{w}, \theta^\tau)$ are obtained. The 3DMM parameters from the edited em-

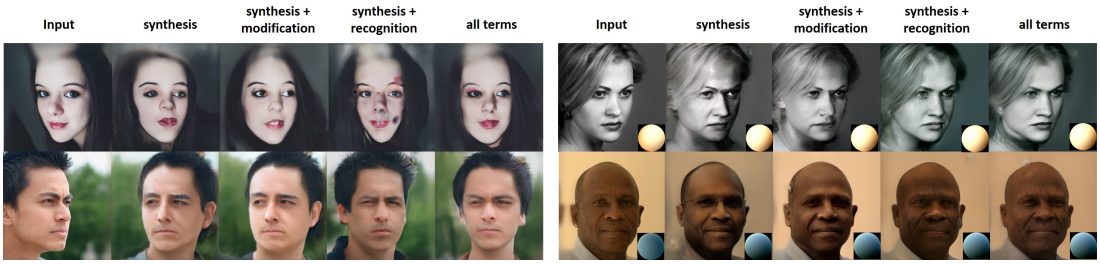

Figure 8.6: Ablative analysis of the different loss functions. *Modification* refers to the edit, invariance and identity terms simultaneously. The left block shows results for editing the head pose and the right block shows results for editing scene illumination. All losses are required to obtain high-fidelity edits. Images from Shen et al. (2016).

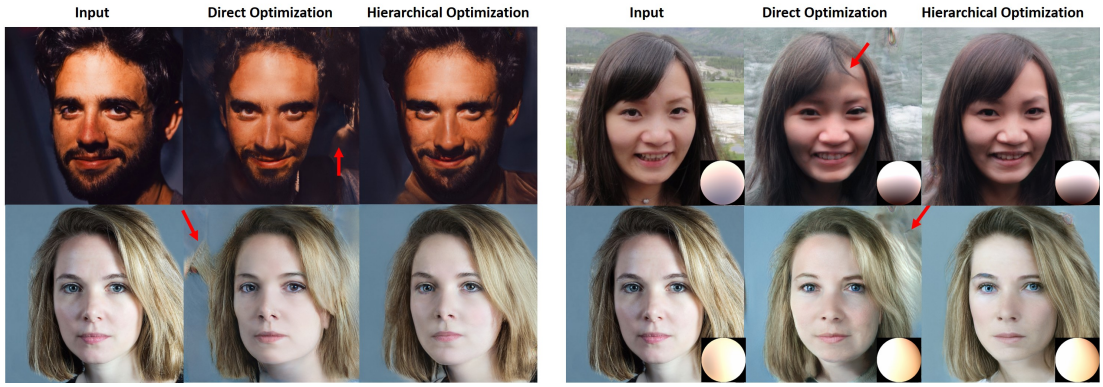

Figure 8.7: Ablative analysis with and without hierarchical optimization. The left block shows the results for pose editing and the right block for illumination editing. Without the hierarchical optimization, the obtained embedding cannot be easily edited and artifacts appear in the modified images. Images from Shen et al. (2016).

bedding, $\theta_{\text{est}} = \theta(\text{rignet}(\mathbf{w}, \theta^{\tau}))$ are computed. The final embedding is then computed as $\text{rignet}(\mathbf{w}, \theta^{\tau}_{\text{new}})$ with $\theta_{\text{new}} = \theta + (\theta - \theta_{\text{est}})$.

8.4.1 *High-Fidelity Semantic Editing*

The approach is evaluated on a large variety of portrait images taken from Shen et al. (2016) and Shih et al. (2014). The images are preprocessed as in StyleGAN (Karras et al., 2019a). Figs. 8.3, 8.4, 8.5 show results of controlling the head pose, scene illumination, and facial expressions, respectively. The projections onto the StyleGAN space are detailed, preserving the facial identity. The approach also produces photo-realistic edits. Fig. 8.3 shows that the approach can handle a large variety of head pose modifications, including out-of-plane rotations. It also automatically inpaints uncovered background regions in a photo-realistic manner. Fig. 8.4 demonstrates the relighting results. The approach can handle complex light material interactions, resulting in high photo-realism. The relighting effects are not restricted to just the face region, with hair and even eyes being relit. The approach also allows for editing facial expressions, see Fig. 8.5.

Table 8.2: Different settings are quantitatively compared using several metrics for pose editing. All numbers are averaged over more than 2500 pose editing results. The quality of the fit is measured by comparing them to the input image using PSNR and SSIM metrics. Editing error is measured as the angular difference between the desired and achieved face poses. Recognition error measures the value of the facial recognition error for the edited images. There is usually a trade-off between the quality and accuracy of editing, as lower recognition errors correspond to higher editing errors. We also compare to Image2StyleGAN (Abdal et al., 2019) embeddings using these metrics. While it achieves the highest quality fitting, the editing results do not preserve the facial identity well.

	synthesis	synthesis + recognition	synthesis + modification	all terms (PIE)	all terms (direct opt.)	Image2StyleGAN
PSNR (dB) ↑ / SSIM ↑	30.15 / 0.70	29.84 / 0.69	30.15 / 0.70	29.96 / 0.70	29.76 / 0.69	**31.21** / **0.75**
Editing Error (rad) ↓	0.06	0.11	**0.036**	0.08	0.037	0.07
Recognition Error ↓	95.76	43.64	90.10	**42.82**	51.65	275.40

8.4.2 *Ablation Studies*

Here, we evaluate the importance of the different proposed loss functions, and also evaluate the hierarchical optimization strategy.

LOSS FUNCTIONS Fig. 8.6 shows qualitative ablative analysis for the different loss functions. We group the edit, invariance and identity terms as *modification terms*. Adding face recognition consistency without the modification terms lead to incorrect editing in some cases. Adding the modification terms without face recognition consistency leads to the method being able to accurately change the specified semantic property, but the identity of the person in the image is not preserved. Using all terms together leads to results with photorealistic edits with preservation of identity. We do not evaluate the importance of the individual components of the modification terms, as it was already evaluated in Tewari et al. (2020b).

HIERARCHICAL OPTIMIZATION Hierarchical optimization is an important component of PIE. Fig. 8.7 shows results with and without this component. Without hierarchical optimization, the method directly optimizes for $\mathbf{w} \in W^+$. While this leads to high-quality fits, the obtained embedding can be far from the training distribution of StyleRig. Thus, the quality of edits is poor. For example in Fig. 8.7 (top), the StyleGAN network interprets the ears as background, which leads to undesirable distortions. With hierarchical optimization, the results do not suffer from artifacts.

QUANTITATIVE ANALYSIS We also analyze the effect of different design choices quantitatively, see Tab. 8.2. We look at three properties, the quality of recostruction (measured using PSNR and SSIM between the projected image and the input), the accuracy of edits (measured as the angular distance between the desired and estimated head poses), and idenity preservation under edits (measured using the second term in Eq. 8.15) during editing. The numbers reported are averaged over more than 2500 pose editing results. We can see that removing the recognition term changes the identity of the face during editing, and removing the modification loss increases the editing and recognition error. Hierarchical optimization also leads to better facial identity preservation, compared to direct optimization. This is expected, since the results with

Figure 8.8: Comparison of head pose editing for self-reenactment (first two rows) and cross-identity reenactment (last two rows). We compare the approach to Wiles et al. (2018), Wang et al. (2019c), Siarohin et al. (2019) and Geng et al. (2018). The pose from the reference images is transferred to the input. The approach obtains higher quality head pose editing results, specially in the case of cross-identity transfer. All approaches other than ours are incapable of *disentangled* edits, i.e., they cannot transfer the pose without also changing the expressions. The implementation of Geng et al. (2018) does not handle cross-identity reenactment. Note that while the three competing approaches require a reference image in order to generate the results, we allow for explicit control over the pose parameters. Image from Shen et al. (2016).

direct optimization often have artifacts. Note that the artifacts outside of the face region (hair, ears) would not increase the recognition errors significantly. The recognition term introduces a clear trade-off between the quality of identity preservation under edits and the accuracy of edits. The modification terms allow for slight improvements in both identity preservation as well as the accuracy of the edits.

8.4.3 *Comparison to the State of the Art*

8.4.3.1 *Image2StyleGAN*

Image2StyleGAN (Abdal et al., 2019) also projects real images to the StyleGAN latent space, and is thus a closely related approach. The source code of Image2StyleGAN was kindly provided by the authors. We show editing results using Image2StyleGAN embeddings in Figs. 8.1, 8.3, 8.4 and 8.5. Since these embeddings are optimized only using the synthesis terms and without using hierarchical optimization, the results are often implausible, as is most evident when editing the head pose and scene illumination. However, Image2StyleGAN projections are more detailed than ours. We also quantitatively compare to Image2StyleGAN in Tab. 8.2. Image2StyleGAN obtains the highest quality projections in terms of PSNR and SSIM. When combined with StyleRig, it also leads to low editing errors. However, the recognition errors are very high due to the artifacts in the results, as shown in the qualitative results.

