3D Morphable Face Models - Past, Present and Future

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al.

3D Morphable Face Models - Past, Present and Future

BERNHARD EGGER, Massachusetts Institute of Technology, USA
WILLIAM A. P. SMITH, University of York, UK
AYUSH TEWARI, Max Planck Institute for Informatics & Saarland Informatics Campus, Germany
STEFANIE WUHRER, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
MICHAEL ZOLLHOEFER, Stanford University, USA
THABO BEELER, Disney Research|Studios, Switzerland
FLORIAN BERNARD, Max Planck Institute for Informatics & Saarland Informatics Campus, Germany
TIMO BOLKART, Max Planck Institute for Intelligent Systems, Germany
ADAM KORTYLEWSKI, Johns Hopkins University, USA
SAMI ROMDHANI, IDEMIA, France
CHRISTIAN THEOBALT, Max Planck Institute for Informatics & Saarland Informatics Campus, Germany
VOLKER BLANZ, University of Siegen, Germany
THOMAS VETTER, University of Basel, Switzerland

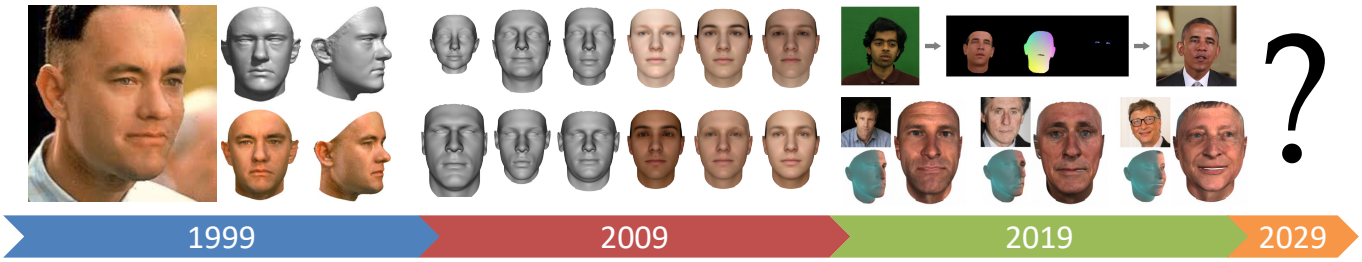


Fig. 1. 20 years of 3D Morphable Models. Fitting results from the original paper [Blanz and Vetter 1999], the first publicly available Morphable Model [Paysan et al. 2009a], and state-of-the-art facial re-enactment results [Kim et al. 2018a] and GAN-based models [Gecer et al. 2019b].

In this paper, we provide a detailed survey of 3D Morphable Face Models over the 20 years since they were first proposed. The challenges in building and applying these models, namely capture, modeling, image formation, and image analysis, are still active research topics, and we review the state-of-the-art in each of these areas. We also look ahead, identifying unsolved challenges, proposing directions for future research and highlighting the broad range of current and future applications.

Keywords: 3D Computer Vision, Computer Graphics, Statistical Modelling, Analysis-by-Synthesis, Generative Models

*Institute of Engineering Univ. Grenoble Alpes.

Authors' addresses: Bernhard Egger, Massachusetts Institute of Technology, USA, egger@mit.edu; William A. P. Smith, University of York, UK, william.smith@york.ac.uk; Ayush Tewari, Max Planck Institute for Informatics & Saarland Informatics Campus, Germany, atewari@mpi-inf.mpg.de; Stefanie Wuhrer, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France, stefanie.wuhrer@inria.fr; Michael Zollhoefer, Stanford University, USA, michael@zollhoefer.com; Thabo Beeler, Disney Research|Studios, Switzerland, thabo.beeler@disneyresearch.com; Florian Bernard, Max Planck Institute for Informatics & Saarland Informatics Campus, Germany, fbernard@mpi-inf.mpg.de; Timo Bolkart, Max Planck Institute for Intelligent Systems, Germany, tbolkart@tuebingen.mpg.de; Adam Kortylewski, Johns Hopkins University, USA, akortyl1@jhu.edu; Sami Romdhani, IDEMIA, France, sami.romdhani@idemia.com; Christian Theobalt, Max Planck Institute for Informatics & Saarland Informatics Campus, Germany, theobalt@mpi-inf.mpg.de; Volker Blanz, University of Siegen, Germany, blanz@informatik.uni-siegen.de; Thomas Vetter, University of Basel, Switzerland, thomas.vetter@unibas.ch.

1 INTRODUCTION

It is 20 years since 3D Morphable Face Models were first presented at SIGGRAPH '99. They were proposed as a general face representation and a principled approach to image analysis. Blanz and Vetter [1999] introduced and tackled many subsidiary problems and the results were considered groundbreaking. The impact of the original paper has been long term, recognized by an impact paper award, and the approach and applications are accessible to a wide audience (the original supplementary video was one of the most popular videos in the early days of YouTube). However, the approach is not just of historical interest. In the past two years, 3D Morphable Face Models have been re-discovered in the context of deep learning and are incorporated into many state-of-the-art solutions for face analysis. This survey aims to build a starting point for researchers new to the topic, act as a reference guide for the community around 3D Morphable Models and to introduce exciting open research questions.

1.1 Definition

A 3D Morphable Face Model is a generative model for face shape and appearance that is based on two key ideas: First, all faces are in dense point-to-point correspondence, which is usually established on a set of example faces in a registration procedure and then maintained throughout any further processing steps. Due to this

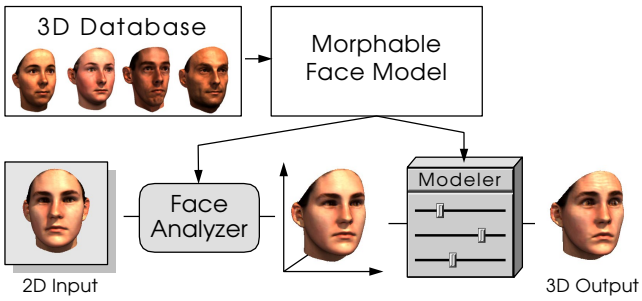


Fig. 2. The visual abstract of the seminal work by Blanz and Vetter [1999]. It proposes a statistical model for faces to perform 3D reconstruction from 2D images and a parametric face space which enables controlled manipulation.

correspondence, linear combinations of faces may be defined in a meaningful way, producing morphologically realistic faces (*morphs*). The second idea is to separate facial shape and color and to disentangle these from external factors such as illumination and camera parameters. The Morphable Model may involve a statistical model of the distribution of faces, which was a principal component analysis in the original work [Blanz and Vetter 1999] and has included other learning techniques in subsequent work.

1.2 History

The initial research question behind the idea of 3D Morphable Models (3DMM) was how a visual system, biological or artificial, can cope with the high variety of images that a single class of objects can generate, and how objects are represented to solve vision tasks. The leading assumption for the development of 3DMMs was that prior knowledge about object classes plays an important role in vision and helps to solve otherwise ill-posed problems. 3DMMs are designed to capture such prior knowledge, and they are learned automatically from a set of examples. The representation is general, so it may be applied to different objects and tasks.

Representations of faces and the task of face recognition have been in the focus of vision research for a long time. An important and very influential paradigm shift in this field was the Eigenfaces approach by Sirovich and Kirby [1987] and Turk and Pentland [1991], which learned an explicit face representation from examples and operated entirely on grey-levels in the image domain. Eigenfaces treated images of faces as a vector space and performed a principal component analysis, with the eigenvectors representing the main modes of variation in that space. The drawback of Eigenfaces was not only that it was limited to a fixed pose and illumination, but that it had no effective representation of shape differences: When the coefficients in linear combinations of eigenvectors are changed continuously, structures will fade in and out, rather than shift along the image plane. As a consequence, the model fails to find a single parameter for, say, the distances between the eyes. The Eigenfaces approach was also extended to 3D face surfaces by Atick et al. [1996] to model shading variations in faces, yet with essentially the same limitation.

Several research groups proceeded by adding an Eigendecomposition of 2D shape variations between individual faces. This provided both an explicit shape model, and - after warping the images - an aligned Eigenface model without blurring and ghosting artifacts. While in the original Eigenface approach, the images were only aligned by a single point (e.g., the tip of the nose), the new methods established correspondence on significantly more points. Landmark-based face warping for image analysis was introduced by Craw and Cameron [1991]. Using approximately 200 landmarks, the first statistical shape model was proposed in Active Shape Models [Cootes et al. 1995]. While this model used shape only, Active Appearance Models [Cootes et al. 1998] proposed a combination of shape and appearance that turned out to be very successful and influential. Other groups computed dense pixel-wise image correspondences with optic-flow algorithms for modeling the facial shape variations [Hallinan et al. 1999; Jones and Poggio 1998]. In all these correspondence-based approaches, images are warped to a common template, and the appearance variation is then performed in the same way as the original Eigenfaces, but on the shape-normalized images. The shape model, on the other hand, provides a powerful and compact representation of shape differences by shifting pixels in the image plane. However, compared to the simple linear projection in Eigenfaces, the image analysis task is transformed into a more challenging nonlinear model fitting problem.

These 2D models were efficient to cover the shape variation for a fixed pose and illumination setting. The framework was extended to variations across pose by Vetter and Poggio [1997] and to other object classes, such as images of cars [Jones and Poggio 1998]. All this groundwork demonstrated that a separation of shape and texture information in images can model the variation of faces. On the other hand, the price to pay for taking pose and illumination variations into account was high: eventually, it would require many separate models, each limited to a small range of poses and illuminations. In contrast, the progress of 3D Computer Graphics in the 1990s demonstrated that variations in pose and illumination are easy to simulate, including self-occlusion and shadowing. Adapting methods from graphics to face modeling and computer vision led to the new face representation in 3DMMs and the idea of using analysis-by-synthesis to map between the 3D and 2D domain. Those were the two key contributions in the first paper on 3DMMs [Blanz and Vetter 1999], compare Figure 2. The name Morphable Model was derived from their 2D counterpart [Jones and Poggio 1998], and in fact, Jones and Poggio strongly influenced the ideas that led to 3DMMs.

3DMMs and 2D Morphable Models rely on dense correspondence, rather than only a set of facial feature points. In the original work, this was established by an optical flow algorithm for image registration. The image synthesis algorithm used a standard rendering model with perspective projection, ambient and directional lighting, and a Phong model of surface reflectance that includes a specular component. However, in analysis-by-synthesis, this approach comes at a computational price because shape-camera [Smith 2016] and illumination-albedo [Egger 2018] ambiguities lead to a hard ill-posed optimization problem. Moreover, the optimization is costly and is prone to end in unwanted local optima. Just as it is already dramatically more complicated to fit an Active Appearance Model

to a 2D image, compared to the simple projection needed for Eigenfaces, the complexity of 3DMM fitting raises additional problems which have remained challenging to researchers after 20 years of development.

At the time the initial 3DMM was developed, image-based models were dominating computer vision and even animation [Ezzat et al. 2002], and they were rather elaborate at that time. It was a key decision to take the best of both the 2D and the 3D world, by using 3D models to manipulate existing images on the one hand, and applying 2D algorithms to 3D surfaces: Unlike mesh-based algorithms, the original 3DMM used optical flow, multi-resolution approaches and interpolation algorithms on parameterized surfaces of faces. With the initial face scanner delivering surfaces in a two-dimensional cylinder parameterization, all those steps were performed in 2D, and most of the methods involved were replaced with their 3D equivalent only many years later. It is interesting to see that after a development towards 3D, the computer vision community came back to 2D representations by using deep learning, and now evolves again to 3D, e.g., by integrating 3DMMs.

Over the past years, 3DMMs were applied beyond faces. Models were built for the surface of the human body [Allen et al. 2003; Anguelov et al. 2005; Loper et al. 2015] and for other specific parts of the body like ears [Dai et al. 2018] and hands [Khamis et al. 2015], animals [Sun and Murata 2020; Zuffi et al. 2018] and even cars [Shelton 2000]. In this survey, we focus on 3DMMs to model the human face, though many of the techniques and challenges are the same across different object classes.

The 3DMM was developed in a time where algorithms and data were rarely shared across researchers and institutions. 10 years later the first publicly available 3DMM was released [Paysan et al. 2009a] and in the last 10 years, all individual data and algorithmic components needed to build and use 3DMMs were released by various researchers. We collected a list of all available resources and will further maintain it [Community 2019].

The 3DMM was built as a general representation for faces, not just aiming at one specific task. Even though the model is outperformed for some very specific applications such as face recognition, it is unique in its generality across different tasks and applications.

1.3 Organization

There is a recent state of the art report on monocular 3D reconstruction, tracking and applications [Zollhoefer et al. 2018]. This focuses on the most recent advances, particularly related to the specific task of tracking and reconstruction. In contrast, in this paper, we instead focus on the 3DMM, all involved methods, and reflect the major contributions over the past 20 years while at the same time highlighting challenges and future directions.

This survey is organized from building to applying a 3DMM. We start with Section 2 where we present methods to acquire 3D facial data for model building. We then describe in Section 3 the various approaches to model the 3D shape and facial appearance. In Section 4 we discuss the methods to generate a 2D image from our 3D model using computer graphics. Our Section 5 surveys the major application of 3DMMs, namely the reconstruction of a 3D face from a 2D image. Section 6 summarizes the impact of 3DMMs in the recent advances in the field of deep learning and how deep learning can be used to improve the modeling and analysis. Section 7 summarizes the various applications where 3DMMs were used in the past 20 years. Every Section summarizes the major challenges the authors see regarding the current limitations of 3DMM. We also collect challenges that are shared across multiple Sections in Section 8, where we also venture an outlook on what we expect to see in the next 10 to 20 years and how the 3DMM will keep impacting how faces are represented.

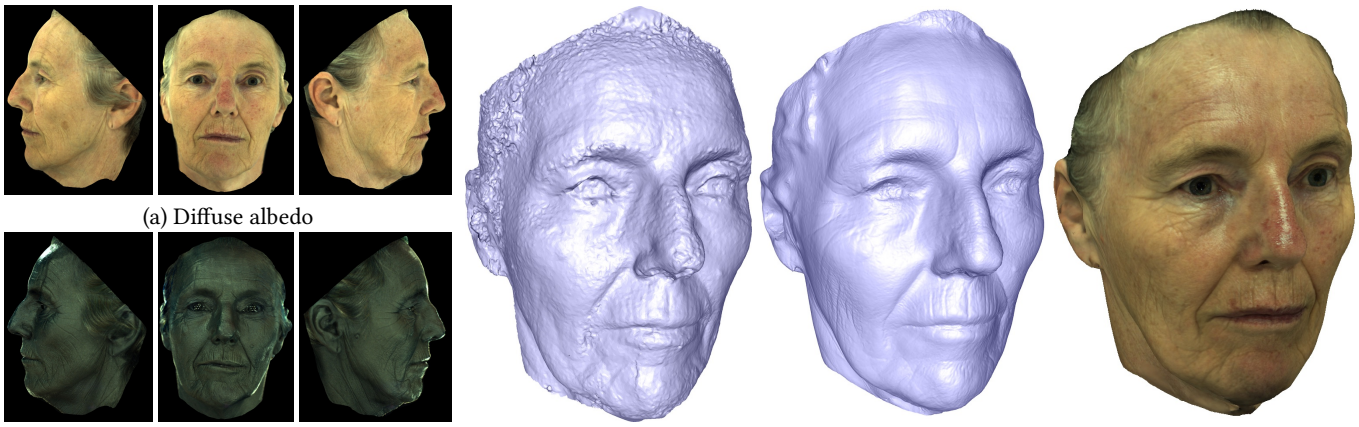
2 FACE CAPTURE

The key ingredient to any 3DMM is a representative set of 3D shapes, usually coupled with corresponding appearance data. The typical way to construct such a sample pool is by acquiring data from the real world. In this section, we give a brief overview of different approaches that have been used to acquire facial data as well as data of facial parts. As we are concerned with the creation of input datasets for 3DMMs, we limit the discussion to acquisition under controlled conditions, as opposed to the more challenging in-the-wild setting. Note that controlled 3D face capture may not always be necessary. There have been attempts to learn 3DMMs directly from images [Cashman and Fitzgibbon 2012] and state-of-the-art deep learning-based methods simultaneously learn a 3DMM and regression-based fitting from 2D training data (see Section 6.3). In this section we begin by covering shape acquisition methods in Section 2.1 including geometric, photometric and hybrid methods. Sections 2.2, 2.3 and 2.4 describe methods for capture of appearance, face parts and dynamics respectively. Section 2.5 lists publicly available 3D face datasets that could be exploited for building 3DMMs. Finally, we consider open challenges related to face capture in Section 2.6.