8.4.3.2 *Other Aproaches*

We also compare PIE to a number of related techniques, X2Face (Wiles et al., 2018), Geng et al. (2018) and Siarohin et al. (2019). We compare the relighting capabilities of PIE to the single-image relighting approach of Zhou et al. (2019). The source codes of these techniques are publicly available. For Geng et al. (2018), we estimated the landmarks using the dlib tracker (King, 2009) as suggested by the authors. We also

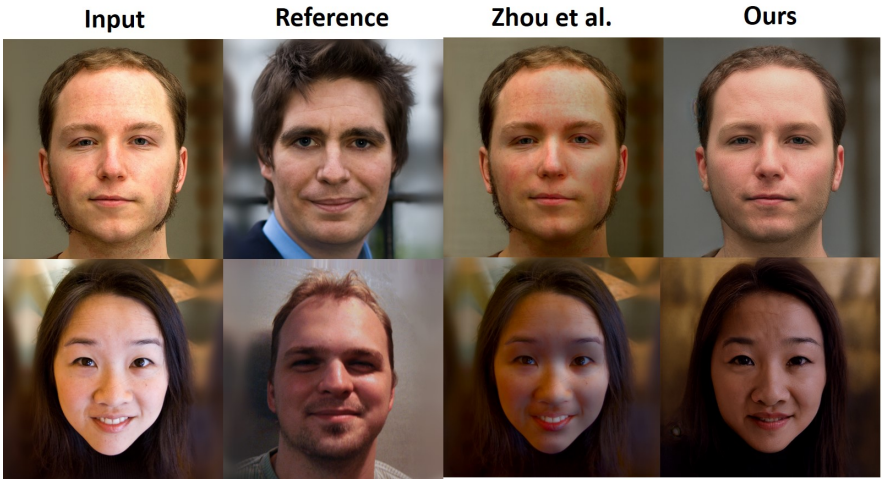

Figure 8.9: Comparison of the relighting results of PIE with Zhou et al. (2019). The illumination in the reference image is transferred to the input. The results of PIE are more natural and achieve more accurate relighting. While PIE can edit colored illumination while Zhou et al. (2019) can only edit monochrome light. In addition, we can also edit the head pose and facial expressions, while Zhou et al. (2019) is trained only for relighting. Images from Shih et al. (2014).

trained the few shot video-to-video translation method of Wang et al. (2019b) for portrait image editing. We trained on 700 videos from the FaceForensics dataset (Rossler et al., 2019). Landmarks were extracted using the dlib tracker as recommended by the authors. The approaches of Geng et al. (2018), Wiles et al. (2018) , Wang et al. (2019b) and Siarohin et al. (2019) are trained on a video corpus. In contrast, PIE does not use any direct supervision of the edited images. We compare to these methods in two different settings, self-reenactment and cross-identity reenactment.

SELF-REENACTMENT For self-reenactment, we capture several images of a person in different poses. We pick the first image and use the other images of the person as reference to edit the head pose. We captured 9 people in different poses, resulting in 31 images in the test set. Fig. 8.8 shows some qualitative results. Geng et al. (2018) use a warp-guided algorithm. While this enables expression changes and in-plane head motion, out-of-plane motion cannot be handled as shown in Fig. 8.8. We also compare to X2Face (Wiles et al., 2018), which samples a learned embedded face in order to synthesize portrait images with different poses and expressions. As such, it shares its limitations with Geng et al. (2018) and produces artifacts for strong pose changes. All approaches do not share the same cropping method, which makes it difficult to quantitatively evaluate the results. In addition, translation of the head during capture can lead to different illumination conditions. Thus, instead of directly computing errors in the image space, we first detect 66 facial landmarks (Saragih et al., 2009) on all results, as well as the reference images. We then compute the landmark alignment error, which is the averaged ℓ_2-distance between the landmarks after 2D Procrustes alignment (including scale). The implementation of Geng et al. (2018) often fails to generate such large pose edits, so we do not consider this approach in the quantitative evaluation. Due to artifacts, the landmark detector fails on 29% images for the approach of Wiles et al. (2018) and on 23% for Wang et al. (2019b). All the results

Table 8.3: Evaluation of pose edits: We measure landmark alignment errors for same-subject reenactment on 31 images, and facial recognition distances for cross-subject reenactment on 49 images. Existing landmark detection (Saragih et al., 2009) and facial recognition (King, 2009) often fail on images from competing methods, implying higher realism of PIE.

	Landmark Alignment (number of images)	Recognition (number of images)
Wiles et al. (2018)	**10.9 (22)**	0.52 (42)
Wang et al. (2019b)	28.19 (24)	0.49 (45)
Siarohin et al. (2019)	11.97 (31)	0.51 (46)
Ours	20.12 (31)	**0.40 (49)**

of PIE, as well as those of Siarohin et al. (2019) pass through the detector. This can be considered as a pseudo-metric of realism, since the landmark detector is trained on real portrait images, implying that the results are better than those of Wiles et al. (2018) and Wang et al. (2019b), and on par with Siarohin et al. (2019). Table 8.3 shows the errors for different methods. The low errors for Wiles et al. (2018) are possibly due to the landmark detector failing in challenging cases. We obtain only slightly worse results compared to Siarohin et al. (2019), even though PIE does not have access to ground truth during training. Siarohin et al. (2019) train on videos allowing for supervised learning. In addition, their edits are at a lower resolution of 256×256, compared to the image resolutions of 1024×1024 used in this chapter.

CROSS-IDENTITY REENACTMENT We also compare to others in cross-identity reenactment, which is closer to the setting of this chapter of semantically disentangled editing. Here, the image being edited and the reference image have different identities. Fig. 8.8 shows some qualitative results. The implementation of Geng et al. (2018) does not support this setting. Wiles et al. (2018) and Wang et al. (2019b) result in similar artifacts as discussed before. Unlike other approaches, Siarohin et al. (2019) uses two driving images in order to edit the input image, where they use the deformations between the two images as input. In the case of self-reenactment, we provide the input image as the first driving image. We do the same here, which leads to the two driving images with different identities. This significantly alters the facial identity in the output image. We also quantitatively evaluate the extent of identity preservation for different methods using a facial recognition tool (King, 2009), see Table. 8.3. All methods other than ours do not support semantically disentangled editing. As can be seen in Fig. 8.8 (bottom), other methods simultaneously change the expressions in addition to the head pose.

INTERACTIVE USER INTERFACE While all existing approaches need a driving image(s) for editing, we allow for explicit editing, using intuitive controls. An interactive user interface to edit images is developed. The user can change the head pose using a trackball mouse interface. Spherical harmonic coefficients and blendshape coefficients are changed using keyboard controls. All editing results run at around 5fps on a TITAN X Pascal GPU.

RELIGHTING We compare the relighting results of PIE to the single-image relighting approach of Zhou et al. (2019), see Fig. 8.9. PIE allows for colored illumination changes,

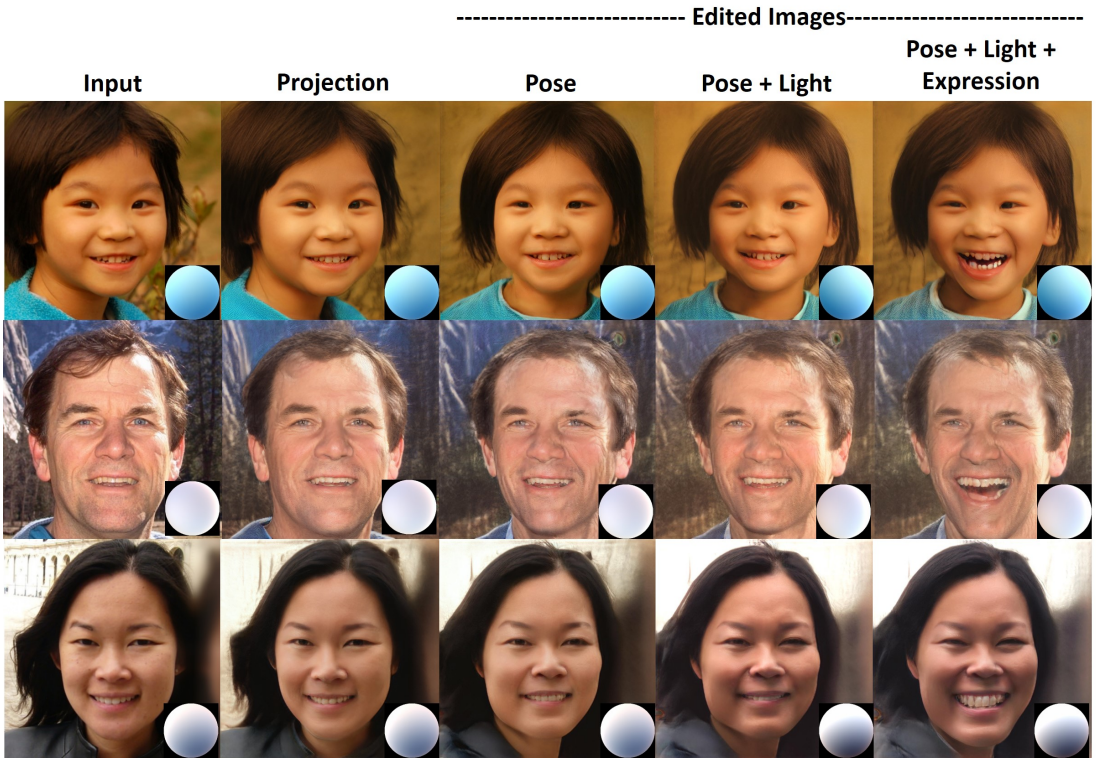

Figure 8.10: PIE also allows for sequential editing. We optimize for the StyleGAN embedding using the pose RigNet. We can then use the edited pose results with the RigNets for other semantic components for sequential editing. Images from Shen et al. (2016).