2.1 Shape Acquisition

The three-dimensional shape is arguably the most important ingredient to a 3DMM. The issue of shape representation has not been widely considered in the context of 3DMMs. By far the most commonly used representation is a triangle mesh. Rare exceptions include cylindrical [Atick et al. 1996] and orthographic [Dovgard and Basri 2004] depth maps (though these representations do not permit meaningful dense correspondence), per-vertex surface normals [Aldrian and Smith 2012], and, more recently, volumetric orientation fields [Saito et al. 2018] and signed distance functions [Park et al. 2019]. Using a triangle mesh representation, dense correspondence requires that all samples exhibit the same topology and that the vertices encode the same semantic point on all samples. Establishing correspondence across the samples is a challenging topic in itself, discussed in Section 3.5. In this section, we focus on the acquisition of raw 3D data, before establishing correspondence.

2.1.1 Geometric methods. Geometric methods estimate directly the 3D coordinates of a shape either by observing the same surface point from two or more viewpoints (in which case the challenge is identifying corresponding points between images) or by observing a projected pattern (in which case the challenge is identifying the correspondence between the known pattern and an image of its projection). Methods can either be considered active, i.e. they emit light or other signals into the scene, or passive. Laser scanners,



(a) Diffuse albedo (b) Specular albedo (c) MVS geometry (d) Hybrid geometry (e) Rendering

Fig. 3. Capture of intrinsic face properties using a hybrid geometric/photometric method [Seck et al. 2016; Smith et al. 2020]. Multi view stereo (MVS) is used to reconstruct a coarse mesh (c). A photometric light stage [Ma et al. 2007] is used to capture diffuse and specular albedo maps (a,b) and surface normals that are merged with the MVS mesh to produce a mesh with fine surface detail (d). Together, these can be used to synthesize highly realistic images of the face (e).

Time-of-Flight sensors, and Structured Light systems are active systems, where multi-view photogrammetry is a passive alternative. Active multi-view photogrammetry may be considered a hybrid active/passive approach, as it relies on passive photogrammetry to reconstruct the shape, but augments the object with a well-defined texture projection that benefits the reconstruction [Zhang et al. 2004]. Unlike structured light, the origin of the light does not matter as the projected texture is solely meant to augment the texture used for multi-view stereo matching. This type of technology is used by the Intel® RealSense™ D435 camera for example[1]. In the early days of 3DMMs, active systems were the only real option to acquire 3D shapes at a reasonable quality. The original paper of Blanz and Vetter [1999] relied on laser scanning [Levoy et al. 2000], where the face is rasterized via one or more laser-beams. The laser beam illuminates the face surface at a point and using the known camera/laser arrangement the 3D position of this point may be triangulated. The biggest drawback of laser scanners is the acquisition time, as only very few samples are gathered at any given time – even at very high frame rates, such systems require the subjects to sit still for several seconds.

Structured light scanners [Geng 2011] overcome this limitation to some extent by injecting not only a few beams but leveraging projectors that offer millions of them. The challenge here is to identify which beam is illuminating the object at a given point. This is addressed by structuring the projected light in a way that allows to clearly identify the origin of any ray. The simplest approach is binary encoding, which projects black and white patterns assigning a unique binary code to each pixel. The required number of patterns is still quite substantial, for VGA resolution one needs 19 distinct patterns and for 4K resolution 23 patterns, and hence this approach is most suited for capturing static objects. However, technical improvements have begun to make these approaches viable for dynamic capture of faces. The Intel® RealSense™ SR300 uses only 9 binary patterns to obtain VGA, while the most recent RealSense depth camera produces VGA resolution at 60 depth FPS with a scanning laser technology. Other more complex structured light methods have been proposed, such as gray codes or (colored) fringe patterns, which can reduce the number of required frames further, in extreme cases even to a single frame. A very popular commercial system that was used to create face datasets ([Cao et al. 2014b]) and that employs structured light is the first generation Kinect sensor[2]. The device employs a structured dot pattern, which allows reconstructing depth from a single frame by sacrificing spatial resolution. Resolution may be improved by accumulating several frames [Newcombe et al. 2011]. With the increased resolution and quality of consumer cameras, passive systems have become the method of choice in most cases, since they are simpler to assemble and operate and off-the-shelf photogrammetry software solutions, both commercial such as Agisoft[3] or RealityCapture[4], as well as open-source solutions such as Meshroom[5], provide very good results on human faces. Also, complete systems can be purchased that come with both hard- and software[6][7][8]. These methods typically do not require the aggregation of information over time and hence offer themselves for single-shot acquisition [Beeler et al. 2010] as well as full-frame rate performance capture [Beeler et al. 2011; Bradley et al. 2010; Furukawa and Ponce 2009]. A potential disadvantage of the aforementioned systems is their form factor since they all require at least some separation between the different participating components, i.e. the cameras or lights, often referred to as the baseline. An alternative which becomes more and more viable due to the push of the mobile industry are time-of-flight sensors, where the elements can be located close to each other. The second-generation Kinect

[1] https://www.intelrealsense.com/depth-camera-d435/

[2] https://en.wikipedia.org/wiki/Kinect#Kinect_for_Xbox_360_(2010)
[3] https://www.agisoft.com
[4] https://www.capturingreality.com
[5] https://alicevision.org/
[6] https://www.canfieldsci.com/imaging-systems/vectra-m3-3d-imaging-system/
[7] http://www.di4d.com
[8] http://www.3dmd.com/

sensor[9] belongs to this family, as well as many depth sensors that are shipped with modern mobile phones. A challenge that time-of-flight sensors share with most of the active systems is that color information has to be acquired separately and is not intrinsically aligned with the 3D data, which is another advantage of passive setups.

2.1.2 Photometric methods. Photometric methods typically estimate surface orientation, from which the 3D shape may be recovered via integration. The challenge here is to select models that accurately capture reflectance properties of the surface and obtaining sufficient measurements that the inversion of these models is well-posed. Compared to geometric methods, photometric methods typically offer higher shape detail and do not rely on the presence of matchable features (so are applicable to smooth, featureless surfaces) but often suffer from low-frequency bias in the reconstructed positions caused by modeling errors in reflectance and illumination. Photometric stereo [Ackermann et al. 2015] estimates the surface normal at each pixel by observing a scene from a fixed position under at least three different illumination conditions, which can be spectrally multiplexed [Hernández et al. 2007] in order to reduce the number of frames required. Early work assumed known lighting directions and perfectly diffuse reflectance. When illumination is uncalibrated and a more suitable glossy reflectance model is used, generic face priors can be used to resolve the resulting ambiguity [Georghiades 2003]. Typically more lighting conditions are used to increase robustness and coverage, such as four [Zafeiriou et al. 2013] or even nine [Gotardo et al. 2015]. Gradient-based illumination takes the number of conditions to the extreme, by illuminating the subject not with discrete individual point lights, but by an ideally continuous, omnidirectional incident illumination gradient. An advantage of this set up is that hard light source occlusions (cast shadows) are replaced by soft partial occlusions of the illuminating hemisphere (ambient occlusion). In practice, the omnidirectional illumination is realized via a light-stage [Debevec et al. 2000], which discretizes the gradient with a large number (several hundred) of light sources. The original work of Ma et al. [2007] suggests the use of four distinct gradients, which has later been extended using complementary gradients [Wilson et al. 2010]. Again, variants of temporal, spectral and polarization multiplexing have been proposed to reduce the number of required conditions.

2.1.3 Hybrid methods. Hybrid methods combine the strength of geometric and photometric methods, specifically, they reduce the low-frequency bias typically present in photometric methods and increase the high-frequency details when compared to geometric methods. Nehab et al. [2005] propose a method for merging the low frequencies of positional information and the high frequencies of surface normals. The method is particularly efficient, involving only the solution of a sparse linear system of equations, and has been used in the context of 3DMM fitting [Patel and Smith 2012]. Various combinations of geometric and photometric methods have been considered. For example, Zivanov et al. [2009] combine structured light with photometric stereo, Ma et al. [2007] combine structured light with gradient-based illumination, Ghosh et al. [2011] combine multi-view stereo with gradient-based illumination, and Beeler et al. [2010] combine passive multi-view photogrammetry with shape-from-shading. Figure 3(d) shows the output of a hybrid method in which photometric surface normals are merged with a multi-view stereo mesh.

[9]https://en.wikipedia.org/wiki/Kinect#Kinect_for_Xbox_One_(2013)

2.2 Appearance Capture

In addition to shape, appearance is also required for many 3DMM tasks, such as synthesizing images (see Section 4) and inverse rendering (see Section 5). Unlike shapes, which are almost exclusively represented as triangular meshes, appearance representation varies substantially. While in theory, every vertex of the mesh could have an associated appearance property, typically shapes are parameterized to the 2D domain and textures are used to store appearance properties. Appearance can be as simple as backprojecting the color of the images onto the shapes, which causes shading effects to be baked in. Self-occlusion, in particular when only a single viewpoint is available, results in missing data in the occluded areas, which must be hallucinated somehow. Booth et al. [2018b] use 3DMM fits to in-the-wild images and Principal Component Pursuit with missing values to complete the unobserved texture. They build their appearance model directly on the sampled textures. Such a simplistic approach, however, does not allow intrinsic face appearance properties to be separated from shading/shadowing (and hence illumination/geometry). A partial solution to this problem is to control illumination conditions during capture, for example by using multiple light sources to create approximately ambient lighting. Note that a truly Lambertian convex surface observed under truly ambient light gives exactly the albedo [Lee et al. 2005]. The appearance models in the most popular 3DMMs [Booth et al. 2018a; Dai et al. 2017; Paysan et al. 2009a] use this approach, combining images from multiple cameras to provide full coverage of the face with diffuse lighting to approximate albedo. A better approach is to explicitly separate shading from skin color, often referred to as intrinsic decomposition. This allows relighting of the face under novel incident illumination conditions and a 3DMM built on such data truly models intrinsic characteristics of the face. Several approaches have been presented over the years to acquire reflectance data suited for parametric rendering, measuring surface reflectance [Marschner et al. 1999] and even subsurface scattering properties [Ghosh et al. 2008]. The polarised spherical illumination environment used by Ma et al. [2007] enables diffuse albedo to be captured in a single shot and specular albedo in two images (see Figure 3(a) and (b)). While such approaches have predominately used active setups, recently capture under passive conditions has been demonstrated [Gotardo et al. 2018].

2.3 Face part specific methods

Certain parts of the human face require more targeted acquisition methods and devices since they do not conform with the assumptions typically made by abovementioned approaches. For example, the frontmost part of the eye, the cornea, is for obvious reasons fully transparent and distorts the appearance of the underlying iris due to refraction. Bérard et al. [2014] leverage a combination of several specialized algorithms, including shape-from-specularity,

in order to reconstruct all visible components of the eye. Another challenging example are teeth [Wu et al. 2016a], which exhibit extremely challenging appearance [Velinov et al. 2018]. Hair violates the common assumption that the reconstructed shape is a smooth continuous surface, and requires specialized approaches that estimate hair fibers [Beeler et al. 2012], hair strands [Hu et al. 2014a; Luo et al. 2013] and braiding [Hu et al. 2014b], or even encode hair as a surface [Echevarria et al. 2014] for manufacturing. While most hair acquisition focuses on static reconstruction, some do capture hair in motion [Xu et al. 2014] or estimate physical properties for hair simulation [Hu et al. 2017a]. Especially challenging is the acquisition of partially or completely hidden properties, such a the tongue [Hewer et al. 2018], the skull [Achenbach et al. 2018; Beeler and Bradley 2014], or the jaw [Zoss et al. 2019, 2018], where oftentimes specialized imaging systems are required, such as Computer Tomography (CT), Magnetic Resonance Imaging (MRI), or Electromagnetic Articulography (EMA). Lastly, even skin itself requires specialized treatment in some areas, such as lips [Garrido et al. 2016b] or eyelids [Bermano et al. 2015], where the local appearance and deformation exceed the capabilities of the more generic methods.

2.4 Dynamic capture

Historically, 3DMMs have been mostly concerned with static shapes, for example with a set of neutral shapes from different individuals or with a discrete set of expressions per individual, neglecting how the face transitions between expressions. Most capture systems used to build 3DMMs were hence static systems, focused on capturing individual shapes rather than full performances. As the field begins to integrate more temporal information into the models, the need for dynamic capture systems will rise. Active systems have been considered, both geometric [Zhang et al. 2004] and photometric [Wilson et al. 2010]. However, passive systems [Beeler et al. 2011; Bradley et al. 2010] are currently the technologies of choice, since they do not require temporal multiplexing and still deliver high-quality shapes, and more recently even per-frame reflectance data [Gotardo et al. 2018]. A beneficial side-effect of such technologies is that they often provide shapes that are already in correspondence, removing the need to establish correspondence in a post-processing step (Section 3.5), and making them attractive solutions even when only a discrete set of shapes is desired. Available commercial solutions include Di4D[10], 3dMD[11], or the Medusa system[12].

2.5 Publicly available face datasets

A relatively large number of publicly available datasets exist that could be leveraged in the construction of 3DMMs, though many have never been used for this purpose. We believe there is not broad awareness of the range of 3D datasets available and so collect them together in Table 1. We hope that this will encourage work that seeks to exploit multiple datasets for 3DMM building.

2.6 Open challenges

The field of face capture is far ahead of face modeling in general and 3DMMs in particular. There is a large gap between the quality of data that can be captured and the data actually used to build 3DMMs. There is a further gap between the quality of this already-deficient data and what a 3DMM is able to synthesize (see Section 3). Hence, from the perspective of 3DMMs, the open challenges in capture do not generally relate to improving the acquisition quality, but to the lack of publicly available data. While there is a decent number of datasets publicly available (see Section 2.5), most of these contain only moderate quality shape data and no appearance information, with the exception of [Stratou et al. 2011], which consists of 23 identities only. We believe that the lack of high-quality datasets is due to a variety of reasons. On the one hand, high-quality acquisition devices that can capture both shape and appearance are not readily available. Most of them are custom-built, cannot easily be purchased or licensed, and require expert knowledge for operation. On the other hand, acquiring and processing data may be a time and resource-intense effort, since many systems in the research community were not conceived for scalable deployment but for experimental use; slow capture methods are not applicable to young or elderly people, expensive setups are challenging to replicate on a global scale to capture whole populations, and methods requiring very bright illumination makes it unpleasant to be captured with eyes open. Furthermore, most high-quality systems, in particular ones that also measure appearance, generally require controlled lab conditions which makes it difficult to capture large numbers of the general public. Advances in face capture may alleviate some of these issues.

Additionally, there are many important broader questions related to data acquisition that remain unanswered. How many faces do we really need to capture in order to build a representative (universal) model? How can we ensure we capture natural expressions? Most people are not trained to perform specific expressions (i.e. FACS[13]), and will have difficulties performing naturally when put in a capture setup, leading to a biased dataset. How should we deal with bias in general and what is the right sampling strategy with respect to age, gender, ethnicity and so on? Are the capture methods themselves biased? For example, capturing faces with very dark skin is challenging for both photometric and geometric methods. Should we accept that we cannot hope to capture sufficiently broad data and therefore rely on synthesizing additional data or using captured data to build a bootstrap model that is refined on large 2D datasets? These approaches are discussed in Section 6.

Finally, there are some philosophical and ethical issues to consider. The human face is unique and highly personal. Once a face has been captured in high detail, it is possible to synthesize new images that are almost indistinguishable from photos. If captured datasets are made publicly available, it is very difficult to control the distribution and use of such data. Obtaining proper informed consent is, therefore, both legally and ethically important but perhaps even this does not go far enough, particularly when consent for minors is given by parents. These issues are beyond the expertise

[10] http://www.di4d.com/
[11] http://www.3dmd.com/
[12] https://studios.disneyresearch.com/medusa/
[13] https://en.wikipedia.org/wiki/Facial_Action_Coding_System

dataset	format and resolution	coverage	no. samples	scanner
Spacetime faces [Zhang et al. 2004]	triangle mesh (23k vertices, consistent topology)	inner face only	1 individual × 384 frame dynamic sequence	structured light
CASIA 3D Face Database [cas 2005]	640×480 depth map and texture image	face, neck, sometimes ears	123 individuals × 37-38 scans (expression, pose, illumination)	Minolta Vivid910
BU-3DFE [Yin et al. 2006]	triangle mesh (20k-35k triangles), two texture images (1,300 × 900)	face, neck, sometimes ears	100 individuals × 25 expressions	3dMD
BU-4DFE [Yin et al. 2008]	triangle mesh (35k vertices), texture image (1,040 × 1,329)	face, neck, sometimes ears	101 individuals × six 100 frame expression sequences	Dimensional Imaging
Bosphorus [Savran et al. 2008]	1, 600 × 1, 200 depth map and texture image	inner face only	105 individuals × up to 35 expressions per subject + 13 poses	Inspeck Mega Capturor II
York 3D Face Database [Heseltine et al. 2008]	depth map containing 5k-6k points, texture image	inner face only	350 individuals × 15 expressions	projected pattern stereo
B3D(AC)ˆ2 [Fanelli et al. 2010]	raw scan: triangle mesh (55k vertices), 780 × 580 texture image; processed: triangle mesh (23k vertices, consistent topology), 1,024 × 768 UV texture map	inner face only	14 individuals × around 80 dynamic sequences (speech-4D)	structured light stereo
Florence 3D Faces [Bagdanov et al. 2011]	triangle mesh (60k-80k triangles), 4 MPixel texture, additonal 2D HD video	face, neck, sometimes ears	53 individuals	3dMD
D3DFACS [Cosker et al. 2011]	triangle mesh (30k vertices), 1,024 × 1,280 UV texture map	face, neck, sometimes ears	10 individuals × around 52 dynamic sequences, FACS coded	3dMD
3DRFE [Stratou et al. 2011]	triangle mesh (1.2M vertices), 1,296 × 1,944 diffuse and specular albedo maps and hybrid normal maps	inner face, neck	23 individuals × 15 expressions	light stage
Hi4D-ADSIP [Matuszewski et al. 2012]	triangle mesh (20k vertices), texture image	inner face only	80 individuals × around 42 dynamic sequences	Dimensional Imaging
BP4D-Spontaneous [Zhang et al. 2014]	triangle mesh (30k-50k vertices), texture image (1,040 × 1,329)	face, neck, sometimes ears	41 individuals × eight one minute dynamic sequences	Dimensional Imaging
3D Dynamic Database for Unconstrained Face Recognition [Alashkar et al. 2014]	3.5k vertices for dynamic, 50k vertices for static, texture image	inner face only	58 individuals × one static scan + seven dynamic sequences	Artec
FaceWarehouse [Cao et al. 2014b]	raw: 640 × 480 RGBD; processed: triangle mesh (11k vertices, consistent topology)		150 individuals × 20 expressions	Microsoft Kinect
MMSE [Zhang et al. 2016a]	triangle mesh (30k-50k vertices), 1,040 × 1,392 texture image	inner face only	140 individuals × four dynamic sequences	Dimensional Imaging
Headspace [Dai et al. 2017]	triangle mesh (180k vertices), 2,973 × 3,055 UV texture map	full head including face, neck, ears	1,519 individuals	3dMD
4DFAB [Cheng et al. 2018]	triangle mesh (60k-75k vertices), UV texture map	face, neck and ears	180 individuals × 4k-16k frames of dynamic sequences	Dimensional Imaging
CoMA [Ranjan et al. 2018]	triangle mesh (80k-140k vertices), texture images (avg resolution 3, 700 × 3, 200), six raw camera images (each 1, 600 × 1, 200), alignments in FLAME topology	full head including face, neck, ears	12 individuals × 12 extreme expression sequences	3dMD
VOCASET [Cudeiro et al. 2019]	triangle mesh (80k-140k vertices), texture images (avg resolution 3, 700 × 3, 200), six raw camera images (each 1, 600 × 1, 200), alignments in FLAME topology	full head including face, neck, ears, speech	12 individuals × 40 dynamic sequences (speech-4D)	3dMD

Table 1. Overview of publicly available 3D shape and/or appearance scans of human faces.

of computer graphics and vision researchers and perhaps suggest a need for discussion and debate with other disciplines.

3 MODELING

This section outlines how to compute a 3DMM by modeling the variations of digitized 3D human faces. In particular, the following three types of variations are commonly considered. First, geometric variations across different identities are captured in a *shape model*, as outlined in Section 3.1. Commonly used models include global models, which represent variations of the entire face surface, and local models, which represent variations of facial parts. Second, geometric variations across different facial expressions are captured in an *expression model*, as outlined in Section 3.2. Commonly used models can be mainly classified into additive and multiplicative models. More recently, nonlinear expression models are starting to be explored. Third, variation in appearance and illumination are captured in a separate *appearance model* as outlined in Section 3.3.

It is interesting to note that the landmark paper on 3DMMs published 20 years ago [Blanz and Vetter 1999] proposed first models for all three types of variation that are still commonly used today.

To compute shape, expression, or appearance models, statistics are performed over a database of face data, where traditionally 3D scans of faces were used, and more recently some approaches also learn face models directly from 2D images, as outlined in Section 6.3. This computation of statistics requires correspondence information, that is, anatomically corresponding parts of the faces need to be compared, and hence known either explicitly or implicitly. An overview of how correspondence information is computed for faces is given in Section 3.5. The most commonly used approach is to compute correspondence information explicitly before computing the 3DMM. Some recent methods compute correspondence information at the same time while the 3DMM is built.

3DMMs are generative models and the ability to synthesize novel faces is a key feature and briefly discussed in Section 3.6. Finally, this section provides a list of available models and discusses open challenges on 3D face modeling in Sections 3.7 and 3.8, respectively.

3.1 Shape models

This section considers modeling geometric variation across different subjects computed using classical modelling approaches that use 3D data. To use a set of 3D scans as training data, we require a distance measure between any pair of scans, and computing a distance between raw scans consisting of different numbers of unstructured vertices is a complex problem. Most commonly, the community proceeds by first pre-processing the dataset by deforming a template mesh to all scans, which establishes anatomic correspondences between the points of the scans (see Section 3.5). We denote the surface of such a pre-processed mesh by $\mathcal{S}$ in the following. The i-th vertex of $\mathcal{S}$ is denoted by $\mathbf{v}_i \in \mathbb{R}^3$, and its associated vector $\mathbf{c} \in \mathbb{R}^{3n}$ contains the coordinates of $\mathbf{v}_i$ in a fixed order. All meshes share a common triangulation. We denote the ith triangle by $\mathbf{t}_i = (t_i^1, t_i^2, t_i^3) \in \{1, \dots, n\}^3$, where t_i^1, t_i^2, t_i^3 provide indices to the associated vertices $\mathbf{v}_{t_i^1}, \mathbf{v}_{t_i^2}, \mathbf{v}_{t_i^3}$, and we denote the complete triangulation by $\mathcal{T} = (\mathbf{t}_1, \dots, \mathbf{t}_m)$. Distances between shapes $\mathcal{S}_1$ and $\mathcal{S}_2$ are computed as difference between $\mathbf{c}_1$ and $\mathbf{c}_2$ after rigidly aligning $\mathcal{S}_1$ and $\mathcal{S}_2$ in $\mathbb{R}^3$.

3DMMs most often follow Dryden and Mardia [2002] for their definition of *shape* $\mathcal{S}$ as containing the geometric information remaining after having removed differences caused by translation, rotation, and sometimes uniform scaling. While scaling is typically not removed for human faces, this is often done in geometric morphometrics (e.g., [Dryden and Mardia 2002, Section 2]).

A *shape space* is traditionally defined as the set of all configurations of n vertices in $\mathbb{R}^3$ with fixed connectivity. Since we are interested in modeling human faces only in the context of 3DMMs, in the following, the term shape space refers to a d-dimensional parameter space (with $d \ll n$) that represents *plausible* 3D human faces. In this way, each 3D face has an associated parameter vector $\mathbf{w} \in \mathbb{R}^d$.

In 3DMMS, statistical shape analysis is used as generative model, i.e. the *shape space* has an associated probability distribution called prior that is defined by a density function $f(\mathbf{w})$ and that measures the likelihood that a realistic 3D face would be represented by a particular vector $\mathbf{w}$ in *shape space*. With a slight abuse of notation, we interpret $\mathbf{c}$ as a generator function in the following as

$$\mathbf{c} : \mathbb{R}^d \to \mathbb{R}^{3n} \tag{1}$$

that maps the low-dimensional parameter vector $\mathbf{w}$ to the vector of all vertex coordinates $\mathbf{c}(\mathbf{w}) \in \mathbb{R}^{3n}$. We again use $\mathbf{v}_i(\mathbf{w}) \in \mathbb{R}^3$ to refer to the ith vertex of the mesh given by $\mathbf{w}$. While the resolution (number of vertices) of the model is usually fixed, a progressive mesh representation based on edge collapse simplification of the generator function has been considered [Patel and Smith 2011].

This part considers the case where all faces in the training data have a similar (typically neutral) expression; generator functions that additionally model varying expressions are discussed in Section 3.2. As in [Brunton et al. 2014b], our discussions distinguishes *global models* that model the entire face or head area from *local models* that perform statistics over localized areas.

3.1.1 Global models. Let $\{\mathcal{S}_i\}_i$ denote the training shapes and $\{\mathbf{c}_i\}_i$ their associated coordinate vectors. The seminal work on 3DMMs [Blanz and Vetter 1999] proposed a global shape model that uses principal component analysis (PCA) to compute the linear generator function as

$$\mathbf{c}(\mathbf{w}) = \bar{\mathbf{c}} + \mathbf{E}\mathbf{w}\,, \tag{2}$$

where $\bar{\mathbf{c}}$ is the mean computed over the training data, $\mathbf{E} \in \mathbb{R}^{3n \times d}$ is a matrix that contains the d most dominant eigenvectors of the covariance matrix computed over the shape differences $\{\mathbf{c}_i - \bar{\mathbf{c}}_i\}$, and $\mathbf{w}$ is the low-dimensional shape parameter vector. One hypothesis of this model is that training faces can be linearly interpolated to generate new 3D faces. Another hypothesis is that the 3D faces in the reduced parameter space $\mathbb{R}^d$ follow a multivariate normal distribution, which can be directly deduced from the eigenvalues corresponding to $\mathbf{E}$. This implies that the density function $f(\mathbf{w})$ evaluating the likelihood of the parametric representation $\mathbf{w}$ in shape space is simply the Mahalanobis distance of $\mathbf{w}$ to the origin.

The 3DMM was originally computed over 200 subjects and has proven to be useful in a variety of applications thanks to its power to generate plausible shapes, and its simple underlying model. A

recent study rebuilds such a model from a very large dataset containing 9,663 3D scans and revisits best practices [Booth et al. 2016], demonstrating that the originally proposed generator function for shape remains highly relevant in the research community.

One observation by Blanz and Vetter [1999] is that moving the representation vector $\mathbf{w}$ away from the mean face increases their distinctiveness, eventually leading to caricatures of the identity. In order to model distinctive facial identities, Patel and Smith [2016] propose an alternative density function $f(\mathbf{w})$ based on the following observation. Consider the squared Mahalanobis distances from the mean for a set of d-dimensional vectors that follow a multivariate Gaussian distribution. These distances form a χ^2_d-distribution, which has expected value d. Hence, to preserve the shape distinctiveness related to identity, Patel and Smith restrict the representation $\mathbf{w}$ to have Mahalanobis distance $\sqrt{d}$ from the mean. Lewis et al. [2014b] propose a similar argument showing that, even if faces are truly Gaussian distributed (which has been shown for the Basel data by a Kolmogorov Smirnoff test for shape and per-vertex color, where the marginal distribution for the shape is close to a Gaussian [Egger et al. 2016b]), methods that make the assumption that typical faces lie near the mean are not valid.

Recently, Lüthi et al. [2018] proposed a nonlinear shape space that models deformations from the mean as Gaussian processes.

3.1.2 Local models. Using a global generator function in Equation (1) is known to lead to representations that do not model fine-scale geometric details. To improve the modeling of important localized areas, such as the eye or nose regions, Blanz and Vetter [1999] initially experimented manually segmenting the face into regions and learning separate PCA models per region. Their results demonstrate that this localized modeling allows for reconstructions of higher fidelity. This idea has been extended since with representations that achieve much higher accuracy than the global PCA model, and this comes in general at the cost of a less compact representation $\mathbf{w}$.

First local models segmented the face manually [Basso and Verri 2007; Kakadiaris et al. 2007; ter Haar and Veltkamp 2008]. Smet and Gool [2010] and Tena et al. [2011] propose automatic ways of segmenting the faces into areas based on information learned over the displacements of corresponding vertices in the training set. Brunton et al. [2011] propose a model that combines shape variations that are localized in different areas with a multi-resolution framework that uses a wavelet decomposition of the 3D face models. Fine-scale geometric detail can alternatively be modeled using hierarchical pyramids that consider differences between a smooth face and increasingly high-resolution geometry representing e.g., wrinkles [Golovinskiy et al. 2006].

It is also possible to perform localized analysis using different statistical approaches than PCA. Neumann et al. [2013] propose the use of sparse PCA combined with a group sparsity constraint to identify localized deformation components over the training data. Ferrari et al. [2015] follow a related idea and learn a dictionary of deformation components oversampled regions for the application of face recognition. Wu et al. [2016b] combine a local deformation subspace model with an anatomical bone structure that acts as a regularizer of the deformation. The local deformation subspace is computed over overlapping localized patches, and the statistical model explicitly factors the rigid and non-rigid deformations applied to each patch.

3.2 Expression models

As simple linear models similar to the ones described can be used to model expression variation for one subject, this section considers models that capture variations of both identity and expression. Unlike simple linear models learned over a dataset of varying identities and expressions (e.g., [Booth et al. 2017]), our focus is on models that explicitly decouple the influence of identity and expression by modeling them in separate coefficients. We classify these methods into additive, multiplicative, and nonlinear models, depending on how the two sets of coefficients are combined.

3.2.1 Additive models. Given two shapes of the same subject, one with expression $\mathbf{c}_{\text{exp}}$ and one neutral shape $\mathbf{c}_{\text{ne}}$, Blanz and Vetter [1999] transferred expressions between subjects by adding the expression offsets $\boldsymbol{\Delta}_c := \mathbf{c}_{\text{exp}} - \mathbf{c}_{\text{ne}}$ to the neutral shape of another subject.

Several other methods then built on this idea, and model expression variations as an additive offset to an identity model with a neutral expression. Formally, additive models are given by

$$\mathbf{c}(\mathbf{w}^s, \mathbf{w}^e) = \bar{\mathbf{c}} + \mathbf{E}^s\mathbf{w}^s + \mathbf{E}^e\mathbf{w}^e, \tag{3}$$

where $\bar{\mathbf{c}}$ is a mean, $\mathbf{E}^s$ and $\mathbf{E}^e$ are the matrices of basis vectors of the shape and expression space, and $\mathbf{w}^s$ and $\mathbf{w}^e$ are the shape and expression coefficients. Note that the basis vectors of the expression space can be interpreted as a data-driven blendshape model, where the basis vectors are orthogonal and do not carry interpretable semantic meaning in general [Lewis et al. 2014a].

Starting with Blanz et al. [2003], several methods propose to learn two PCA models, one over shape and one over expression to derive $\mathbf{E}^s$ and $\mathbf{E}^e$, and to compute $\bar{\mathbf{c}}$ as the mean over training data, either in neutral expression, or as sum of two means (one over shape and one over expression). Blanz et al. [2003] learned the expression space from a single subject captured in multiple expressions. Amberg et al. [2008] extended this work to include expression data from multiple subjects. This leads to a statistical expression model which does not enable control over specific facial expressions. It is therefore feasible for analysis-by-synthesis tasks but limited for controlling or synthesizing specific interpretable expression variation. Thies et al. [2015] use blendshapes as the basis vectors of the expression space. These expression blendshapes are not orthogonal and hence information of different blendshapes are potentially redundant.

3.2.2 Multiplicative models. Another body of work model shape and expression variations in a multiplicative manner. Li et al. [2010] propose a method to adapt a pre-defined blendshape model to a specific subject given a small number of static face scans in different expressions, which provides a personalized facial rig. Bouaziz et al. [2013] combine a morphable shape model $\mathbf{c}(\mathbf{w}^{\mathbf{s}})$ (Eq. 2) with a set of d_e linear expression transfer operators $\mathbf{T}_j : \mathbb{R}^{3n} \rightarrow \mathbb{R}^{3n}$ that transform the neutral shape to generate personalized blendshapes.

Formally, this model is defined as

$$\mathbf{c}(\mathbf{w}^s, \mathbf{w}^e) = \sum_{j=1}^{d_e} w_j^e \mathbf{T}_j \left(\mathbf{c}(\mathbf{w}^s) + \boldsymbol{\delta}^s\right) + \boldsymbol{\delta}_j^e, \tag{4}$$

where $\boldsymbol{\delta}^s$ and $\boldsymbol{\delta}_j^e$ are corrective vectors to adapt the blendshapes to the tracked subject, and w_j^e is the j-th coefficient of $\mathbf{w}^e$.

A commonly used multiplicative model is the multilinear model that extends the idea of PCA of performing a singular value decomposition to tensor data by performing a higher-order tensor decomposition (HOSVD) of 3D face data stacked into a training tensor. In particular, given a training set of different identities all captured in the same set of expressions, the vertex coordinates are stacked into a data tensor on which HOSVD is performed. This allows to model correlations of shape changes caused by identities and expressions. This model was first applied to 3D face modeling by Vlasic et al. [2005a], and can be defined as

$$\mathbf{c}(\mathbf{w}_2, \mathbf{w}_3) = \mathcal{M} \times_2 \mathbf{w}^s \times_3 \mathbf{w}^e, \tag{5}$$

where $\mathcal{M} \in \mathbb{R}^{3n \times d_s \times d_e}$ denotes the multilinear model tensor, and $\times_i$ denotes the tensor mode-product. Thanks to its expressiveness and simplicity, this model is being used extensively for various applications [Bolkart and Wuhrer 2015a; Dale et al. 2011; Fried et al. 2016; Mpiperis et al. 2008; Yang et al. 2012]. To allow modeling localized variations, the multilinear model has been applied to wavelet coefficients at different levels of detail [Brunton et al. 2014a].