as shown in Fig. 8.4. PIE produces higher-quality and more realistic output images. PIE is also quantitatively compared to the relighting quality of these approaches in an illumination transfer setting, where the illumination in a reference image is transferred to a given input image. Since we do not have ground truth data available, the results are compared using a network which predicts the illumination from the reference and the relighted results. A model-based face autoencoder (Tewari et al., 2017), trained on the VoxCeleb dataset (Chung et al., 2018) is used. This network predicts a 27 dimensional spherical harmonics coefficients. The predictions are compared using a scale-invariant ℓ_2-loss. PIE obtains higher quality (0.34), compared to Zhou et al. (2019) (0.36). The numbers are averaged over 100 relighting results. While the method of Zhou et al. (2019) is only trained for relighting, PIE allows us to also edit the head pose and facial expressions.

8.4.4 *Generality of the embeddings*

SEQUENTIAL EDITING PIE also allows for sequential editing of the different semantic parameters, see Fig. 8.10. Here, the embedding is optimized using the pose RigNet network. After editing the pose, the new embedding can be used as input to the illumination and expression RigNets. Since all three versions of RigNet were trained on the same training data, this still produces plausible results.

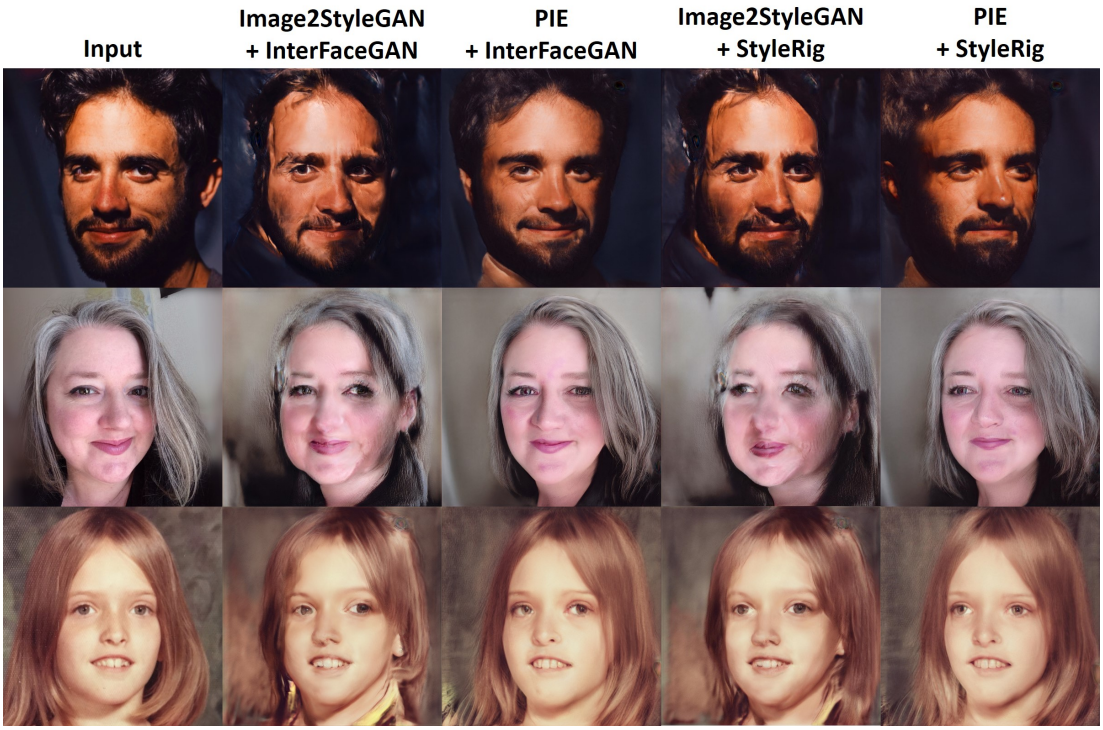

Figure 8.11: The embeddings of PIE obtain similar quality editing results with the InterFaceGAN (Shen et al., 2020) editing approach. Similar improvements over Image2StyleGAN (Abdal et al., 2019) embeddings can be noticed. Images from Shen et al. (2016).

OTHER STYLEGAN EDITING METHODS PIE obtains a StyleGAN embedding which can be edited using StyleRig. In order to test the generality of these embeddings, we attempt to edit them using InterFaceGAN (Shen et al., 2020), see Fig. 8.11. The improvements over Image2StyleGAN generalize to InterFaceGAN editings. PIE better preserves the facial identity and produce fewer artifacts. The editing results with InterFaceGAN are of a similar quality to those obtained using StyleRig. However, InterFaceGAN cannot change the scene illumination.

8.5 LIMITATIONS

Even though we have demonstrated a large variety of compelling portrait image editing results, there is still room for further improvement of our approach: (1) At the moment, our approach has a limited expressivity, i.e., it does not allow the artifact-free exploration of the whole parameter space of the underlying 3D morphable face model. For example, we cannot change the in-plane rotation of the face or arbitrarily change the lighting conditions. The main limiting factor is the training corpus (FFHQ (Karras et al., 2019a)) that has been used to pretrain the StyleGAN-generator, since it does not contain such variations. Due to the same reason, our approach is also not yet suitable for video-based facial reenactment, since the variety of facial expressions in the training corpus is severely limited. This problem could be alleviated by pretraining the generator on a larger and less biased training corpus that covers all dimensions well. (2) Our method only allows for independent control over the semantic parameters, which is important for editing applications. While sequential control is possible, simultaneous

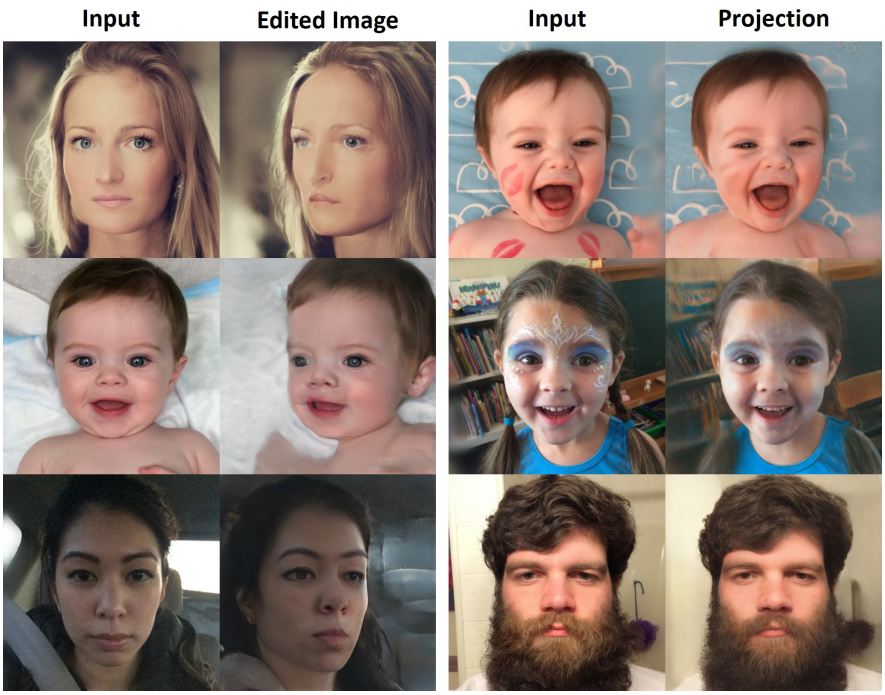

Figure 8.12: Limitations: Large edits can lead to artifacts. High-frequency texture on the foreground or background is difficult to fit. Our method also cannot handle cluttered backgrounds or occlusions. Images from Shen et al. (2016).