Computing a multilinear model with HOSVD requires a complete tensor of data, where each identity needs to be present in all expressions, and the data need to be in semantic correspondence specified by expression labels. This severely limits the kind of data that can be used for training. Recently, a number of methods have been proposed to address this limitation using an optimization approach [Bolkart and Wuhrer 2016], a custom tensor decomposition method [Wang et al. 2017], and an autoencoder structure [Fernández Abrevaya et al. 2018], respectively.

3.2.3 Nonlinear models. Facial shape and expression are mostly modeled with a linear subspace, often assuming a Gaussian prior distribution. Few methods exist to model facial variations with nonlinear transformations. Li et al. [2017] introduce FLAME, an articulated expressive head model that provides nonlinear control over facial expressions by combining jaw articulation with linear expression blenshapes. Ichim et al. [2017] use a muscle activation model driven by physical simulation. Koppen et al. [2018] instead of a single Gaussian distribution use a Gaussian mixture model to represent facial shape and texture. In another line of works, Shin et al. [2014] capture facial wrinkles in multi-scale maps and nonlinearly transfer them to other faces to enhance realism.

Recently, several deep learning-based models were published that fall into this group of nonlinear models [Bagautdinov et al. 2018; Lombardi et al. 2018; Ranjan et al. 2018; Tewari et al. 2019, 2018; Tran and Liu 2018a]. Section 6 covers these models in more detail.

3.3 Appearance models

This section describes approaches for modelling the facial appearance, where we distinguish between *linear* and *nonlinear* models. The appearance of a face is influenced by its albedo and illumination. However, most 3DMMs do not completely separate these factors, so that oftentimes the illumination is baked into the albedo. Hence, in the following, we call the problem of statistically capturing this information *appearance modeling*. The most common way to build an appearance model is by performing statistics on appearance information of the training shapes, where the appearance information is usually either represented in terms of per-vertex values or as a texture in *uv*-space.

3.3.1 Linear per-vertex models. Usually, color information is modeled as a low-dimensional subspace that explains the color variations. This leads to an analogous model to the linear shape model:

$$\mathbf{d}(\mathbf{w}^t) = \bar{\mathbf{d}} + \mathbf{E}^t \mathbf{w}^t, \tag{6}$$

where $\bar{\mathbf{d}}$ and $\mathbf{E}^t$ shares the same number of rows as $\bar{\mathbf{c}}$ and $\mathbf{E}$ and $\mathbf{w}^t$ is the low-dimensional texture parameter vector.

Booth et al. [2017] and Booth et al. [2018b] use a convex matrix factorization formulation for learning a per-vertex appearance model from images based on back-projection, where it is assumed that the 3D geometry of the face in the image is known. Their appearance model is not built using the color images directly but rather features computed from the images, for example, SIFT. This brings advantages that the features may be somewhat invariant to illumination changes and also that they depend on a local neighborhood which may widen the basin of convergence. In a similar vein, Wang et al. [2009] construct a linear model of spherical harmonic bases (see Section 4). This jointly models texture (more precisely diffuse albedo) and fine-scale shape (surface normal orientation) such that appearance under any illumination can be synthesized as a linear function of the basis.

3.3.2 Linear texture-space models. A downside of per-vertex models is that they require compatible resolutions between the shape and appearance representation. This is rather uncommon in computer graphics, where usually a low(er) resolution geometry model (oftentimes including normals) is used in conjunction with a high(er) resolution 2D texture map. Working with a 2D texture also has other advantages, such as the possibility of using image processing techniques to modify the texture maps. With that, such a representation is also amenable for being processed by convolutional neural networks (CNNs), as will be addressed in the next section.

We now turn our attention towards works that build linear appearance models in texture space. The original work by Blanz and Vetter [1999] used a texture-based representation by representing the face in a cylindrical way. Later, texture-based representations were used to add textural details like wrinkles [Pascal 2010], or to segment skin and detect moles [Pierrard 2008]. Cosker et al. [2011] model appearance variation in *uv*-space based on sequences of facial images recorded from different views. The images of the dynamic sequences are aligned based on a non-rigid registration so that the color variation can be modeled using a linear subspace model based on PCA. Dai et al. [2017] also use a *uv*-space appearance representation that is defined for the entire head. Huber et al. [2016] use a per-vertex appearance variation model based on PCA, but in addition, also define a common *uv*-mapping so that the model can be textured based on given facial images. Moschoglou et al. [2018] formulate a robust matrix factorization problem in order to learn

attributed facial *uv*-maps from a collection of training textures. A study on the effect of different *uv*-space embeddings of the texture was presented by Booth and Zafeiriou [2014].

3.3.3 Nonlinear models. Traditionally, the facial appearance is modeled as a linear subspace, where oftentimes a Gaussian distribution is assumed. However, as empirically shown by Egger et al. [2016a], the Gaussian assumption is not very accurate and may lead to a sub-optimal facial appearance model. Hence, the authors proposed to replace a PCA-based appearance model with a *Copula Component Analysis* model [Han and Liu 2012]. Subsequently, this idea was extended to jointly model facial shape, texture, and attributes [Egger et al. 2016b]. Recent work learned a joint shape and texture model using neural networks with an adverserial loss [Gecer et al. 2019a]. Alotaibi and Smith [2017] use the observation that skin color forms a nonlinear manifold in RGB space, approximately spanned by the colors of the pigments melanin and hemoglobin. They inverse render maps of these parameters and then construct a linear statistical model in the parameter space. The resulting biophysical 3DMM is guaranteed to produce plausible skin colors. In addition to global facial appearance models, there are also approaches that consider models of local skin variations. For example, Dessein et al. [2015] use a texture model based on small overlapping patches that are extracted from a face database, and Schneider et al. [2018] have presented a stochastic model that is able to synthesize freckles.

More recently, a range of appearance modeling approaches based on deep learning have been proposed, where many of these methods are also built within an analysis-by-synthesis framework. These aspects will be discussed in-depth in Secs. 5 and 6.

3.4 Joint shape and appearance models

Blanz and Vetter [1999] originally proposed building separate, independent models for shape and texture. Interestingly, in 2D the Active Appearance Model [Cootes et al. 1998] was originally proposed with a combined shape and appearance model. The advantage of such a joint model is that correlations between shape and texture can be learned and exploited as a constraint during fitting with fewer parameters. On the other hand, separate models are more flexible and, since shape and texture parameters can be adjusted independently, sequential algorithms can fit the two models independently. However, 3DMMs that jointly model shape and texture have subsequently been considered. Schumacher and Blanz [2015] use canonical correlation analysis to study shape/texture correlations and also correlations between face parts. Egger et al. [2016a] use copula component analysis that can deal with the different scales of shape and texture data. Zhou et al. [2019] propose a deep convolutional colored mesh autoencoder that learns a joint nonlinear model of shape and texture.

3.5 Correspondence

The previously discussed models typically require the data with point-to-point correspondence between all shapes. We refer to the process of establishing such a dense correspondence between scans as registration in the following.

Many methods exist to establish point-to-point correspondence for general classes of objects (e.g., [Tam et al. 2013; van Kaick et al. 2011]), yet the space of face deformations is strongly constrained. Most commonly used face registration methods follow the principle of deforming a template mesh to each scan in the dataset. This registration process typically starts with a rough alignment (often using sparse correspondences) and leads to dense correspondences in the end.

While several image-based methods can also be seen as jointly learning correspondence (between images) and building a statistical model (e.g., [Tewari et al. 2019; Tran and Liu 2018a]), we cover such deep learning-based methods in Section 6 in more details.

3.5.1 Sparse correspondence computation. Several methods exist to establish a sparse correspondence for a dataset of 3D scans by predicting landmarks, i.e. a common set of salient points, for each scan. This sparse correspondence then typically serves as automatic initialization for dense correspondence methods.

Most of the methods use some local descriptors, or combination of local descriptors and connectivity information between descriptors, to predict salient points. While landmark localization in images is widely researched (e.g., Bulat and Tzimiropoulos [2017]), our focus is on methods that establish sparse correspondence between 3D scans.

Existing methods use combinations of different geometric descriptors. Passalis et al. [2011] use shape index and spin image features, Berretti et al. [2011] use curvature and scale-invariant feature transform (SIFT) features, and Creusot et al. [2013] consider combinations of local features such as Gaussian curvature, mean curvature, and a volumetric descriptor, and learn the statistical distribution of these descriptors for each landmark.

Further, existing methods use geometric relations or relations between landmarks along with geometric feature descriptors. Guo et al. [2013] project a scan into an image and predict landmarks with a 2D PCA model and geometric relations with additional texture constraints. Salazar et al. [2014], similarly to Creusot et al. [2013], learn the statistical distribution of local surface descriptors with additional Markov network to additionally consider connections between landmarks. Bolkart and Wuhrer [2015a] extend this further to sequences by additionally considering temporal edges within the Markov network.

3.5.2 Dense correspondence computation. Methods that deform a template to establish correspondence mostly differ in the parameterization of the deformation. We group existing methods according to the type of scan data they register. We distinguish here between static methods, i.e. methods that aim at registering static 3D scan, and dynamic methods that register 3D motion sequences. Blanz and Vetter [1999] use a bootstrapping approach to iteratively fit a 3DMM to a scan, refine the correspondence between the model fit and the scan with a flow field, and refine the model. Blanz et al. [2003] later extend this approach to expressive scans. Amberg et al. [2008] register expressive scans with a non-rigid ICP. Hutton et al. [2001] establish a thin-plate spline (TPS) mapping to warp each scan to a reference and establish correspondence using nearest neighbor search.

Passalis et al. [2011] register scans by deforming an annotated face model (AFM) [Kakadiaris et al. 2005], i.e. an average 3D face template that is segmented into different annotated areas, by solving

a second-order different equation. Mpiperis et al. [2008] initially fit a subdivision surface to a scan, where the deformation of the base mesh (i.e. the mesh of the lowest subdivision level) is guided by a sparse landmark correspondence. After registering a training set, they parametrize the deformation of the base mesh with a PCA model over the training data. Salazar et al. [2014] use a generic expression blendshape model to fit the expression of the scan, followed by a non-rigid ICP to closely fit the surface of the scan. [Gerig et al. 2018] establish dense correspondence with a Gaussian process deformation model with the spatially varying kernel.

Several methods exist to sequentially register motion sequences. Weise et al. [2009] use an identity PCA model to register a neutral scan, and then track motion sequences by optimizing sparse and dense optical flow between consecutive frames. Fang et al. [2012] and Li et al. [2017] initialize the optimization by the registration of the previous frame to exploit temporal information. Fang et al. [2012] use an AFM, Li et al. [2017] a non-rigid ICP regularized by FLAME. Fernández Abrevaya et al. [2018] uses a spatiotemporal method to register entire motion sequences by iteratively refining the registration of entire sequences by explicitly encoding temporal information with a Discrete Cosine Transform (DCT).

Further, non-template fitting based registration methods exist. Sun et al. [2010] use a conformal mapping to parameterize two meshes and establish dense correspondence between the resulting planar meshes by extrapolating sparse landmark correspondences. Ferrari et al. [2015] segment the face scans into non-overlapping parts divided by geodesic curves between selected landmark pairs, and consistently re-sample each part.

3.5.3 Jointly solving for correspondence and statistical model. Li et al. [2013] and Bouaziz et al. [2013] jointly update person-specific blendshape models and register motion sequences in a real-time facial animation framework. During tracking, Li et al. [2013] use an adaptive PCA model that combines the person-specific blendshapes with additional corrective basis vectors that are successively updated, and Bouaziz et al. [2013] optimizes for corrective deformation fields (Equation 4).

Bolkart and Wuhrer [2015b] and Zhang et al. [2016b] instead optimize correspondence for a dataset of different subjects in multiple expression in a groupwise fashion. Bolkart and Wuhrer [2015b] jointly update the point correspondence within the mesh surface by minimizing an objective function that measures the compactness of a multilinear face model. Zhang et al. [2016b] optimize functional maps across the entire dataset.

3.6 Synthesis of novel model instances

3DMMs can be used to synthesize new 3D faces that are different from any of the observed training data, yet realistic. This can be achieved by altering the coefficients in parameter space (i.e. shape space, expression space or appearance. Common operations in parameter space include interpolating or extrapolating between the coefficients of training samples. Furthermore, any of the generative models presented in this section can be used to directly synthesize new 3D faces by drawing random samples in parameter space according to the prior distribution. Depending on the model, this sampling allows to synthesize or alter identity, expression, or appearance of a static 3D face. Synthesis works are heavily used for entertainment purposes, and these works are discussed in Section 7.2.

Synthesis of static 3D faces notably includes the generation of face caricatures by moving the identity coefficient linearly away from the mean [Blanz and Vetter 1999] which is mainly explored to study the human face processing system as discussed in Section 7.5.

With 3DMMs that encode and decouple identity and expression information, it is easy to synthesize dynamic sequences by fixing the identity coefficients while modifying the expression coefficients. Some works aim to synthesize coherent dynamic 3D face videos of a fixed identity with the help of 3DMMs. These include works that synthesize 4D videos from a static 3D mesh paired with semantic label information [Bolkart and Wuhrer 2015a], and from a static 3D mesh and audio information [Cudeiro et al. 2019].

3.7 Publicly available models

In Table 2 we list publicly available shape and/or appearance models of human faces. Figure 4 visualizes geometry or appearance variations of some models. We also refer to the curated list of 3DMM software and data that we collected, share and update [Community 2019].

3.8 Open challenges

While 3D face modeling has received considerable attention during the past two decades, some challenges remain. First, the statistics of most models are limited to the face and do not include information on eyes, mouth interior or hair. These details are however crucial for many applications, and it is not straightforward to combine a 3DMM with specific models e.g., for hair. Second, the interpretability of the representations would benefit from being improved. PCA is the most commonly used method to perform statistics on 3D faces, and as it is an unsupervised method, the principal components do not coincide with attributes that humans would use to describe a face. Third, methods that incorporate different levels of detail typically come at the cost of a less compact representation, and it is unknown how many parameters are required to accurately represent facial geometry and appearance at varying levels of detail. Fourth, the different models presented in this section have different advantages and drawbacks, making them most suitable for specific applications. It is unknown whether one integrated optimal model for all applications exists. Fifth, all currently available models, even the large scale ones have a very strong racial bias towards white. This can be alleviated in the future by scanning efforts in different parts of the world. Another potential solution to overcome a racial bias can be the generative model itself, as these allow to generate and add synthetic data to biased datasets. Sixth, learning from inhomogeneous data presents another open challenge. There are many available datasets with different resolution, coverage, noise characteristics, biases and so on (see Table 1). To make the best use of this data requires methods that can learn models from all data sources simultaneously but this requires explicit ways to deal with data inhomogeneity. Some very recent work begins to look at this problem [Liu et al. 2019b]. Finally, there are some fundamental open questions related to the statistical modeling of shape. Two face shapes differ by nonlinear shape deformation superposed on

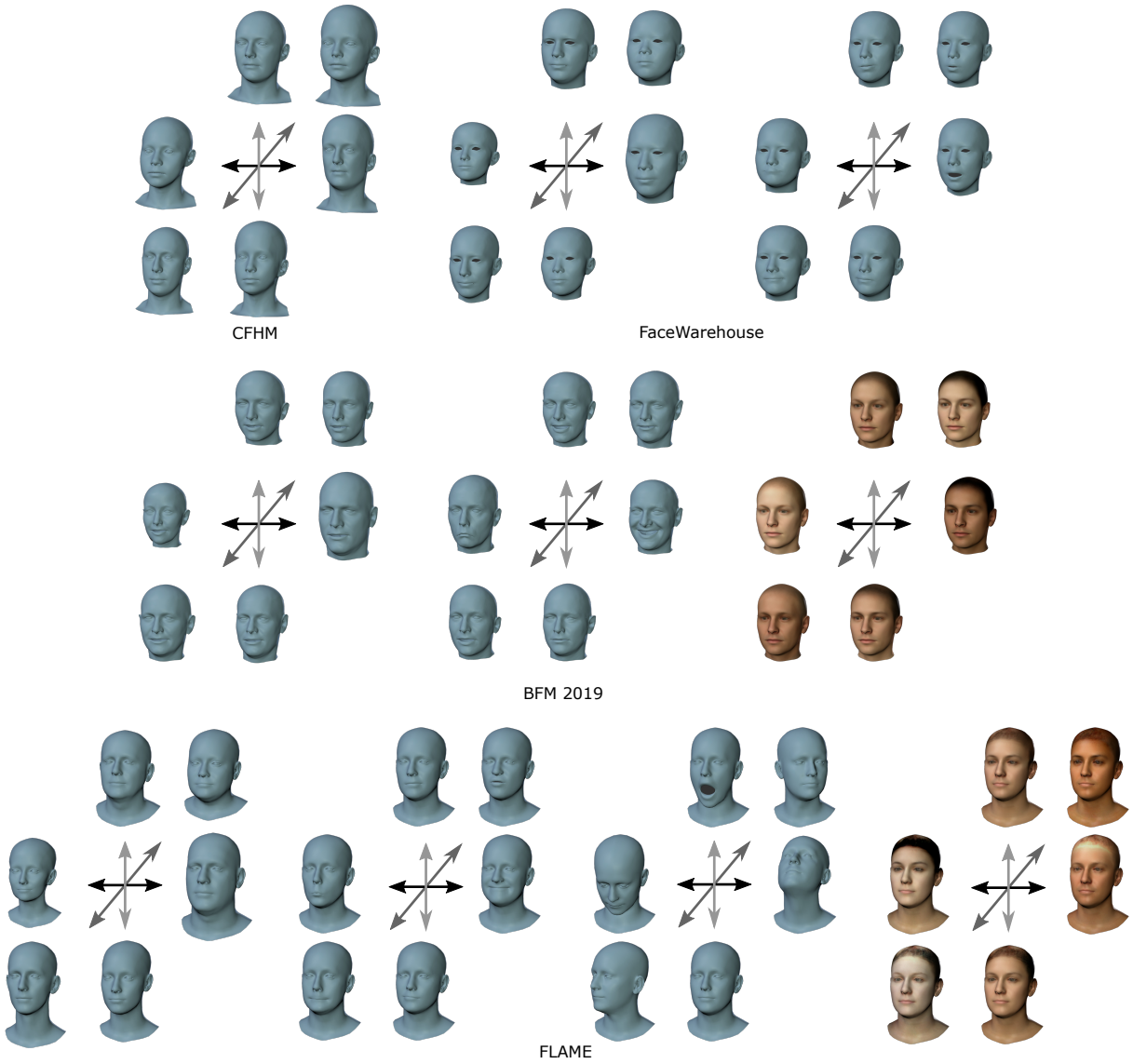


Fig. 4. Model variations of existing face models. Top left: CFHM [Ploumpis et al. 2019] shape variations. Top right: FaceWarehouse [Cao et al. 2014b] shape and expression variations (while the original model is not available to the best of our knowledge, the visualized multilinear face model is trained from the published FaceWarehouse dataset). Middle: BFM 2019 [Gerig et al. 2018] shape, expression, and appearance variations. Bottom: FLAME [Li et al. 2017] shape, expression, pose, and appearance variation. For shape, expression, and appearance variations, three principal components are visualized at ±3 standard deviations. The FLAME pose variations are visualized at $\pm\pi/6$ (components three and four) and at 0, $\pi/8$ (component six).

top of rigid body motion. Conventionally, this is dealt with by first rigidly aligning, then modeling the residual shape differences but this makes the model dependent on the choice of alignment metric. For faces specifically, estimated skull position has been used for rigid alignment [Beeler and Bradley 2014]. Although not applied to faces, a recent method uses a rigid body motion invariant distance measure to learn nonlinear principal components [Heeren et al. 2018].

4 IMAGE FORMATION

A 3DMM provides a parametric representation of face geometry and appearance. One key usage of such a model is synthesis, which involves two steps. First, generating a new model instance via sampling from the parameter space or manual interaction with model parameters (see Section 3.6). Second, rendering the generated model into a 2D image via a simulation of the image formation process, i.e. the computer graphics pipeline. The synthesis also forms an important part of 3DMM-based face analysis, either through classical analysis-by-synthesis (see Section 5) or as a model-based decoder

model	geometry	appearance	data	comment
Basel Face Model (BFM) 2009 [Paysan et al. 2009b]	shape	per-vertex	200 individuals, each in neutral expression	includes separate models for facial parts
FaceWarehouse [Cao et al. 2014b]	shape, expression	-	150 individuals, each with 20 expressions	
Global and local linear model [Brunton et al. 2014b]	shape	-	100 individuals	
Multilinear Wavelet model [Brunton et al. 2014a]	shape, expression	-	99 individuals, 25 expressions	
Multilinear face model [Bolkart and Wuhrer 2015b]	shape, expression	-	2500 scans (100 individuals, 25 expressions)	
Multilinear face model [Bolkart and Wuhrer 2016]	shape, expression	-	2510 scans (205 individuals, up to 23 expressions)	
Large Scale Facial Model (LSFM) [Booth et al. 2016]	shape	-	9663 individuals	
Surrey Face Model [Huber et al. 2016]	shape, expression	per-vertex	169 individuals	multi-resolution
Liverpool-York Head Model (LYHM) [Dai et al. 2017]	shape	per-vertex	1212 individuals	full head (no hair, no eyes)
Faces Learned with an Articulated Model and Expressions (FLAME) [Li et al. 2017]	shape, expression, head pose	texture	3800 individuals for shape, 8000 for head pose, 21000 frames for expression	female, male, gender neutral model, full head (no hair)
Basel Face Model (BFM) 2017 [Gerig et al. 2018]	shape, expression	per-vertex	200 individuals for shape and appearance, a total of 160 expression scans	BFM 2019 with full head and multi-resolution
York Ear Model [Dai et al. 2018]	shape	-	20 3D ear scans, augmented with 605 landmark-annotated 2D ear images	ear only
Multilinear autoencoder [Fernández Abrevaya et al. 2018]	shape, expression	-	5000 scans from 195 individuals, 500000 after augmentation	
Convolutional Mesh Autoencoder (CoMA) [Ranjan et al. 2018]	shape, expression	-	12 individuals, 12 extreme expressions, 20466 meshes in total	full head (no hair)
Combined Face & Head Model (CFHM) [Ploumpis et al. 2019]	shape	-	Merged from LYHM and LSFM models	full head (no hair)
Morphable Face Albedo Model [Smith et al. 2020]	-	per-vertex diffuse and specular albedo	73 individuals (50 scanned + 23 3DRFE [Stratou et al. 2011])	extends BFM2017

Table 2. Overview of publicly available 3D shape and/or appearance models of human faces.

within a deep learning architecture (see Section 6). In this section, we focus on modeling the image formation process. This potentially encompasses the whole of the rendering literature, so we restrict our attention to techniques and models that have been applied in the context of 3DMMs. We cover the geometry and photometry of image formation in Sections 4.1 and 4.2, the rendering pipelines used for 3DMM fitting in Section 4.3 and finally in Section 4.4 we highlight where there are future opportunities for exploiting state-of-the-art rendering techniques to improve 3DMM synthesis.

4.1 Geometric image formation

A camera model describes the *geometry* of image formation, specifically, how positions in the 3D world project to 2D locations in the image plane. A variety of camera models have been used in the

3DMM literature which are described here in order of increasing accuracy with respect to a real camera. We denote the projection of a 3D point $\mathbf{v} = [u, v, w]^{\mathrm{T}}$ onto the 2D point $\mathbf{x} = [x, y]^{\mathrm{T}}$ by $\mathbf{x} = \mathbf{project}[C, \mathbf{v}] \in \mathbb{R}^2$, where **project** represents one of the camera projection models below and $C = (C_{\text{intrinsic}}, C_{\text{extrinsic}})$ contains the camera parameters. $C_{\text{extrinsic}} = (\mathbf{R}, \mathbf{t})$ describes the pose in terms of a rotation $\mathbf{R} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ that transform from world to camera coordinates. $C_{\text{intrinsic}}$ is a set of internal parameters specific to each projection model. The task of estimating C is known as camera calibration or camera resectioning and is usually done from known or estimated 2D-3D correspondences. Estimating $C_{\text{extrinsic}}$ with known $C_{\text{intrinsic}}$ is called pose estimation or, in the case of a perspective camera model, the perspective-n-point problem.

Scaled orthographic. The scaled orthographic projection model comprises an orthographic projection whose sole parameter is a uniform scaling $s \in \mathbb{R}_{>0}$:

$$\mathbf{ortho}[\mathbf{v}, \mathbf{R}, \mathbf{t}, s] = s\mathbf{P}(\mathbf{R}\mathbf{v}+\mathbf{t}) = s\begin{bmatrix}\mathbf{r}_1 & t_1 \\ \mathbf{r}_2 & t_2\end{bmatrix}\tilde{\mathbf{v}} = \mathbf{C}\tilde{\mathbf{v}}, \quad \mathbf{P} = \begin{bmatrix}1 & 0 & 0 \\ 0 & 1 & 0\end{bmatrix} \tag{7}$$

where $\tilde{\mathbf{v}} = [u, v, w, 1]^{\mathrm{T}}$ is the homogeneous representation of $\mathbf{v}$ and $\mathbf{r}_1$, $\mathbf{r}_2$ are the first two rows of $\mathbf{R}$. This model is not physically meaningful but is linear in vertex position, translation and scale and avoids size/distance/perspective ambiguities introduced by more realistic camera models. Since $\mathbf{R}$ must be restricted to $SO(3)$, the projection is nonlinear in any parameterization of $\mathbf{R}$. In the context of 3DMMs, scaled orthographic projection has been used for example by Bas et al. [2017b]; Blanz et al. [2004a]; Knothe et al. [2006]; Patel and Smith [2009]. The scaled orthographic model can be interpreted as an approximation to perspective projection when the distance between the surface and camera is large relative to the depth variation. Concretely, when $\max_w - \min_w \ll \bar{w}$ with $\bar{w} = \text{mean}_w$ the mean distance between the surface and the camera, then the nonlinear division in perspective projection can be approximated by a fixed scale $s = f/\bar{w}$ where f is the focal length of the camera. This gives physical meaning to the scaled orthographic model.

Affine. The affine camera generalizes the orthographic model by allowing arbitrary affine transformations. Specifically, this additionally allows non-uniform scaling and skew transformations which can approximate perspective effects whilst remaining linear. An affine camera can be represented by an arbitrary matrix $\mathbf{C} \in \mathbb{R}^{2\times4}$ with the projection given simply by $\mathbf{x} = \mathbf{affine}[\mathbf{v}, \mathbf{C}] = \mathbf{C}\tilde{\mathbf{v}}$. The projection is linear in $\mathbf{C}$ and since its 8 entries are unconstrained, they can be estimated using linear least squares (though note that numerical stability entails first performing a normalization procedure). In the context of 3DMMs, the affine camera has been used for example by Aldrian and Smith [2013]; Huber et al. [2016].

Perspective. A nonlinear perspective projection is given by the pinhole camera model $\mathbf{x} = \mathbf{pinhole}[\mathbf{v}, \mathbf{K}, \mathbf{R}, \mathbf{t}]$. The matrix:

$$\mathbf{K} = \begin{bmatrix}f_x & \gamma & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1\end{bmatrix}$$

contains the intrinsic parameters of the camera, namely the focal length in the x and y directions $f_x, f_y \in \mathbb{R}_{>0}$, the skew $\gamma \in \mathbb{R}$ and the principal point $[c_x, c_y] \in \mathbb{R}^2$. Common assumptions are that the pixels are square (in which case a single focal length $f = f_x = f_y$ parameter is used), that the camera sensor is perpendicular to the camera view vector (in which case $\gamma = 0$) and that the principal point is in the centre of the image ($c_x = w/2$ and $c_y = h/2$). Note that f is actually a product of two quantities: the physical focal length in world units (e.g., mm) and the conversion factor from world units to pixels (i.e. with units of pixels/mm). The nonlinear perspective projection can be written in linear terms by using homogeneous representations:

$$\lambda\tilde{\mathbf{x}} = \mathbf{K}\begin{bmatrix}\mathbf{R} & \mathbf{t}\end{bmatrix}\tilde{\mathbf{v}} = \mathbf{C}\tilde{\mathbf{v}}, \tag{8}$$

where $\tilde{\mathbf{x}} = [x, y, 1]^{\mathrm{T}}$, λ is an arbitrary scaling factor and $\mathbf{C} \in \mathbb{R}^{3\times4}$ is known as the camera matrix. The final image coordinate is obtained by the nonlinear homogenization of $\tilde{\mathbf{x}}$. Unlike the linear models, the pinhole model captures the effect of distance on projected shape. This becomes important when a face is close to the camera. At "selfie" distance (e.g., 0.5m), the difference between perspective and orthographic projection of 3D face landmarks is about 6% of the interocular distance [Bas and Smith 2019]. For this reason perspective projection is commonly used in the context of 3DMMs, for example, in the original Blanz and Vetter [1999] paper and more recently in a shape-from-landmarks setting [Cao et al. 2014a, 2013; Saito et al. 2016]. Unfortunately, since calibration information is rarely available the increased complexity of this model introduces ambiguities between shape, scale and focal length that have only recently been studied [Bas and Smith 2019; Smith 2016], though the ambiguity has often been hinted at in the literature. For example, the original Blanz and Vetter [1999] paper relied on a fixed, manually provided subject-camera distance. Booth et al. [2018b] state "we found that it is beneficial to keep the focal length constant in most cases, due to its ambiguity with t_z". Schönborn et al. [2017] explored the ambiguity of estimated distance from the camera under perspective and observed a very high posterior standard deviation and the distance can not be resolved even by using a strong prior for the face shape. In Tewari et al. [2017], a similar effect is observed, indicated by the learning rate on the z translation (i.e. subject-camera distance), which is set three orders of magnitude lower than all other parameters. Both approaches in practice fix the face distance to avoid this difficulty.

4.2 Photometric image formation

The appearance of a face is determined by the interaction of light with the material of the face, predominately skin. Hence, the *photometry* of both illumination and reflectance must be modeled in order to simulate the image formation process.

Reflectance models. The reflection of light from a surface is often described using a Bidirectional Reflectance Distribution Function (BRDF). This describes the directional dependence of local light reflection from an opaque surface. It is represented by a four dimensional function $f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o)$ that gives the ratio of *outgoing* reflected radiance in direction $\boldsymbol{\omega}_o$ to *incoming* incident irradiance from direction $\boldsymbol{\omega}_i$. A BRDF allows us to express irradiance $L_o(\boldsymbol{\omega}_o)$ in direction

$\boldsymbol{\omega}_o$ as a function of light reflected from all incident directions:

$$L_o(\boldsymbol{\omega}_o) = \int_{\Omega(\mathbf{n})} f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) L_i(\boldsymbol{\omega}_i) \cos\theta_i \mathrm{d}\boldsymbol{\omega}_i, \tag{9}$$

where $L_i(\boldsymbol{\omega}_i)$ is incident irradiance from direction $\boldsymbol{\omega}_i$, $\Omega(\mathbf{n})$ is the hemisphere around the local surface normal $\mathbf{n}$ and θ_i is the angle between $\boldsymbol{\omega}_i$ and $\mathbf{n}$. Note $\cos\theta_i = \mathbf{n} \cdot \boldsymbol{\omega}_i$, where $\cdot$ denotes the inner product. Physically-valid BRDFs must exhibit a number of properties: nonnegativity ($f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) \geq 0$), Helmholtz reciprocity ($f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) = f_r(\boldsymbol{\omega}_o, \boldsymbol{\omega}_i)$) and conservation of energy:

$$\forall \boldsymbol{\omega}_i, \int_{\Omega(\mathbf{n})} f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) \cos\theta_i \mathrm{d}\boldsymbol{\omega}_o \leq 1. \tag{10}$$

A particularly simple and commonly used physically valid BRDF is the Lambertian model for a perfectly diffuse reflector. The Lambertian model assumes incident light is scattered equally in all directions resulting in a constant BRDF: $f_{\text{Lambert}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) = \rho_d/\pi$. $\rho_d \in [0, 1]$ is the diffuse reflectivity or *albedo*, that is usually wavelength dependent and can be thought of as the color of the object. Work predating the original 3DMM used a linear statistical 3D face shape model with the Lambertian reflectance model in a shape-from-shading context [Atick et al. 1996]. Subsequently, the Lambertian model has been used for 3DMM fitting in the context of the spherical harmonic lighting model [Zhang and Samaras 2006] (see below) where its simplicity yields closed-form expressions. This is now very common, including in the current state-of-the-art, e.g., [Tran et al. 2019; Tran and Liu 2018b]. In general, the Lambertian model is a poor approximation for the complex reflectance properties of facial skin, hair, eyes, etc., and so more sophisticated models have been considered.