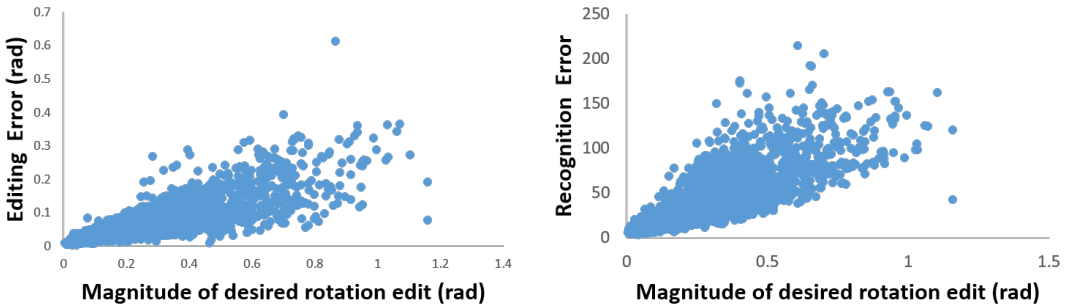

Figure 8.13: Scatterplot of the editing (left) and recognition errors (right), with respect to the magnitude of the desired pose edits for over 2500 pose editing results. Larger edits lead to both higher editing and recognition errors.

control is a more challenging problem. (3) Our approach does not provide explicit control over the synthesized background. At the moment, the background changes during the edits and does not remain static as it should, since the network has learned correlations between the face and the background. This could potentially be alleviated by learning an explicit foreground-background segmentation and having a consistency loss on the static background region. (4) In challenging cases with large deformations, cluttered backgrounds or occlusions and high-frequency textures, our method can fail to faithfully fit to the input image and preserve editing properties at the same time, see Fig. 8.12. In addition, 3D face reconstruction also often fails under occlusions which would lead to incorrect data for our approach. (5) Larger edits generally correspond to worse results, and can often lead to artifacts, as shown in Fig. 8.12. This can also be seen in Fig. 8.13, where larger pose edits correlate with higher editing and facial recognition errors. (6) Similar to StyleGAN, our approach also sometimes shows droplet-like artifacts. This could be alleviated by switching to a higher quality generator architecture, such as StyleGAN2 (Karras et al., 2019b), which has been shown to solve this problem. (7) While we show results for people of different ethnicities, genders and ages, we did not extensively study the biases present in the method. Some of the components used, such as the 3DMM are known to have racial biases, see Chapter 5. (8) Our results are not guaranteed to be temporally consistent. The results could be made more temporally consistent by employing a temporal network architecture and space-time versions of our losses. Nevertheless, our approach, already now, enables the intuitive editing of portrait images at interactive frame rates.

8.6 CONCLUSION

This chapter presented the first approach for embedding portrait photos in the latent space of StyleGAN, which allows for intuitive editing of the head pose, facial expression, and scene illumination. To this end, a hierarchical optimization scheme was devised that embeds a real portrait image in the latent space of a generative adversarial network, while ensuring the editability of the recovered latent code. Semantic editing is achieved by mapping the control space of a 3D morphable face model to the latent space of the generator. In addition, a novel identity preservation loss enables to better preserve the facial identity.

This approach is a first step towards intuitive and interactive editing of portrait images using a semantic control space akin to computer animation controls. In addition, the approach provides more insights into the inner workings of GANs, since it allows the intuitive and interactive exploration of the space of face images. This can shed light on the biases the model has learned from the employed training corpus. The methods presented in Chapters 7 and 8 bring the two different domains of 2D and 3D face models together, thus opening the road towards even more interesting edits.

9 CONCLUSION

This thesis proposed several methods for self-supervised 3D face reconstruction and controllable synthesis of portrait images. Chapter 4 demonstrated self-supervised monocular 3D reconstruction using a pretrained 3DMM prior. Chapter 5 proposed a method for the refinement of the pretrained 3DMM using a dataset of monocular images. This direction was extended in Chapter 6, where the identity components of the 3DMM were learned entirely from videos, without using any 3D supervision. Chapter 7 built connections between the self-supervised 3D reconstruction pipeline and a high-quality neural generative model (StyleGAN), allowing for semantic control of StyleGAN generated images. Finally, Chapter 8 presented a method for building an intuitive and interactive image editing system that can process existing real images. All methods proposed in this thesis did not use supervised training. The methods instead relied on different priors, such as a 3DMM for reconstruction, and StyleGAN for synthesis.

9.1 INSIGHTS AND OUTLOOK

SELF-SUPERVISED LEARNING This thesis took steps towards learning a 3D morphable model from in-the-wild 2D data. While the identity components of the model could be learned entirely from 2D data, all methods relied on an existing 3D expression model. The recent work of B R et al. (2021b) has shown the possibility of learning all components of the model from 2D data. Only a template face mesh is used as a prior. Learning from in-the-wild data makes lifelong learning (Parisi et al., 2019) a possibility. Developing methods that can continually learn from the virtually endless stream of images and videos is an interesting direction for future work.

Learning from in-the-wild data comes with some limitations. In the absence of the correct camera parameters, all camera intrinsics and extrinsics are assumed to be identical. In addition, in-the-wild data is often of low quality with noise, motion blur and compression artifacts. These effects are not modeled in the learning process, which results in lower quality 3D reconstructions. Disentangling the different 3D components from in-the-wild data is also challenging, as shown in this thesis. While the methods presented in the thesis learn disentangled models, this disentanglement is not perfect. For example, facial hair is often learned only in the appearance component, and illumination and albedo are not disentangled perfectly. While these limitations can be avoided by training on high-quality datasets where all capture parameters are known, and the face is captured under multiple light conditions with multiple cameras, such datasets can be expensive to acquire and not readily available at a large scale. Thus, it is important to develop methods that can better disentangle the different components without using large supervised datasets.

Chapter 8 enabled intuitive 3D editing of portrait images at photorealistic quality. The method does not use any supervised data and only relies on a single image of the person. While this demonstrates high-quality editing under very constrained

settings, this also leaves some room for improvements. For example, the structure of the teeth has to be hallucinated if the input image has a face with its mouth closed. A better balance between accuracy and generalization could be achieved by using more images of the person in different scene conditions. In addition, as mentioned in Chapter 7, the StyleGAN prior limits the expressivity of the face. Learning more expressive generative models is an important problem for future work.

This thesis used a variety of in-the-wild data during training, such as images and video datasets. However, other related data such as the speech or transcript of the video were not used. Such multi-modal data could be used to build more accurate and better disentangled models, as well as interesting applications such as text-to-video (Fried et al., 2019) and speech-to-video (Thies et al., 2020) synthesis.

LEARNING HIGHER-QUALITY MODELS This thesis only models the surface of the face, without the skull. Extending the methods proposed in this thesis to also consider the skull of the person would allow for anatomically constrained deformations (Wu et al., 2016), as well as applications in digital forensics (Gietzen et al., 2019; Ubelaker, 2015).

Identity and expressions are modeled as independent components, and person-specific expressions are not captured. High-frequency wrinkles are also person-specific (Cao et al., 2015; Garrido et al., 2016a); thus, methods developed in this thesis cannot directly be used for modeling them. While Chapter 4 presented a method for reconstructing high-frequency details, a generative model of face wrinkles is an open problem. Learning face dynamics by building temporal generative models is also an open problem, important for video editing tasks.

Most existing 3DMMs only focus on the diffuse albedo of the face. The recent model of Smith et al. (2020) also includes the specular component. However, skin appearance also includes higher-order light transport effects. Priors over higher quality skin appearance are necessary for photorealistic reconstructions. First steps have been taken by B R et al. (2021a), where a neural representation of skin appearance is developed, along with a method for its reconstruction from monocular images. This approach requires a light stage dataset for training, which is expensive to capture and not publicly available. Learning high quality appearance models from in-the-wild data is still an open problem.

Most 3DMMs only model the frontal face region, and not the full head. Hair is especially challenging due to its complex geometry and appearance. A fixed mesh template might be insufficient for capturing hair due to its large deformations. First steps have been taken by Yenamandra et al. (2021) for building template-free implicit morphable models of full heads including hair. However, high-quality parameteric models of hair geometry and appearance are still an open problem.

It is not clear what representation is ideal for learning morphable models. This thesis relied on a mesh-based representation. However, other representations could be better suited for this task. Recent volumetric representations (Lombardi et al., 2019; Mildenhall et al., 2020) could be ideal for representing hair. Exploring these representations for higher-quality and more complete models is an important direction for future work.

Learning 3D morphable models directly from 2D data would also allow for modeling different classes of objects, where 3D scans are difficult to capture, such as animals. The methods in this thesis relied on components such as face keypoint detectors.

Building such detectors for other deformable object classes would allow for extending the applicability of these methods. Alternatively, developing new methods which do not require such annotations would make them more widely applicable.

DIFFERENTIABLE RENDERING Differentiable rendering is a crucial component which enables self-supervised learning, see Fig. 2.2. The differentiable renderer used in this thesis allowed for learning high-quality geometry and appearance models of human faces without the use of supervised training datasets. However, the approximations made in the image formation process (explained in Chapter 2) can limit the reconstruction quality. For example, the monocular reconstruction methods can struggle in the presence of cast shadows, since the differentiable renderer only models direct illumination. Renderers which can model higher-order light transport effects will allow for more accurate reconstructions, as well as finer control over them. Ray tracing-based differentiable renderers (Li et al., 2018; Nimier-David et al., 2019) can model complex light transport effects, but are significantly more computationally expensive compared to the renderer used in this thesis. Developing efficient differentiable renderers which account for global illumination would be important for 3D reconstruction tasks.