Blanz and Vetter [1999] originally used the Phong model which augments the Lambertian term with a constant ambient term and a phenomenological specular model enabling the simulation of glossy reflectance. The Phong model can be described in terms of the following BRDF:

$$f_{\text{Phong}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) = \frac{\rho_a + \rho_s(\mathbf{r} \cdot \omega_o)^\eta}{\mathbf{n} \cdot \boldsymbol{\omega}_i} + \rho_d \tag{11}$$

where $\mathbf{r}$ is the reflection of $\boldsymbol{\omega}_i$ about $\mathbf{n}$, η is the *shininess* that controls the width of the specular lobe and ρ_a, ρ_s are ambient and specular "albedos". In the context of 3DMMs, usually only ρ_d is allowed to vary spatially. Note that the Phong BRDF does not satisfy the constraints above for physical validity. In graphics, extremely complex, physically-valid BRDF models have been developed specifically for materials of relevance to face 3DMMs, for example for skin [Krishnaswamy and Baranoski 2004] and hair [Marschner et al. 2003]. Note that skin is a layered, partially translucent material and so a local BRDF model is inadequate to describe the actual subsurface scattering effects that take place. More complex 8-dimensional bidirectional subsurface scattering reflectance distribution functions (BSSRDF) have been proposed for such materials. However, both these and the more complex BRDFs have proven to be too complex to integrate into 3DMM fitting pipelines and so the majority of work has used Lambertian or non-physical models of moderate complexity.

Lighting. In (9), $L_i(\boldsymbol{\omega}_i)$ represents the hemispherical incident illumination environment at the surface point. Natural illumination is usually complex, comprising multiple, possibly extended sources as well as secondary illumination reflected from other surfaces. A common assumption is that the illumination environment is distant relative to the size of the object in which case it can be represented by a constant 2D *environment map*, a discrete approximation of $L_i(\boldsymbol{\omega}_i)$ that is used for every point on the surface. However, the space of possible natural illumination is very high dimensional and rendering with an environment map is computationally expensive, so a number of further simplifications are commonly used.

The simplest illumination model is a point source, in which $L_i(\boldsymbol{\omega}_i)$ is a delta function characterized by a unit vector in the light source direction, $\mathbf{s}$, and an intensity, L_i. Ignoring constants and assuming image intensity is proportional to surface radiance, we can plug in the simple BRDF models above and obtain the following shading models:

$$I_{\text{Lambert}} = L_i \rho_d \mathbf{n} \cdot \mathbf{s}, \quad I_{\text{Phong}} = L_i \left[\rho_a + \rho_d \mathbf{n} \cdot \mathbf{s} + \rho_s(\mathbf{r} \cdot \mathbf{v})^\eta\right], \tag{12}$$

where $\mathbf{v}$ is a unit vector in the viewer direction. Usually, the light source intensity and albedos would be RGB values. Often $\rho_a = \rho_d$, representing the intrinsic color of the surface. In a 3DMM, this is described by the statistical texture model (6). Note that these simple models are purely local, this means that they neglect self occlusion of the light source, i.e. cast shadows. These can be added at the cost of computing the occlusion function which is not differentiable.

A better approximation of complex natural illumination is provided by the spherical harmonic illumination model [Ramamoorthi and Hanrahan 2001]. Spherical harmonics provide an orthonormal basis for functions on the sphere, analogous to a Fourier basis in Euclidean space:

$$I_{\text{SH}}(\mathbf{n}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} l_{l,m} \mathcal{B}_{l,m}(\mathbf{n}), \tag{13}$$

where $\mathcal{B}_{l,m}(\mathbf{n})$ are the orthonormal basis functions, and $l_{l,m}$ are coefficients describing reflectance and illumination. The subscript l denotes the degree and m the order of the spherical harmonics. In the Lambertian case, the contribution from the reflectance function is constant and 98% of the energy of the reflectance function can be captured for any illumination environment using an order 2 ($l = \{0, 1, 2\}$) approximation. In practice, this means that a good approximation for appearance can be obtained using 9 illumination coefficients per color channel. Combining this model with the linear texture model for albedo $\mathbf{d}(\mathbf{w}^t)$ yields the following:

$$\mathbf{i}_{\text{SH}} = \mathbf{d}(\mathbf{w}^t) \odot \text{vec}(\mathbf{B}(\mathbf{w}^s)\mathbf{L}), \tag{14}$$

where $\odot$ is the Hadamard (element-wise) product. The matrix $\mathbf{B}(\mathbf{w}^s) \in \mathbb{R}^{n \times 9}$ contains the spherical harmonic basis for each vertex which depends on the vertex normal direction and hence the geometry which in turn is determined by the shape parameter vector. The matrix $\mathbf{L} \in \mathbb{R}^{9 \times 3}$ contains the lighting coefficients for each color channel. Zhang and Samaras [2006] were the first to fit this model in the context of 3DMMs. Aldrian and Smith [2013] additionally used the same model for specular reflection, showing that the coarse structure of the illumination environment can be recovered

from a face image. They also introduced priors to help resolve lighting/texture ambiguities. Egger et al. [2018] went further by learning a low dimensional illumination prior from spherical harmonic lighting coefficients estimated from real in-the-wild images.

An alternate model that takes a step towards capturing global illumination effects is the *ambient occlusion* model. Here, it is assumed that $L_i(\boldsymbol{\omega}_i)$ is constant everywhere, i.e. that illumination is perfectly diffuse. In this case, shading depends only upon the degree to which the incident hemisphere is occluded. The ambient occlusion, $A_{\mathbf{v}}$, at vertex $\mathbf{v}$ is given by:

$$A_{\mathbf{v}} = \frac{1}{\pi} \int_{\Omega(\mathbf{n})} V(\mathbf{v}, \boldsymbol{\omega})(\mathbf{n} \cdot \boldsymbol{\omega}) \mathrm{d}\omega, \tag{15}$$

where $V(\mathbf{v}, \boldsymbol{\omega})$ is the visibility function defined as zero if vertex $\mathbf{v}$ is occluded in direction $\boldsymbol{\omega}$, and one otherwise. One can also define the *bent normal* as the average unoccluded direction. Using the bent normal with the spherical harmonic illumination model and scaling the result by the ambient occlusion provides a rough approximation of global illumination effect. Ambient occlusion and bent normal direction depends on the geometry and hence the 3DMM shape parameters. Aldrian and Smith [2012] proposed to learn a linear model of ambient occlusion and bent normals and included this in their 3DMM synthesis model. Zivanov et al. [2013] similarly construct a joint linear model of spherical harmonic bases and ambient occlusion.

The most complex global model of appearance considered in the context of 3DMMs is the precomputed radiance transfer (PRT) model [Sloan et al. 2002]. This uses an efficient representation (such as spherical harmonics) to approximate the local light transport at each vertex, accounting for shadowing and inter-reflection. These are precomputed but can then be used with any incident illumination at render time. Schneider et al. [2017] learn a linear model of PRT transfer matrices as a function of the 3DMM shape coefficients and use this in a rendering framework.

We denote by $\mathbf{L}$ the set of illumination parameters for any of the above illumination models.

Color transformation. In a real camera, the actual image irradiance measured by the sensor is usually transformed in a complex way in order to achieve a pleasing visual appearance. Often, this amounts to multiplication by a 3×3 color transformation matrix followed by a nonlinearity. The color transformation matrix can be decomposed into a product of three 3×3 matrices: $\mathbf{T} = \mathbf{T}_{\text{xyz2rgb}}\mathbf{T}_{\text{raw2xyz}}\mathbf{T}_{\text{wb}}$, where $\mathbf{T}_{\text{wb}}$ is a diagonal matrix that performs white balancing (compensating for the color of the illumination), $\mathbf{T}_{\text{raw2xyz}}$ is specific to each camera and maps from the native color space to the standardized XYZ space and $\mathbf{T}_{\text{xyz2rgb}}$ is a fixed matrix that transforms to sRGB space. Unfortunately, introducing such a color transformation into the 3DMM image formation model further exacerbates the lighting/albedo ambiguity by providing an additional explanation for observed color. Finally, a nonlinearity is applied, which can be approximated by $\mathbf{i}_{\text{sRGB}} = \mathbf{i}_{\text{linRGB}}^{1/\gamma}$, where usually $\gamma = 2.2$. This nonlinear transformation is important because it means the, often linear, reflectance and illumination models described above cannot explain the final image appearance.

Despite their importance, camera color transformations and nonlinear gamma are almost always ignored in the context of 3DMMs. There are some notable exceptions. Schneider et al. [2017] apply gamma correction to input images to transform back to a linear space. Blanz and Vetter [2003] estimate a per-channel scale and offset as well as scalar color contrast, allowing them to synthesize grayscale images. The same model was used by Aldrian and Smith [2013] and Hu et al. [2013].

4.3 Rendering and visibility

3DMM fitting algorithms differ in whether they synthesize a discrete image in image space (i.e. one color per-pixel) or perform rendering in object space (i.e. one color per-vertex or per-triangle). The former holds the advantage that it is straightforward to incorporate a texture model built in a high-resolution UV space and also that the output is a regular pixel grid that can be passed to CNN, for example for an adversarial loss [Shamai et al. 2020]. Methods that work in object space compute an appearance error by projecting the model vertices into the image and sampling image intensities onto the visible vertices. Visibility can also be computed in either object space or image space, with the latter usually being more efficient.

The original Blanz and Vetter [1999] paper used object space rendering in which a single color was computed for each triangle center (equivalent to flat shading) with image space z-buffering used for visibility testing. Many subsequent methods also worked in object space but usually with per-vertex colors computed using the reflectance models described above with per-vertex surface normals. This has begun to change recently when more conventional rasterization pipelines have been included in 3DMM synthesis. Rasterization associates with each pixel $(x, y) \in \mathcal{I}$, where $\mathcal{I} = \{1, \ldots, w\} \times \{1, \ldots, h\}$, a triangle index or a `NULL` value if the pixel is not covered by a triangle:

$$\mathbf{raster}_{C,\mathcal{T},\mathbf{w}^s,\mathbf{w}^e} : \mathcal{I} \mapsto \{1, \ldots, m, \texttt{NULL}\}, \tag{16}$$

recalling that $\mathbf{w}^s$, $\mathbf{w}^e$ are the shape and expression parameters respectively and $\mathcal{T}$ the mesh triangulation. Since this is a discrete function it is not smooth and not differentiable. In addition, for each pixel, three weights are calculated that are associated with the vertices of the rasterized triangle: $\mathbf{a}_{C,\mathcal{T},\mathbf{w}^s,\mathbf{w}^e}(x, y) \in \mathbb{R}^3_{\geq 0}$. These weights depend on the projected positions of the vertices

$$\mathbf{v}_{t^i_{\mathbf{raster}_{C,\mathcal{T},\mathbf{w}^s,\mathbf{w}^e}(x,y)}}, i \in \{1, 2, 3\}. \tag{17}$$

Often, these weights are barycentric coordinates of the pixel center within the triangle. These weights are a smooth function of the vertex positions and hence of the shape and camera parameters. Hence, rendering is differentiable up to a change in rasterization, i.e. so long as the triangle index associated with each pixel does not change. Tran and Liu [2018a] incorporate such a conventional rasterization pipeline into an in-network differentiable renderer.

Collecting together all of the parameters relating to the camera, illumination, face geometry and texture, $\Theta = (C, \mathcal{L}, \mathbf{w}^s, \mathbf{w}^e, \mathbf{w}^t)$, we can write the rendered appearance in object space of vertex j as $I^j_{\text{model}}(\Theta)$. For an image space rendering we denote the appearance of the model at pixel (x, y) by $I^{x,y}_{\text{model}}(\Theta)$. In the simplest case, the image space rendering is computed directly from the object space

rendering using Gouraud interpolation shading:

$$I^{x,y}_{\text{model}}(\Theta) = \mathbf{a}_{C,\mathcal{T},\mathbf{w}^s,\mathbf{w}^e}(x,y)^T \begin{bmatrix} I^{t^1_j}_{\text{model}}(\Theta) \\ I^{t^2_j}_{\text{model}}(\Theta) \\ I^{t^3_j}_{\text{model}}(\Theta) \end{bmatrix}, \tag{18}$$

where $j = \mathbf{raster}_{C,\mathcal{T},\mathbf{w}^s,\mathbf{w}^e}(x,y)$. Other rasterization strategies may be more complex. For example, Genova et al. [2018] use rasterization in a differentiable deferred shading renderer more akin to Phong interpolation shading. Here, vertex normals and colors are rasterized and interpolated, then reflectance calculations are done in image space.

Note that overcoming the non-differentiable nature of rasterization is an open problem. Hiroharu Kato and Harada [2018] present an approximately differentiable renderer based on rasterization. Liu et al. [2019a] propose a rasterizer in which triangles make a soft (and hence differentiable) contribution to image appearance. More ambitiously, differentiable rendering using other pipelines is now also being considered, for example, differentiable path tracing [Li et al. 2018].

Very recently, the explicit fixed models used in conventional rendering are being augmented or replaced by learning components, so-called *neural rendering*. For example, Kim et al. [2018a] train an image to image network that transforms a low-quality 3DMM rendering into a photorealistic video frame.

4.4 Open challenges

The image formation models used in the context of 3DMMs are much simpler than those used in graphics and many other areas of computer vision. For example, we are not aware of any work that allows for a center of projection different to the center of the image, even though many face image datasets consist of images cropped (probably non-centrally) from larger images. Similarly, nonlinear distortion is always ignored. The effect of this assumption is not understood. In other fields like structure-from-motion, it is standard to impose constraints derived from metadata, knowledge of physical camera parameters and so on. This is not currently being done to a significant extent with 3DMMs.

Advances in rendering in computer graphics are slowly propagated into the world of 3DMMs and especially into the analysis-by-synthesis process. One of the reasons is that almost every model extension makes the model adaptation more complicated and a lot of methods rely on the rendering process to be differentiable. There is a dramatic gap between what current computer graphics or also deep learning-based image generation methods are capable of and what is state of the art for 3DMMs. Also generated instances usually lack facial details like wrinkles or moles which are challenging to render properly. Recent work aims at those challenges by using generative adversarial networks as texture models [Slossberg et al. 2018] but they are not modeled in the shape and not specially treated during rendering. A possible future direction is to either model or learn the gap between current 3DMM renderings and state of the art computer graphics or real-world 2D images.

An interesting open challenge is to better exploit the constraint of the 3DMM. Existing work uses generic pipelines for tasks such as rasterization or visibility calculation. However, the geometry is defined by a low dimensional parameter vector from which the per-vertex visibility could presumably be inferred more efficiently than treating the resulting mesh as a generic shape. The attempt of [Schneider et al. 2017] to learn the relationship between PRT coefficients and shape parameters is a first step in this direction.

5 ANALYSIS-BY-SYNTHESIS

3DMMs have been widely used for image-based reconstruction. Reconstructing a 3D face from an observed image(s) involves estimating the 3DMM coefficients which can best explain the observation. This is the inverse of the image synthesis process covered in the previous section.

Analysis-by-synthesis refers to a class of optimization problems which solves this by minimizing the difference between the observed image(s) and the synthesis of an estimated 3D face. Such an optimization problem can be ill-posed with several ambiguities and multiple minima. This is a widely researched problem, with a variety of solutions exploring different input modalities (Sec. 5.1), energy functions (Sec. 5.2) and optimization strategies (Sec. 5.3). We present publicly available approaches in Table 3.

Analysis-by-synthesis techniques have also recently been used in combination with deep learning architectures for learning-based reconstruction algorithms. We will discuss these methods in Sec. 6.

5.1 Input Modalities

Analysis-by-synthesis methods have been explored using multiple image modalities, from multi-view to monocular images and videos. While multi-view methods produce very detailed and high-quality results, capturing such data requires expensive setups. A lot of recent focus has been on obtaining similar quality reconstructions with much lower cost solutions, e.g., using a single RGB image. This has also led to an increase in commercial applications for the mass market. Fitting a 3DMM to 3D scans can also be considered as analysis-by-synthesis. This is related to registration techniques, covered in Sec. 3.