9.2 SOCIAL IMPLICATIONS

The biases present in the proposed methods have not been studied extensively. Chapter 5 demonstrated that the baseline 3D morphable model (Blanz and Vetter, 1999) learned from 3D scans has racial biases. One reason is that the diversity of races in the training dataset was limited. While the method presented in the same chapter reduces this bias by using larger-scale datasets, the extent of the remaining bias is unclear. Racial and gender biases have been demonstrated for face recognition (Phillips et al., 2003) and gender classification (Buolamwini and Gebru, 2018) problems. It would be important to inspect the biases present in the methods proposed in this thesis, and develop methods to mitigate them. Training 3D models from 2D data is a good direction, as it removes the need for very diverse and large 3D datasets, which can be very difficult to capture.

Editing of portrait images has applications in casual photography and content creation. However, these editing methods could also be misused to misrepresent people in images, for example, by changing their expressions. Detection of synthesized images is thus an important problem. Several such methods exist, including active methods which modify the imaging pipeline (Blythe and Fridrich, 2004; Korus and Memon, 2019; Yan and Pun, 2017), as well as automatic passive methods (Cozzolino et al., 2018; Fox et al., 2020; Rossler et al., 2019; Wang et al., 2019a, 2020). The work of Wang et al. (2020) showed that images generated by StyleGAN can be correctly distinguished from real images with high accuracy. Since the methods developed in Chapters 7 and 8 rely on StyleGAN for synthesis, their results can also likely be distinguished from real images. Further analysis is required to compute the precise detection rates for these methods.

Self-supervised learning is an exciting and promising problem, as it allows for utilizing virtually unlimited images and videos available online for training. The ideas presented in this thesis will hopefully inspire follow-up work on self-supervised learning of photorealistic and semantically controllable models of the different objects around us.

BIBLIOGRAPHY

Abadi, Martín, 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, Dan 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 (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: http://tensorflow.org/.

Abdal, Rameen, Yipeng Qin, and Peter Wonka (2019). "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?" In: *International Conference on Computer Vision (ICCV)*.

– (2020a). "Image2StyleGAN++: How to Edit the Embedded Images?" In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abdal, Rameen, Peihao Zhu, Niloy Mitra, and Peter Wonka (2020b). *StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows*. arXiv: 2008.02401 [cs.CV].

Alexander, Julia (2019). *James Dean, who died in 1955, just landed a new movie role, thanks to CGI*. https://www.theverge.com/2019/11/6/20951485/james-dean-new-movie-cgi-recreation-finding-jack.

Alexander, Oleg, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec (2009). "The Digital Emily Project: Photoreal Facial Modeling and Animation." In: *ACM SIGGRAPH 2009 Courses*. SIGGRAPH '09. Association for Computing Machinery. ISBN: 9781450379380. DOI: 10.1145/1667239.1667251. URL: https://doi.org/10.1145/1667239.1667251.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein gan." In: *arXiv preprint arXiv:1701.07875*.

Averbuch-Elor, Hadar, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen (2017). "Bringing Portraits to Life." In: *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)* 36.6, p. 196.

B R, Mallikarjun, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt (2021a). "Monocular Reconstruction of Neural Face Reflectance Fields." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

B R, Mallikarjun, Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt (2021b). "Learning Complete 3D Morphable Face Models from Images and Videos." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bansal, Aayush, Shugao Ma, Deva Ramanan, and Yaser Sheikh (2018). "Recycle-GAN: Unsupervised Video Retargeting." In: *ECCV*.

Bas, Anil and William A. P. Smith (2018). "Statistical transformer networks: learning shape and appearance models via self supervision." arXiv:1804.02541.

Beeler, Thabo, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross (July 2010). "High-quality Single-shot Capture of Facial Geometry." In: *ACM Trans. Graph.* 29.4, 40:1–40:9. ISSN: 0730-0301.

Beeler, Thabo, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross (2011). "High-quality Passive Facial Performance Capture Using Anchor Frames." In: *ACM SIGGRAPH 2011 Papers*. SIGGRAPH '11. ACM, 75:1–75:10. ISBN: 978-1-4503-0943-1.

Bernard, Florian, Peter Gemmar, Frank Hertel, Jorge Goncalves, and Johan Thunberg (2016). "Linear shape deformation models with local support using graph-based structured matrix factorisation." In: *CVPR*.

Blanz, Volker, Curzio Basso, Tomaso Poggio, and Thomas Vetter (2003). "Reanimating faces in images and video." In: *Computer graphics forum*. Wiley Online Library, pp. 641–650.

Blanz, Volker, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel (2004). "Exchanging faces in images." In: *Computer Graphics Forum*. Wiley Online Library, pp. 669–676.

Blanz, Volker and Thomas Vetter (1999). "A morphable model for the synthesis of 3D faces." In: *Proc. SIGGRAPH*. ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

Blythe, Paul and Jessica Fridrich (2004). "Secure digital camera." In: *Digital Investigation*.

Bogo, Federica, Javier Romero, Matthew Loper, and Michael J. Black (2014). "FAUST: Dataset and Evaluation for 3D Mesh Registration." In: *CVPR '14*. IEEE Computer Society, pp. 3794–3801.

Bolkart, Timo and Stefanie Wuhrer (2015). "A Groupwise Multilinear Correspondence Optimization for 3D Faces." In: *ICCV*. IEEE Computer Society, pp. 3604–3612.

– (2016). "A Robust Multilinear Model Learning Framework for 3D Faces." In: *CVPR*. IEEE Computer Society, pp. 4911–4919.

Bonneel, Nicolas, Kalyan Sunkavalli, James Tompkin, Deqing Sun, Sylvain Paris, and Hanspeter Pfister (2014). "Interactive Intrinsic Video Editing." In: *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2014)* 33.6.

Booth, James, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou (July 2017). "3D Face Morphable Models "In-The-Wild"." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Booth, James, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway (2016). "A 3d morphable model learnt from 10,000 faces." In: *CVPR*.

Botsch, Mario and Olga Sorkine (Jan. 2008). "On Linear Variational Surface Deformation Methods." In: *IEEE Transactions on Visualization and Computer Graphics* 14.1, pp. 213–230. ISSN: 1077-2626. DOI: 10.1109/TVCG.2007.1054. URL: http://dx.doi.org/10.1109/TVCG.2007.1054.

Bouaziz, Sofien, Yangang Wang, and Mark Pauly (2013). "Online Modeling for Realtime Facial Animation." In: *ACM Trans. Graph.* 32.4, 40:1–40:10.

Bradski, G. (2000). "The OpenCV Library." In: *Dr. Dobb's Journal of Software Tools*.

Buolamwini, Joy and Timnit Gebru (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification." In: *Conference on fairness, accountability and transparency*, pp. 77–91.

Cao, Chen, Derek Bradley, Kun Zhou, and Thabo Beeler (July 2015). "Real-time High-fidelity Facial Performance Capture." In: *ACM Trans. Graph.* 34.4, 46:1–46:9. ISSN: 0730-0301.

Cao, Chen, Qiming Hou, and Kun Zhou (2014). "Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation." In: *ACM Trans. Graph.* 33.4, 43:1–43:10.

Cao, Chen, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou (2013). "Facewarehouse: A 3d facial expression database for visual computing." In: *IEEE Transactions on Visualization and Computer Graphics* 20.3, pp. 413–425.

Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros (2019). "Everybody Dance Now." In: *International Conference on Computer Vision (ICCV)*.

Chaudhuri, Bindita, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang (2020). "Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting." In: *IEEE European Conference on Computer Vision (ECCV)*.

Chrysos, Grigoris G., Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape (Dec. 2015). "Offline Deformable Face Tracking in Arbitrary Videos." In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Chung, J. S., A. Nagrani, and A. Zisserman (2018). "VoxCeleb2: Deep Speaker Recognition." In: *INTERSPEECH*.

Cootes, Timothy F., Gareth J. Edwards, and Christopher J. Taylor (June 2001). "Active appearance models." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6, pp. 681–685.

Cozzolino, Davide, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva (2018). "ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection." In: *arXiv*.

Delaunoy, Amaël and Emmanuel Prados (Nov. 2011). "Gradient Flows for Optimizing Triangular Mesh-based Surfaces: Applications to 3D Reconstruction Problems Dealing with Visibility." In: *Int. J. Comput. Vision* 95.2, pp. 100–123. ISSN: 0920-5691.

Deng, Yu, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong (2019). "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0.

Doersch, Carl (2016). "Tutorial on variational autoencoders." In: *arXiv preprint arXiv:1606.05908*.

Dou, Pengfei, Shishir K. Shah, and Ioannis A. Kakadiaris (July 2017). "End-To-End 3D Face Reconstruction With Deep Neural Networks." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Egger, Bernhard (2017). "Semantic morphable models." PhD thesis. University_of_Basel.

Egger, Bernhard, 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 (2020). "3D Morphable Face Models - Past, Present and Future." In: *ACM Transactions on Graphics (TOG)*.

Ekman, Paul and Erika L Rosenberg (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Elgharib, Mohamed, Mohit Mendiratta, Justus Thies, Matthias Nießner, Hans-Peter Seidel, Ayush Tewari, Vladislav Golyanik, and Christian Theobalt (Dec. 2020). "Ego-

centric Videoconferencing." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*.

Fox, Gereon, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt (2020). "VideoForensicsHQ: Detecting High-quality Manipulated Face Videos." In: *arXiv preprint arXiv:2005.10360*.

Fried, Ohad, Eli Shechtman, Dan B. Goldman, and Adam Finkelstein (July 2016). "Perspective-aware Manipulation of Portrait Photos." In: *ACM Trans. Graph.* 35.4.

Fried, Ohad, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala (2019). "Text-based Editing of Talking-head Video." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH)*.

Garg, Ravi, Anastasios Roussos, and Lourdes Agapito (2013). "Dense Variational Reconstruction of Non-rigid Surfaces from Monocular Video." In: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, pp. 1272–1279.

Garrido, Pablo, Levi Valgaerts, Chenglei Wu, and Christian Theobalt (Nov. 2013). "Reconstructing Detailed Dynamic Face Geometry from Monocular Video." In: *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*. Vol. 32. 6, 158:1–158:10.

Garrido, Pablo, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt (June 2016a). "Reconstruction of Personalized 3D Face Rigs from Monocular Video." In: *ACM Transactions on Graphics* 35.3, 28:1–15.

Garrido, Pablo, Michael Zollhöfer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt (2016b). "Corrective 3D Reconstruction of Lips from Monocular Video." In: *ACM Trans. Graph.* 35.6, 219:1–219:11.

Gatys, L. A., A. S. Ecker, and M. Bethge (2016). "Image Style Transfer Using Convolutional Neural Networks." In: *CVPR*, pp. 2414–2423.

Geng, Jiahao, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou (2018). "Warp-guided GANs for single-photo facial animation." In: *ACM Trans. Graph.* 37, 231:1–231:12.

Genova, Kyle, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman (June 2018). "Unsupervised Training for 3D Morphable Model Regression." In: *CVPR*.

Gerig, Thomas, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter (2018). "Morphable face models-an open framework." In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 75–82.

Gietzen, Thomas, Robert Brylka, Jascha Achenbach, Katja zum Hebel, Elmar Schömer, Mario Botsch, Ulrich Schwanecke, and Ralf Schulze (2019). "A method for automatic forensic facial reconstruction based on dense statistics of soft tissue thickness." In: *PloS one* 14.1, e0210257.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680.

Grant, Edward, Pushmeet Kohli, and Marcel van Gerven (2016). "Deep disentangled representations for volumetric reconstruction." In: *ECCVW*.

Hinton, G E and R R Salakhutdinov (July 2006). "Reducing the dimensionality of data with neural networks." In: *Science* 313.5786, pp. 504–507.

Hsieh, Pei-Lun, Chongyang Ma, Jihun Yu, and Hao Li (2015). "Unconstrained realtime facial performance capture." In: *CVPR*. IEEE Computer Society, pp. 1675–1683.

Huang, Gary B., Manu Ramesh, Tamara Berg, and Erik Learned-Miller (Oct. 2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst.

Huber, Patrik, Philipp Kopp, Matthias Rätsch, William Christmas, and Josef Kittler (May 2016). "3D Face Tracking and Texture Fusion in the Wild." arXiv:1605.06764.

Huynh, Loc, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li (June 2018). "Mesoscopic Facial Geometry Inference Using Deep Neural Networks." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Härkönen, Erik, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris (2020). *GANSpace: Discovering Interpretable GAN Controls*. arXiv: 2004.02546 [cs.CV].

Ichim, Alexandru Eugen, Sofien Bouaziz, and Mark Pauly (2015). "Dynamic 3D Avatar Creation from Hand-held Video Input." In: *ACM Trans. Graph.* 34.4, 45:1–45:14.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). "Image-to-Image Translation with Conditional Adversarial Networks." In: *CVPR*.

Itzkoff, Dave (2016). *How 'Rogue One' Brought Back Familiar Faces*. https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html.

Jackson, Aaron S., Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos (Oct. 2017). "Large Pose 3D Face Reconstruction From a Single Image via Direct Volumetric CNN Regression." In: *The IEEE International Conference on Computer Vision (ICCV)*.

Jacobson, Alec, Zhigang Deng, Ladislav Kavan, and JP Lewis (2014). "Skinning: Real-time Shape Deformation." In: *ACM SIGGRAPH 2014 Courses*.

Jahanian, Ali, Lucy Chai, and Phillip Isola (2019). "On the "steerability" of generative adversarial networks." In: *arXiv preprint arXiv:1907.07171*.

Jensen, Henrik Wann, Stephen R Marschner, Marc Levoy, and Pat Hanrahan (2001). "A practical model for subsurface light transport." In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 511–518.

Jia, Yangqing, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell (2014). "Caffe: Convolutional Architecture for Fast Feature Embedding." In: *arXiv preprint arXiv:1408.5093*.

Jin, Xin and Xiaoyang Tan (Aug. 2016). "Face Alignment In-the-Wild: A Survey." arXiv:1608.04188.

Johnson, Justin, Alexandre Alahi, and Li Fei-Fei (2016). "Perceptual losses for real-time style transfer and super-resolution." In: *European Conference on Computer Vision*.

Kajiya, James T (1986). "The rendering equation." In: *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pp. 143–150.

Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen (2018). "Progressive Growing of GANs for Improved Quality, Stability, and Variation." In: *International Conference on Learning Representation (ICLR)*.

Karras, Tero, Samuli Laine, and Timo Aila (2019a). "A style-based generator architecture for generative adversarial networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila (2019b). "Analyzing and Improving the Image Quality of StyleGAN." In: *CoRR* abs/1912.04958.

Kemelmacher-Shlizerman, Ira (2013). "Internet-based Morphable Model." In: *International Conference on Computer Vision (ICCV)*.

Kemelmacher-Shlizerman, Ira, Aditya Sankar, Eli Shechtman, and Steven M. Seitz (2010). "Being John Malkovich." In: *ECCV*.

Kemelmacher-Shlizerman, Ira and Steven M. Seitz (2011). "Face Reconstruction in the Wild." In: *ICCV*.

Kim, H., M. Elgharib, M. Zollhöfer, H.-P. Seidel, T. Beeler, C. Richardt, and C. Theobalt (2019). "Neural Style-Preserving Visual Dubbing." In: *ACM Transactions on Graphics (TOG)*.

Kim, Hyeongwoo, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt (2018a). "Deep Video Portraits." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH)*.

Kim, Hyeongwoo, Michael Zollöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt (2018b). "InverseFaceNet: Deep Single-Shot Inverse Face Rendering From A Single Image." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

King, Davis E. (2009). "Dlib-ml: A Machine Learning Toolkit." In: *Journal of Machine Learning Research* 10, pp. 1755–1758.

Kingma, Durk P and Prafulla Dhariwal (2018). "Glow: Generative flow with invertible 1x1 convolutions." In: *Advances in neural information processing systems*, pp. 10215–10224.

Klaudiny, Martin, Steven McDonagh, Derek Bradley, Thabo Beeler, and Kenny Mitchell (2017). "Real-Time Multi-View Facial Capture with Synthetic Training." In: *Comput. Graph. Forum*.

Korus, Pawel and Nasir Memon (2019). "Content authentication for neural imaging pipelines: End-to-end optimization of photo provenance in complex distribution channels." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8621–8629.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks." In: *NIPS*.

Kulkarni, Tejas D., Will Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum (2015). "Deep Convolutional Inverse Graphics Network." In: *NIPS*.

Laine, Samuli, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen (2017). "Production-level Facial Performance Capture Using Deep Convolutional Neural Networks." In: *SCA*. ACM, 10:1–10:10.

Lewis, J. P., Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng (2014a). "Practice and Theory of Blendshape Facial Models." In: *Eurographics STARs*, pp. 199–218.

Lewis, John P, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng (2014b). "Practice and Theory of Blendshape Facial Models." In: *Eurographics (State of the Art Reports)* 1.8, p. 2.

Li, Shiwei, Sing Yu Siu, Tian Fang, and Long Quan (2016). "Efficient Multi-view Surface Refinement with Adaptive Resolution Control." In: *Computer Vision - ECCV 2016 -*

14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, pp. 349–364.

Li, Tianye, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero (2017). "Learning a Model of Facial Shape and Expression from 4D Scans." In: *ACM Trans. Graph.* 36.6, 194:1–194:17.

Li, Tzu-Mao, Michaël Gharbi, Andrew Adams, Frédo Durand, and Jonathan Ragan-Kelley (2018). "Differentiable programming for image processing and deep learning in halide." In: *ACM Transactions on Graphics (TOG)* 37.4, p. 139.