Multi-View Systems. We will start our discussion with multi-view solutions which minimize the photometric consistency between the multi-view images and the synthesis of the estimated reconstruction. Most multi-view methods, such as those covered in Sec. 2 do not require a strong prior in the form of 3DMMs. However, there are several methods which use 3DMMs to aid reconstruction in stereo camera systems. Model-based stereo reconstruction was explored in Wallraven et al. [1999]. The reconstruction quality was improved by eliminating the estimation of illumination and reflectance in Amberg et al. [2007]; Fransens et al. [2005]. 3DMMs also prove to be very valuable in low-resolution settings where high-quality image textures cannot be exploited, or under occlusions [Romeiro and Zickler 2007; Thies et al. 2018b]. Most of the methods discussed here solve very large optimization problems, and are not real-time. Thies et al. [2018a] is one real-time method which has a data-parallel implementation on a GPU.

publication	input	estimates	approach	comment
Edge fitting [Bas et al. 2017b]	2D image, landmarks	pose, shape	edge features, ICP	
Eos fitting library [Huber et al. 2016]	2D image, landmarks	pose, shape	landmark and contour fitting	Huber [2017] handles expressions
Basel Face Pipeline [Gerig et al. 2018]	2D image, landmarks	pose, shape, expression, texture, illumination	MCMC Sampling	estimates posterior distribution, Egger et al. [2018] handles occlusion
Deep 3D Face Reconstruction [Deng et al. 2019]	2D image(s)	pose, shape, expression, texture, illumination	deep (ResNet)	
PRNet [Feng et al. 2018]	2D image	pose, shape	deep (convolutional)	outputs mesh in BFM topology
Expression-Net [Chang et al. 2018]	2D image	pose, shape, expression, texture	deep (ResNet)	bundles [Chang et al. 2017; Tran et al. 2017]
RingNet [Sanyal et al. 2019]	2D image	pose, shape, expression	deep (ResNet)	handles occlusion
Pix2vertex [Sela et al. 2017]	2D image	pose, shape, expression	deep + shape from shading	shape beyond 3DMM
Facial Details Synthesis [Chen et al. 2019]	2D image	pose, shape, expression, appearance	UNet for details	
3DMMs as STNs [Bas et al. 2017a]	2D image	pose, shape, expression	spatial transformer network	
3D Face Reconstruction [Tran et al. 2018]	2D image, output of [Tran et al. 2017]	shape details	estimate bump map using encoder-decoder architecture	handles occlusions
FLAME [Li et al. 2017]	2D / 3D landmarks	pose, shape, expression	landmark fitting	
Basel Face Pipeline [Gerig et al. 2018]	3D scan, landmarks	pose, shape, expression, texture	Gaussian process regression, nonrgid ICP	
LSFM Pipeline [Booth et al. 2016]	3D scan	pose, shape, expression	nonrigid ICP	fully automatic
Model Fitting [Brunton et al. 2014b]	3D scan, landmarks	pose, shape	nonrigid ICP, template and model fitting	handles occlusions
Multilinear Model Fitting [Bolkart and Wuhrer 2015a; Brunton et al. 2014a]	3D scan, landmarks	pose, shape, expression	nonrigid ICP, global model in Bolkart and Wuhrer [2015a], local model in Brunton et al. [2014a]	handles occlusions

Table 3. Overview of publicly available model adaptation and registration frameworks for 3DMMs.

Monocular RGBD. RGB-D sensors capture RGB as well as depth information of the scene. Consumer stereo cameras either use passive stereo, IR projection-mapping, or time-of-flight technology. The depth channel in the input helps in resolving depth ambiguities due to the lack of multiple views. Thus, in addition to photometric consistency, these methods also minimize depth consistencies using point-to-point and point-to-plane distances, see Sec. 3. Since monocular reconstruction methods solve a smaller optimization problem compared to multi-view methods, many real-time solutions exist [Bouaziz et al. 2013; Hsieh et al. 2015; Li et al. 2013; Thies et al. 2015; Weise et al. 2011a]. While most methods heavily rely on 3DMMs, some try to adapt them to capture user-specific details. [Weise et al. 2011a] build a user-specific expression model by adapting a general one. This is done in an offline stage before the online tracking. [Bouaziz et al. 2013; Li et al. 2013] adapt the 3DMM online, thus removing the need for an offline step. [Hsieh et al. 2015] introduced an occlusion robust tracking system using face segmentation masks. [Liang et al. 2014] reconstruct a single image by retrieving instances

of 3D shapes from a dataset and merging them, thus avoiding the need for 3DMMs.

Monocular RGB. Without the presence of the depth channel, the analysis-by-synthesis problem becomes even more ill-posed. These methods cannot easily resolve depth ambiguities. Thus, the prior knowledge of a 3DMM becomes important. Monocular RGB videos can provide more constraints. The identity component, in this case, can be estimated by fusing information from multiple frames in a preprocessing step. Many methods can track the face in real-time [Cao et al. 2015, 2014a, 2013, 2016a; Ichim et al. 2015; Thies et al. 2016]. As in the case of RGB-D based methods, there are methods which try to add details over the 3DMM reconstructions to make the results user-specific and detailed. [Garrido et al. 2016b] add medium-scale correctives based on spectral basis vectors. Cao et al. [2015]; Garrido et al. [2013, 2016b]; Shi et al. [2014]; Suwajanakorn et al. [2014] also add high-frequency wrinkle-level details. Wu et al. [2016b] use local blendshape models to capture more details compared to global blendshape based methods. [Cao et al. 2013, 2016a; Ichim et al. 2015] compute user-specific 3DMMs using images of a person performing specific known expressions.

Photo-collections, i.e., collections of images of a person can also be used to constrain the identity components of the reconstructions [Kemelmacher-Shlizerman and Seitz 2011; Liang et al. 2016; Piotraschke and Blanz 2016; Roth et al. 2015, 2016; Suwajanakorn et al. 2015]. This is a more unconstrained setting compared to multi-view images where all views are captured at the same time in the same environment. Approaches which use photo-collections and videos are more practical than multi-view images since such data is widely available for most people.

Reconstruction from a single image is the most challenging scenario. However, the original work of Blanz and Vetter [1999] already proposed an analysis-by-synthesis solution, see Fig. 5. While they required manual initialization for the optimization problem, several approaches made the approach more robust to enable automatic reconstruction [Aldrian and Smith 2011a; Bas et al. 2017b; Egger et al. 2018; Fried et al. 2016; Hu et al. 2017b; Kortylewski et al. 2018c; Paysan et al. 2009b; Schneider et al. 2017; Schönborn et al. 2017; Tewari et al. 2018]. Most analysis-by-synthesis approaches evaluate the photometric consistency between the observations and the estimates. Some approaches have explored the use of other image features, such as edges, or SIFT [Booth et al. 2017; Romdhani and Vetter 2005], in order to obtain higher fidelity reconstructions. Occlusion robust reconstruction by jointly solving for segmentation has been explored in Egger et al. [2018]. Monocular reconstruction methods primarily differ in their formulated energy functions. We will look at these in detail in Sec. 5.2.

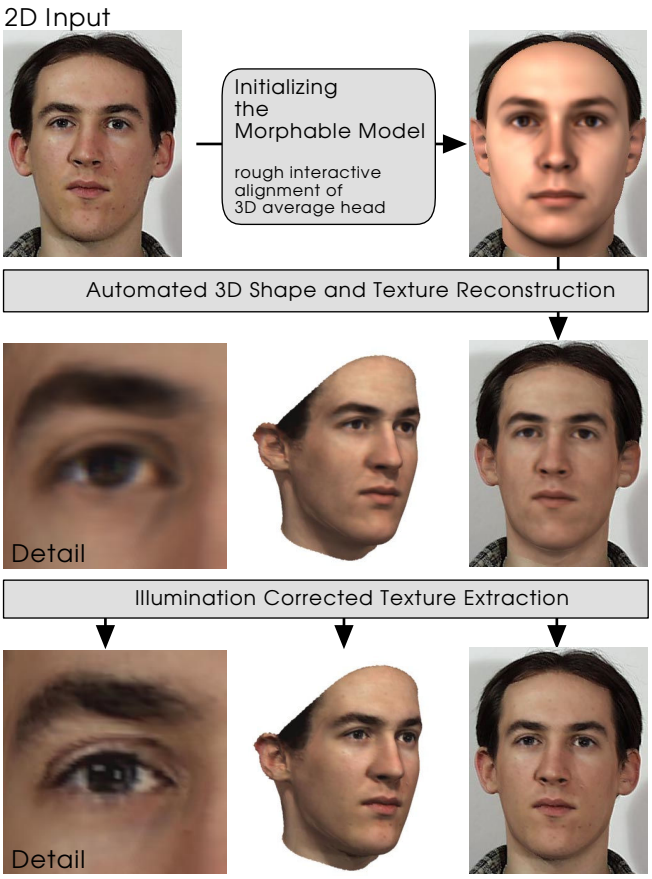


Fig. 5. The analysis-by-synthesis pipeline used by Blanz and Vetter [1999] for reconstruction from a single image. The different steps include initialization, optimization, and refinement of the optimized 3DMM texture.

5.2 Energy Functions

The analysis-by-synthesis paradigm involves the solution of a non-linear optimization problem made up of a number of energy functions. Methods differ in their combination and precise design of these energy functions, their relative weights and (dealt with in the following subsection) the optimization strategy used to minimize the energy. Here we describe the most commonly used energy functions for fitting to RGB images. For single image reconstruction, the energy functions are expressed in terms of a single set of unknown parameters Θ. In the case of multi-view images of a static face, the camera parameters are indexed by viewpoint while all others are fixed across views. In the case of an image sequence of a dynamic face, camera, expression and lighting parameters are indexed by frame while neutral shape and texture parameters are fixed throughout the sequence.

Appearance error. The key ingredient of analysis-by-synthesis is to measure the difference between observed data and a synthesis using the model. Most directly, this is the appearance error between an input image and the rendered face. A number of variants of this term have been used. The *pixel-wise* formulation sums the appearance error over the pixels of the image, necessitating rasterization of the model:

$$E^{\text{pixel}}_{\text{appearance}}(\Theta, I_{\text{obs}}) = \sum_{(x,y)\in\textbf{foreground}} \left\| I_{\text{obs}}(x,y) - I^{x,y}_{\text{model}}(\Theta) \right\|^2 \tag{19}$$

where $\textbf{foreground} = \{(x,y) \in \mathcal{I} | \textbf{raster}_{C,\mathcal{T},\mathbf{w}^s,\mathbf{w}^e}(x,y) \neq \texttt{NULL}\}$ is the set of pixels covered by the union of all triangles. This formulation naturally weights the contribution of model vertices in terms of their contribution to the appearance of a pixel. An alternative is to compute the appearance error *vertex-wise* by rendering in model space and sampling the image intensities onto the vertices:

$$E^{\text{vertex}}_{\text{appearance}}(\Theta, I_{\text{obs}}) = \sum_{j \in \textbf{visible}} \left\| \textbf{interp}[I_{\text{obs}}, \textbf{project}[C, \mathbf{v}_j(\mathbf{w}^s, \mathbf{w}^e)]] - I^j_{\text{model}}(\Theta) \right\|^2 \tag{20}$$

where $\textbf{interp}[X,(x,y)]$ represents differentiable interpolation of 2D object X at location (x,y) and **visible** is the set of visible vertices. A common variant of this approach uses a random subset of vertices rather than all of them. This is more efficient, introduces stochasticity that may help avoid local minima and avoids overly conservative fits near the boundary where background may be sampled. Differentiable interpolation of the image can either be done with explicit differentiable sampling (e.g., bilinear sampling as in Jaderberg et al. [2015]) or precomputing the image gradient and then interpolating this along with the image intensities (this was done in the original Blanz and Vetter [1999] paper). A drawback of the vertex-wise error is that regions of the image with dense coverage from projected vertices are weighted more heavily than more sparsely sampled regions. This can be overcome by using weights related to projected area. Blanz and Vetter [1999] accomplished the same effect by using the triangle area as the probability by which the triangle would be selected in their random sampling. For multi-image methods, the above energies are simply summed over each image.

Feature-based energies. There are other, less direct, ways to compute an error between the observed data and model. This is done by first computing features from observed data and then measuring the difference between those features and the corresponding ones in the model. By far the most commonly used features are landmarks (alternatively known as keypoints or fiducial points) which often used for initialization and are still important in much of the state-of-the-art, e.g., [Sanyal et al. 2019]. A landmark detector returns a set of 2D landmark coordinates $\{\mathbf{x}_j\}_{j=1}^J$ with $\mathbf{x}_j \in \mathbb{R}^2$. As a one-off procedure, each landmark is associated with the corresponding vertex in the 3DMM such that the jth landmark corresponds to vertex index $k_j \in \{1,\ldots,n\}$. The reprojection error of the model landmarks with respect to detected positions is then given by:

$$E_{\text{landmarks}}(\Theta, \{\mathbf{x}_j\}_{j=1}^J) = \sum_{j=1}^{J} \left\| \mathbf{x}_j - \textbf{project}[C, \mathbf{v}_{k_j}(\mathbf{w}^s, \mathbf{w}^e)] \right\|^2. \tag{21}$$

Sometimes the landmarks are allowed to slide on the face surface such that each landmark has a set of vertices to which it could correspond [Zhu et al. 2015].

Edges directly convey geometric information about occluding boundaries and texture edges. Misalignments between model and image edges seriously degrade the perceptual quality of a reconstruction and lead to the wrong part of the face, or the background, being sampled onto the mesh. Moghaddam et al. [2003] were the first to exploit this cue by fitting to multi-view silhouettes. Romdhani and Vetter [2005] computed the distance transform of detected edges in an input image providing a distance-to-edge cost surface that was sampled at projected positions of vertices lying on model texture edges or the occluding boundary. Amberg et al. [2007] extended this to multiple views and improved robustness by averaging the cost surface over different parameters of the edge detector. Keller et al. [2007] showed that these cost functions are neither continuous nor differentiable. Bas et al. [2017b] transformed edge fitting into landmark fitting by alternating between computing an explicit correspondence between edge pixels and model edges and minimizing the resulting landmark energy. Sánchez-Escobedo et al. [2016] directly regress shape parameters from a set of multi-view occluding contours.

Finally, some other features have been considered. Romdhani and Vetter [2005] used the position of specularities in the image to constrain the surface normal direction at the corresponding location on the model via a specular reflection model. Booth et al. [2017] and Booth et al. [2018b] compute dense SIFT features from the input image and compare these to the SIFT features on which their statistical texture model is built in a similar fashion to the vertex-wise appearance error above.

Background Modeling. A common challenge when optimizing for pose and shape is the varying visibility of vertices for *vertex-wise* errors and the varying number of pixels covered by the face for *pixel-wise* errors. This leads commonly to the undesired effect of shrinking. Having the model covering fewer pixels or having fewer vertices visible leads to an undesired local optimum of most error terms. Common strategies to overcome this are fixed visibility, restrictive regularization, relying on landmarks, enforcing edge or contour terms or explicit image segmentation. Schönborn et al. [2015] demonstrated the problems with an implicit background model which is present in all error formulations and have shown that even simple background models b like a constant, a Gaussian or an image histogram-based model can solve this issue. The background model can easily be added to the existing formulations, e.g., for the *pixel-wise* formulation as:

$$E^{\text{image}}_{\text{appearance}}(\Theta, I_{\text{obs}}) = E^{\text{pixel}}_{\text{appearance}}(\Theta, I_{\text{obs}}) + \sum_{(x,y) \in \textbf{background}} b(I_{\text{obs}}(x,y)). \tag{22}$$

Occlusions and Segmentation. Occlusion of faces by other objects, that are not part of the generative model, are a common challenge for the so far presented error terms and for analysis-by-synthesis in general. There are various methods presented on how to identify occlusions. Those methods range from appearance-based methods [De Smet et al. 2006; Pierrard 2008] to detection [Morel-Forster 2016] and segmentation-based methods [Egger et al. 2018; Saito et al. 2016]. They share the basic idea, that occluded pixels are excluded from the model evaluation:

$$E^{\text{semantic}}_{\text{appearance}}(\Theta, I_{\text{obs}}) = \sum_{l \in \textbf{label}} \sum_{(x,y) \in R(l)} E^{\text{pixel}}_{\text{label}}(\Theta, I_{\text{obs}}(x,y), l), \tag{23}$$

where each $E^{\text{pixel}}_{\text{label}}$ is a separate model per label, and $R(l)$ is the image region covered by label l. Those labels could e.g., be face, occlusion and background or also contain more detailed labels like beards. Whilst the segmentation is based on detection and fixed in Morel-Forster [2016]; Pierrard [2008]; Saito et al. [2016], other methods solve for segmentation and model parameter estimation jointly in an Expectation-Maximization-based manner [De Smet et al. 2006; Egger et al. 2018].

Priors. A 3DMM is a statistical model and so provides a natural probabilistic prior over the parameter space. Under the assumption that the original data is Gaussian distributed the natural cost function to express this prior for either the shape or texture model is:

$$E_{\text{prior}}(\Theta) = \sum_{i=1}^{d} \frac{w_i^2}{\sigma_i^2}, \tag{24}$$

where σ_i^2 is the variance associated with the ith principal component. The drawback to this prior is that it is minimized by the mean face and, if weighted heavily, leads to model dominance where recovered faces are too close to the average. There has been in discussion in the literature [Lewis et al. 2014b; Patel and Smith 2016] as to whether this prior is appropriate in high dimensional space and alternatives have been considered, as will be described next.