Liang, Shu, Linda G Shapiro, and Ira Kemelmacher-Shlizerman (2016). "Head Reconstruction from Internet Photos." In: *European Conference on Computer Vision*. Springer, pp. 360–374.

Lin, Jiangke, Yi Yuan, Tianjia Shao, and Kun Zhou (2020). "Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5891–5900.

Liu, Ming-Yu, Thomas Breuel, and Jan Kautz (2017). "Unsupervised Image-to-Image Translation Networks." In: *NIPS*, pp. 700–708.

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang (Dec. 2015). "Deep Learning Face Attributes in the Wild." In: *Proceedings of International Conference on Computer Vision (ICCV)*.

Lombardi, Stephen, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh (July 2019). "Neural Volumes: Learning Dynamic Renderable Volumes from Images." In: *ACM Trans. Graph.* 38.4, 65:1–65:14.

Luan, Fujun, Sylvain Paris, Eli Shechtman, and Kavita Bala (2017). "Deep Photo Style Transfer." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6997–7005.

Lüthi, M., T. Gerig, C. Jud, and T. Vetter (2018). "Gaussian Process Morphable Models." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.8, pp. 1860–1873.

Mao, Xudong, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley (2017). "Least squares generative adversarial networks." In: *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802.

Martin-Brualla, Ricardo, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello (2018). "LookinGood: Enhancing Performance Capture with Real-time Neural Re-rendering." In: *ACM Trans. on Graph. (Proceedings of SIGGRAPH-Asia)* 37.6, 255:1–255:14.

Masci, Jonathan, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber (2011). "Stacked Convolutional Auto-encoders for Hierarchical Feature Extraction." In: *International Conference on Artificial Neural Networks*.

McDonagh, Steven, Martin Klaudiny, Derek Bradley, Thabo Beeler, Iain Matthews, and Kenny Mitchell (2016). "Synthetic Prior Design for Real-Time Face Tracking." In: *3DV* 00, pp. 639–648.

Meka, Abhimitra, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann (July 2019). "Deep Reflectance Fields - High-

Quality Facial Reflectance Field Inference From Color Gradient Illumination." In: vol. 38. 4. DOI: 10.1145/3306346.3323027. URL: http://gvv.mpi-inf.mpg.de/projects/DeepReflectanceFields/.

Meka, Abhimitra, Michael Zollhöfer, Christian Richardt, and Christian Theobalt (2016). "Live Intrinsic Video." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 35.4.

Mescheder, Lars, Andreas Geiger, and Sebastian Nowozin (2018). "Which training methods for GANs do actually converge?" In: *arXiv preprint arXiv:1801.04406*.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng (2020). "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." In: *ECCV*.

Müller, Claus (1966). *Spherical harmonics*. Springer.

NVIDIA (2008). *NVIDIA CUDA Programming Guide 2.0*.

Nagano, Koki, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li (Dec. 2018). "paGAN: real-time avatars using dynamic textures." In: pp. 1–12. DOI: 10.1145/3272127.3275075.

Neumann, Thomas, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt (2013). "Sparse Localized Deformation Components." In: *ACM Trans. Graph.* 32.6, 179:1–179:10.

Nimier-David, Merlin, Delio Vicini, Tizian Zeltner, and Wenzel Jakob (2019). "Mitsuba 2: A retargetable forward and inverse renderer." In: *ACM Transactions on Graphics (TOG)* 38.6, pp. 1–17.

Olszewski, Kyle, Joseph J. Lim, Shunsuke Saito, and Hao Li (2016). "High-Fidelity Facial and Speech Animation for VR HMDs." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2016)* 35.6.

Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu (2016). "Pixel recurrent neural networks." In: *arXiv preprint arXiv:1601.06759*.

Parisi, German I, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter (2019). "Continual lifelong learning with neural networks: A review." In: *Neural Networks* 113, pp. 54–71.

Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu (2019). "Semantic Image Synthesis with Spatially-Adaptive Normalization." In: *CVPR*.

Parkhi, O. M., A. Vedaldi, and A. Zisserman (2015). "Deep Face Recognition." In: *British Machine Vision Conference*.

A 3D Face Model for Pose and Illumination Invariant Face Recognition (2009). IEEE. Genova, Italy.

Peers, Pieter, Naoki Tamura, Wojciech Matusik, and Paul Debevec (July 2007). "Post-production Facial Performance Relighting Using Reflectance Transfer." In: *ACM Trans. Graph.* 26.3.

Pharr, Matt, Wenzel Jakob, and Greg Humphreys (2016). *Physically based rendering: From theory to implementation*. Morgan Kaufmann.

Phillips, P J, Patrick J Grother, Ross J Micheals, D M Blackburn, Elham Tabassi, and Mike Bone (2003). *Face recognition vendor test 2002: Evaluation report*. Tech. rep.

Piotraschke, Marcel and Volker Blanz (2016). "Automated 3D Face Reconstruction from Multiple Images Using Quality Measures." In: *CVPR*. IEEE Computer Society, pp. 3418–3427.

Radford, Alec, Luke Metz, and Soumith Chintala (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." In: *arXiv preprint arXiv:1511.06434*.

Ramamoorthi, Ravi and Pat Hanrahan (2001a). "A signal-processing framework for inverse rendering." In: *Proc. SIGGRAPH*. ACM, pp. 117–128.

– (2001b). "An efficient representation for irradiance environment maps." In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 497–500.

Ranjan, Anurag, Timo Bolkart, Soubhik Sanyal, and Michael J. Black (2018). "Generating 3D Faces Using Convolutional Mesh Autoencoders." In: *ECCV '18*. Vol. 11207. Lecture Notes in Computer Science. Springer, pp. 725–741.

Richardson, Elad, Matan Sela, and Ron Kimmel (2016). "3D Face Reconstruction by Learning from Synthetic Data." In: *3DV*.

Richardson, Elad, Matan Sela, Roy Or-El, and Ron Kimmel (July 2017). "Learning Detailed Face Reconstruction From a Single Image." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Romdhani, Sami and Thomas Vetter (2005). "Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior." In: *CVPR*. Washington, DC, USA: IEEE Computer Society, pp. 986–993.

Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner (2019). "Faceforensics++: Learning to detect manipulated facial images." In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–11.

Roth, Joseph, Yiying Tong, and Xiaoming Liu (Dec. 2016). "Adaptive 3D Face Reconstruction from Unconstrained Photo Collections." In:

Saito, Shunsuke, Tianye Li, and Hao Li (2016). "Real-Time Facial Segmentation and Performance Capture from RGB Input." In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Sanyal, Soubhik, Timo Bolkart, Haiwen Feng, and Michael Black (2019). "Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision." In: *CVPR*, pp. 7763–7772.

Saragih, Jason M., Simon Lucey, and Jeffrey F. Cohn (2009). "Face Alignment through Subspace Constrained Mean-Shifts." In: *Proc. ICCV*, pp. 1034–1041.

– (2011). "Deformable Model Fitting by Regularized Landmark Mean-Shift." In: *IJCV* 91.2.

Schönborn, Sandro, Bernhard Egger, Andreas Forster, and Thomas Vetter (July 2015). "Background Modeling for Generative Image Models." In: *Comput. Vis. Image Underst.* 136.C, pp. 117–127.

Sela, Matan, Elad Richardson, and Ron Kimmel (2017). "Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation." In: *ICCV*.

Selim, Ahmed, Mohamed Elgharib, and Linda Doyle (2016). "Painting Style Transfer for Head Portraits using Convolutional Neural Networks." In: *ACM Trans. on Graph. (Proceedings of SIGGRAPH)*, 129:1–129:18.

Sengupta, Soumyadip, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs (2018). "SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild." In: *Computer Vision and Pattern Regognition (CVPR)*.

Shen, Jie, Stefanos Zafeiriou, Grigoris G. Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic (Dec. 2015). "The First Facial Landmark Tracking In-the-Wild Challenge: Benchmark and Results." In: *ICCVW*.

Shen, Xiaoyong, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia (2016). "Deep automatic portrait matting." In: *European conference on computer vision*. Springer, pp. 92–107.

Shen, Yujun, Jinjin Gu, Xiaoou Tang, and Bolei Zhou (2020). "Interpreting the Latent Space of GANs for Semantic Face Editing." In: *CVPR*.

Shi, Fuhao, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai (2014). "Automatic Acquisition of High-fidelity Facial Performances Using Monocular Videos." In: *ACM Trans. Graph.* 33.6, 222:1–222:13.

Shih, YiChang, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand (July 2014). "Style Transfer for Headshot Portraits." In: *ACM Trans. Graph.* 33.4. ISSN: 0730-0301. DOI: 10.1145/2601097.2601137. URL: https://doi.org/10.1145/2601097.2601137.

Shu, Zhixin, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras (July 2017a). "Portrait Lighting Transfer Using a Mass Transport Approach." In: *ACM Trans. Graph* 36.4.

Shu, Zhixin, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras (2017b). "Neural Face Editing with Intrinsic Image Disentangling." In: *CVPR*.

Siarohin, Aliaksandr, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe (Dec. 2019). "First Order Motion Model for Image Animation." In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Simonyan, Karen and Andrew Zisserman (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *International Conference on Learning Representations*.

Smith, William AP, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger (2020). "A morphable face albedo model." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5011–5020.

Sumner, Robert W and Jovan Popović (2004). "Deformation transfer for triangle meshes." In: *ACM TOG* 23.3, pp. 399–405.

Sumner, Robert W, Johannes Schmid, and Mark Pauly (2007). "Embedded deformation for shape manipulation." In: *ACM Transactions on Graphics (TOG)*. Vol. 26. 3. ACM, p. 80.

Sun, Qianru, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele (2018). "A Hybrid Model for Identity Obfuscation by Face Replacement." In: *European Conference on Computer Vision (ECCV)*.

Sun, Tiancheng, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi (July 2019). "Single Image Portrait Relighting." In: *ACM Trans. Graph.* 38.4, 79:1–79:12. ISSN: 0730-0301. DOI: 10.1145/3306346.3323008. URL: http://doi.acm.org/10.1145/3306346.3323008.

Suwajanakorn, Supasorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz (2014). "Total Moving Face Reconstruction." In: *ECCV*.

Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman (2015). "What Makes Tom Hanks Look Like Tom Hanks." In: *ICCV*.

Tewari, Ayush, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt (2019). "FML: Face model learning from videos." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tewari, Ayush, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt (2020a). "PIE: Portrait Image Embedding for Semantic Control." In: *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*.

Tewari, Ayush, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt (2020b). "StyleRig: Rigging StyleGAN for 3D Control over Portrait Images." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tewari, Ayush, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer (2020c). "State of the Art on Neural Rendering." In: *Computer Graphics Forum (EG STAR)*.

Tewari, Ayush, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt (2018). "Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tewari, Ayush, Michael Zollhöfer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt (2020d). "High-Fidelity Monocular Face Reconstruction Based on an Unsupervised Model-Based Face Autoencoder." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2, pp. 357–370.

Tewari, Ayush, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian (2017). "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction." In: *The IEEE International Conference on Computer Vision (ICCV)*.

Thies, Justus, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner (2020). "Neural Voice Puppetry: Audio-driven Facial Reenactment." In: *European Conference on Computer Vision (ECCV)*.

Thies, Justus, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner (2016a). "Face2Face: Real-time Face Capture and Reenactment of RGB Videos." In: *CVPR*.

Thies, Justus, Michael Zollhöfer, and Matthias Nießner (2019). "Deferred neural rendering: Image synthesis using neural textures." In: *ACM Transactions on Graphics (TOG)* 38.4, pp. 1–12.

Thies, Justus, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt (2015). "Real-time Expression Transfer for Facial Reenactment." In: *ACM Trans. Graph.* 34.6, 183:1–183:14.

Thies, Justus, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner (2016b). "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos." In: *CVPR*.

Tran, Luan, Feng Liu, and Xiaoming Liu (June 2019). "Towards High-fidelity Nonlinear 3D Face Morphable Model." In: *CVPR*.

Tran, Luan and Xiaoming Liu (June 2018a). "Nonlinear 3D Face Morphable Model." In: *In Proceeding of IEEE Computer Vision and Pattern Recognition*. Salt Lake City, UT.

Tran, Luan and Xiaoming Liu (2018b). "On Learning 3D Face Morphable Model from In-the-wild Images." arXiv:1808.09560.

Tretschk, Edgar, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt (2020a). "PatchNets: Patch-Based Generalizable Deep Implicit 3D Shape Representations." In: *European Conference on Computer Vision (ECCV)*.

Tretschk, Edgar, Ayush Tewari, Michael Zollhöfer, Vladislav Golyanik, and Christian Theobalt (2020b). "DEMEA: Deep Mesh Autoencoders for Non-Rigidly Deforming Objects." In: *European Conference on Computer Vision (ECCV)*.

Trigeorgis, George, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou (July 2017). "Face Normals "In-The-Wild" Using Fully Convolutional Networks." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tuan Tran, Anh, Tal Hassner, Iacopo Masi, and Gerard Medioni (July 2017). "Regressing Robust and Discriminative 3D Morphable Models With a Very Deep Neural Network." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tulsiani, Shubham, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik (2017). "Multi-view Supervision for Single-View Reconstruction via Differentiable Ray Consistency." In: *CVPR*. IEEE Computer Society, pp. 209–217.

Tylecek, Radim and R Sara (2010). "Refinement of surface mesh for accurate multiview reconstruction." In: *International Journal of Virtual Reality* 9.1, pp. 45–54. ISSN: 1081-1451.

Tzimiropoulos, Georgios (June 2015). "Project-Out Cascaded Regression With an Application to Face Alignment." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ubelaker, Douglas H (2015). "Craniofacial superimposition: historical review and current issues." In: *Journal of forensic sciences* 60.6, pp. 1412–1419.

Valgaerts, Levi, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt (Nov. 2012). "Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting." In: *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012)*. Vol. 31. 6, 187:1–187:11.

Vinyals, Oriol, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. (2016). "Matching networks for one shot learning." In: *Advances in Neural Information Processing Systems*, pp. 3630–3638.

Vlasic, Daniel, Matthew Brand, Hanspeter Pfister, and Jovan Popović (July 2005). "Face Transfer with Multilinear Models." In: *ACM Trans. Graph.* 24.3, pp. 426–433. ISSN: 0730-0301. DOI: 10.1145/1073204.1073209. URL: http://doi.acm.org/10.1145/1073204.1073209.

Vu, H. H., P. Labatut, J. P. Pons, and R. Keriven (May 2012). "High Accuracy and Visibility-Consistent Dense Multiview Stereo." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.5, pp. 889–901. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.172.

Wang, Nannan, Xinbo Gao, Dacheng Tao, and Xuelong Li (Oct. 2014). "Facial Feature Point Detection: A Comprehensive Survey." arXiv:1410.1037.

Wang, Sheng-Yu, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros (2019a). "Detecting photoshopped faces by scripting photoshop." In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10072–10081.

Wang, Sheng-Yu, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros (2020). "CNN-generated images are surprisingly easy to spot...for now." In: *CVPR*.

Wang, Ting-Chun, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro (2019b). "Few-shot Video-to-Video Synthesis." In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro (2019c). "Video-to-Video Synthesis." In: *Proc. NeurIPS*.

Wiles, O., A.S. Koepke, and A. Zisserman (2018). "X2Face: A network for controlling face generation by using images, audio, and pose codes." In: *European Conference on Computer Vision (ECCV)*, pp. 690–706.

Wu, Chenglei, Derek Bradley, Markus Gross, and Thabo Beeler (2016). "An anatomically-constrained local deformation model for monocular face capture." In: *ACM transactions on graphics (TOG)* 35.4, pp. 1–12.

Wu, Chenglei, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt (2011). "High-quality shape from multi-view stereo and shading under general illumination." In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, pp. 969–976.

Yan, Cai-Ping and Chi-Man Pun (2017). "Multi-scale difference map fusion for tamper localization using binary ranking hashing." In: *IEEE Transactions on Information Forensics and Security* 12.9, pp. 2144–2158.

Yenamandra, Tarun, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt (2021). "i3DMM: Deep Implicit 3D Morphable Model of Human Heads." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yin, Lijun, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato (2006). "A 3D facial expression database for facial behavior research." In: *International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 211–216.

Yu, Ye and William A. P. Smith (2019). "InverseRenderNet: Learning Single Image Inverse Rendering." In: *CVPR*.

Zakharov, Egor, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky (2019). "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models." In: *CoRR* abs/1905.08233. arXiv: 1905.08233. URL: http://arxiv.org/abs/1905.08233.

Zeiler, Matthew D. (2012). "ADADELTA: An Adaptive Learning Rate Method." In: *arXiv preprint arXiv:1212.5701*.

Zhao, Fang, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan (2016). "Robust LSTM-Autoencoders for Face De-Occlusion in the Wild." arXiv:1612.08534.

Zhou, Hao, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs (Oct. 2019). "Deep Single-Image Portrait Relighting." In: *The IEEE International Conference on Computer Vision (ICCV)*.

Zhu, Jiapeng, Yujun Shen, Deli Zhao, and Bolei Zhou (2020). "In-domain GAN Inversion for Real Image Editing." In: *Proceedings of European Conference on Computer Vision (ECCV)*.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks." In: *ICCV*.

Zhu, Xiangyu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li (2015). "High-fidelity Pose and Expression Normalization for face recognition in the wild." In: *CVPR*. IEEE Computer Society, pp. 787–796.

Zollhöfer, Michael, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt (2018). "State

of the art on monocular 3D face reconstruction, tracking, and applications." In: *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library, pp. 523–550.

nVidia (Oct. 2012). *CUBLAS Library User Guide*. v5.0. nVidia. URL: http://docs.nvidia.com/cublas/index.html.