One class of techniques allows reconstructed shape and/or texture to deviate from the 3DMM subspace enabling recovery of fine-scale detail not captured by the model. Allowing arbitrary shape or albedo changes transforms the problem into classical shape-from-shading and becomes highly ill-posed. For this reason, additional generic priors are used. Patel and Smith [2012] use a piecewise smoothness prior on per-vertex diffuse albedo which is allowed to vary per-vertex along with surface normals to satisfy a shape-from-shading constraint. This is regularized using the squared vertex distance between the updated shape and the closest shape in the 3DMM space. Richardson et al. [2017] use the same regularization, though, expressed in terms of per-pixel depth. To ensure smoothness, they also use the L1 norm of the discrete Laplacian of the depth map. The L2 norm of the mesh Laplacian has also been used as a smoothness prior [Garrido et al. 2016a; Tewari et al. 2018]

When reconstructing a dynamic face from video, parameters can either be assumed fixed (if identity dependent) or smoothly varying (pose, expression, lighting). These latter parameters can, therefore, be regularized with generic temporal smoothness priors. A common and simple way to express this prior is to initialize each frame with the estimate from the previous one. This encourages convergence to a local minimum close to the solution for the previous frame. More sophisticated priors have also been considered. For example, Cao et al. [2013]; Weise et al. [2011a] build a Gaussian mixture model over expression parameters from the previous k frames. This model is then used to regularize the estimate for the current frame.

5.3 Optimization

From the perspective of optimizing the energy functions above, there are a number of significant challenges. First, most of the energy terms are nonconvex in theory and we observe in practice that there are many local minima. Second, the appearance error is not even continuous due to rasterization/vertex visibility and shadowing all being noncontinuous functions. Third, the appearance error has a small basin of convergence. When a model feature is completely misaligned to the image (or in the extreme case, the whole model aligned entirely to background), the gradient of the appearance error conveys no useful information. Fourth, all parameters have global influence. Fifth, computing the appearance error and its gradient is computationally expensive, amounting to the rendering of an image. For these reasons, a significant effort has gone into the selection of optimization algorithms and engineering of the optimization schedule to develop methods that are sufficiently fast and robust.

The majority of existing approaches optimize based on gradient information of the energy function. The original Blanz and Vetter [1999] approach used first-order gradient descent, as have other more recent methods [Bouaziz et al. 2013; Fried et al. 2016; Ichim et al. 2015]. Since they computed the appearance error over only a small subset of randomly selected triangles, this is strictly *stochastic* gradient descent (SGD). An interesting parallel here is that modern deep learning-based methods (see Section 6) are usually trained with SGD and use similar energy functions so they are learning from the same signal used in the original method.

Since the energy terms above can easily be formulated as nonlinear least-squares problems, specialized pseudo-second-order methods like Gauss-Newton or Levenberg-Marquardt have often been used [Garrido et al. 2013, 2016a,b; Romdhani and Vetter 2005; Thies et al. 2015, 2016]. Booth et al. [2017] use a "*project-out*" strategy in which appearance parameters are implicitly solved in a least squares sense and optimization takes place only over geometric parameters. General pseudo-second-order methods such as BFGS have been used [Cao et al. 2013; Weise et al. 2011a] as well as genuine second-order methods, specifically a stochastic variant of Newton's method [Blanz and Vetter 2003]. As the problem size increases, as in the case of shape-from-shading, gradient descent becomes the most common optimization approach [Garrido et al. 2016a; Shi et al. 2014; Suwajanakorn et al. 2014; Tewari et al. 2018]. In all the above methods, the discontinuity of the appearance function is dealt with by fixing rasterization/visibility when computing gradients or even keeping them fixed for a certain number of iterations. Importantly, this means that the gradient cannot convey information about a change in visibility. Many other tricks have been considered, for example, hierarchical optimization (both in parameter space and spatially, i.e. multiresolution [Thies et al. 2016]) and using an optimization schedule in which different energy functions are switched on or weighted differently at different phases on the optimization [Blanz and Vetter 1999].

Several approaches have decomposed the energy terms into several smaller, often linear, problems (sometimes with closed-form solutions) that can be solved efficiently and in sequence [Aldrian and Smith 2010, 2011a, 2013, 2011b; Bas et al. 2017b; Cao et al. 2014a, 2013; Hu et al. 2017c; Romdhani et al. 2002; Saito et al. 2016; Zhu et al. 2015]. These alternating approaches are usually very efficient but not guaranteed to obtain the optimum solution that comes from optimizing all parameters simultaneously.

Gradient-based methods are typically initialized by fitting only to landmarks, i.e. to optimize the landmark energy in isolation. Originally, the landmark positions were provided manually but

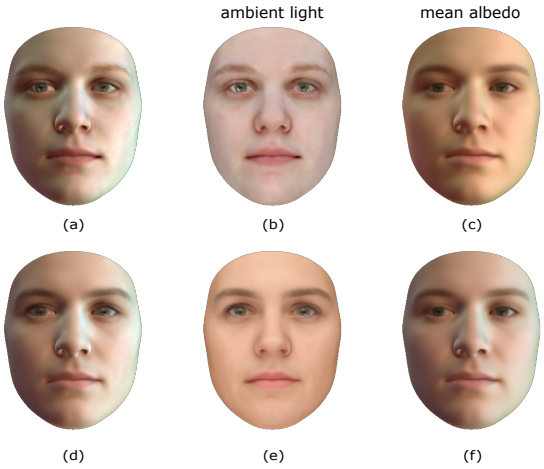


Fig. 6. An example of the albedo-illumination ambiguity presented by Egger [2018]. The target image in the first row (a), its color rendered under ambient illumination (b) and its illumination rendered on the mean albedo of the Basel Face Model (c). The second row shows a model instance with different color (e) and illumination (f) parameters but very similar appearance (d).

combining with an automatic landmark detector provided fully automatic methods [Breuer et al. 2008]. From a landmark detector that outputs many hypothesized landmark locations, including many false positives, Amberg and Vetter [2011] use Branch and Bound to select the subset configuration of landmarks that is most consistent with the 3DMM. Bas and Smith [2019] show how to express the landmark energy as a separable nonlinear least squares problem.

While gradient-based methods are widely used mainly due to computational efficiency and ease of implementation, these methods are sensitive to initialization and often end up in local minima. Probabilistic methods based on Bayesian inference were proposed to deal with these limitations [Egger et al. 2018; Kortylewski et al. 2018c; Schneider et al. 2017; Schönborn et al. 2017]. These methods do not require any gradient computation of the energy terms to update the estimates. They are stochastic and thus, less susceptible to getting stuck in local minima. Different from optimization-based methods which only provide a single solution, these approaches approximate the full posterior distribution and thus provide access to a manifold of possible solutions.

5.4 Open challenges

Reconstructing 3D shape and albedo from a 2D image is an ill-posed problem. Ambiguities like the perspective face shape ambiguity [Smith 2016] and the albedo illumination ambiguity [Egger 2018] have been demonstrated (see Figure 6). These ambiguities can not be resolved completely and priors are our best approach to at least find a reasonable estimation. They are the major reason why there is a huge gap between the estimates we can get from multi-view and 3D data vs. from monocular images. Even in the state-of-the-art, it is often evident that overall skin color is explained using the lighting while the albedo colors are similar for very different skin types [Tewari et al. 2018]. This is somewhat improved by discriminative methods that do not need to synthesize the same appearance as a given image, only an image with the same identity [Genova et al. 2018], thereby sidestepping explicit estimation of illumination and camera parameters. Reporting the geometric errors obtained by the model mean is not common. Only three papers demonstrated their 3D reconstructions to be closer in mesh distance to the ground truth face compared to the model mean [Aldrian and Smith 2013; Sanyal et al. 2019; Schönborn et al. 2017].

Current state of the art techniques also lack dramatically in accuracy across pose and in terms of matching the contours and edges. It is very difficult to evaluate these beyond qualitative evaluation which makes it difficult to compare different approaches. Recently a first benchmark with natural images and ground truth shape was published and will help to better compare competing methods [Sanyal et al. 2019]. However, as methods get more accurate, the mesh distance errors get close to the range of error in computing "ground truth" using multi-view methods. This makes it difficult to quantitatively compare different approaches.

Another challenge which is usually neglected are occlusions. Faces are mostly occluded by objects which are frequently in front of faces like glasses, cigarettes, hands or microphones, but can also be occluded by virtually every other object. Analysis-by-synthesis methods fail when they do not explicitly model occlusions. Furthermore, reconstruction methods based on 3DMMs are limited to the space of faces covered by the models. A lot of residual error in the results stems from the fact that 3DMMs do not model detailed and high-frequency geometry and texture. Furthermore, most approaches use simple lighting models which cannot explain many in-the-wild images. These limitations are also shared with the learning-based methods which use analysis-by-synthesis in their pipeline, see Sec. 6.

Recent techniques have been focused on reconstruction from a single face individually. The aim of face image analysis would, however, go beyond interpreting a single face or each face separately. We would like to analyze and interpret interactions between people and perhaps also ease the analysis task by exploiting scene constraints, such as shared illumination parameters to deal with albedo-illumination ambiguity, or constraints on the perspective face shape ambiguity by analyzing multiple faces jointly.

6 DEEP LEARNING

So far, we have mainly discussed classical face modeling and parameter estimation techniques based on optimization-based inverse graphics. We now discuss how these processes can be replaced by or combined with deep learning, see Figure 7. There are a number of reasons for wanting to do this. On the modeling side, the use of nonlinear, deep representations offers the possibility to surpass classical linear or multilinear models in terms of generalization, compactness and specificity [Styner et al. 2003]. On the parameter estimation side, we can exploit the speed and robustness of deep networks to achieve reliable performance on uncontrolled images.

We begin by discussing deep modeling and deep model fitting before finally discussing methods that simultaneously learn both the model and how to fit it within a single deep network.

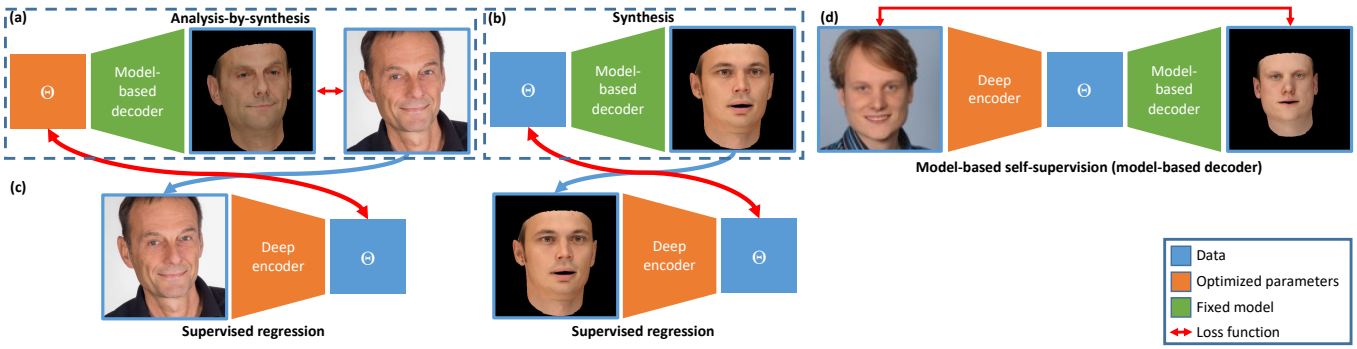


Fig. 7. The relationship between classical analysis-by-synthesis and deep learning approaches. (a) analysis-by-synthesis, e.g., [Blanz and Vetter 1999]. (a)+(c) training a regressor based on the output of an analysis-by-synthesis algorithm, e.g., [Tran et al. 2017], (b)+(c) training a regressor using synthetic data generated by a model, e.g., [Richardson et al. 2016], (d) self-supervision, e.g., [Tewari et al. 2017].

6.1 Deep Face Models

The traditional modeling techniques discussed in Section 3 aim to represent face shape, expression, and appearance as vector $\mathbf{w}$ in a low-dimensional latent space $\mathbb{R}^d$. The projection into (respectively reconstruction from) this latent space is defined by linear or multi-linear operations, and can be thought of as encoding (respectively decoding) the high-dimensional information in $\mathbb{R}^d$. Deep learning provides a new tool for building 3DMMs using nonlinearities both in the encoder and the decoder. This way of building morphable models is currently a very active area of research.

We can see the relationship between the encoder and decoder learned using deep learning and classical works using the example of linear models commonly used for shape and texture modeling. In the context of deep learning, such a linear model formalized in Equation (2), is exactly equivalent to a fully connected layer in a neural network. Concretely, the parameter vector $\mathbf{w}$ plays the role of the input features, the principal components $\mathbf{e}_j$ are the weights and the mean $\bar{\mathbf{c}}$ is the bias. This can be viewed as *decoding* from the latent parameter space to the data space $\mathbf{c}$. Projection onto the model can similarly be viewed as *encoding* with a fully connected layer in which the input features are the data, the weights are the rows of the transposed principal component matrix and the biases are given by $-\mathbf{e}_j^T \bar{\mathbf{c}}$. Concluding the analogy, a PCA can be accomplished by combining the encoder and decoder as a linear autoencoder with a single hidden layer. Such an autoencoder with d neurons in the hidden layer will learn a latent space with the same span as a d dimensional PCA, though without the guarantee of orthogonality (though this could be ensured with appropriate loss functions).

Given this close relationship between classical methods and deep learning, it is natural to ask if there exist more powerful nonlinear models that can be trained based on current advances in deep neural networks. As in the classical work, this has been considered for the 2D case. Duong et al. [2019] propose a deep appearance model for 2D facial images that extends 2D AAMs to model nonlinearities. This is achieved using deep Boltzmann machines to model 2D shape and texture information. For modeling 3D faces, first successful models using autoencoders, GANs, and hybrid structures have been proposed, as detailed in the following.

Fernández Abrevaya et al. [2018] proposed the first encoder-decoder architecture to model the 3D geometry of faces. The encoder first projects the 3D face to a 2D image and uses a standard image-based encoder, while the decoder is fixed to a classical tensor-based face model. This allows decoupling shape variations caused by identity and expression. Bagautdinov et al. [2018] introduced a VAE that models different levels of detail of facial geometry by representing global and increasingly localized shape variations in different layers of the network. The 3D geometry is again represented using a two-dimensional mapping, and convolutions are performed in the image domain. This work allows representing highly detailed geometric information in latent space. Lombardi et al. [2018] extend this work to jointly encode variations in appearance and geometry, for the application of highly detailed facial rendering from novel viewpoints. Ranjan et al. [2018] proposed the first autoencoder architecture for the geometry of faces that performs convolutions in 3D mesh space directly instead of going through a 2D image representation. The model, named CoMA, allows for very compact representations of the facial geometry. This work was recently extended to encode both texture and shape information jointly [Zhou et al. 2019].

An alternative line of work considers learning GANs for 3D face modeling. Slossberg et al. [2018] proposed the first 3DMM using GANs. In this work, the facial texture is mapped to a coherent 2D image domain, and two-dimensional convolutions are employed to build a GAN of facial texture. This is combined with a standard PCA-based 3DMM for facial geometry, where for a generated face texture, a suitable PCA-based geometry is computed. Recently, multiple methods were proposed to generate 3D facial geometry, possibly with texture information. Fernández Abrevaya et al. [2019] proposed to train a GAN for the geometry of 3D faces that is able to decouple different factors of variation such as identity and expression. Shamai et al. [2020] proposed a GAN architecture to generate both facial geometry and texture, with a focus on highly detailed texture information by mapping the face to a unit rectangle. Cheng et al. [2019] proposed the first intrinsic GAN architecture that operates directly on 3D meshes. As in the case of 2D images, GANs are generally able to generate more detailed and realistic 3D faces than autoencoders at the cost of being more difficult to train.

Finally, hybrid structures can be effective to learn nonlinear 3DMMs. Tran and Liu [2018a] jointly learn a 3DMM and 3D reconstruction from a 2D image using a differentiable renderer in the training loss, see also Section 6.3. The network takes as input a 2D image and encodes it into projection, shape and texture parameters. Two decoders are then used to infer 3D shape and texture, respectively. Wang et al. [2019b] proposed an adversarial auto-encoder structure that allows disentangling factors of variation such as identity, expression, or pose of 2D facial images, and that is trained in an unsupervised way. While the method's input and output are 2D images, the 3D geometry of the face can be reconstructed.

Recently, appearance modeling approaches based on deep learning have also been proposed. The rise of deep learning methods facilitated to learn per-vertex appearance models directly from images, such as done by Tewari et al. [2018], who learn per-vertex albedo model offsets in order to improve the generalization ability of an existing PCA-based model. Similarly, Tewari et al. [2019], learn a per-vertex albedo model from scratch based on video data. Zhou et al. [2019] train a mesh decoder that jointly models the texture and shape on a per-vertex basis, which, however, relies on the availability of 3D shape and appearance data. There are also several deep learning approaches that consider a texture-based appearance modelling. Without the need of 3D data, Tran and Liu [2018a] learn a nonlinear facial appearance model represented in *uv*-space based on CNNs, which, however, does not explicitly consider lighting. In follow-up work, the authors considered a more elaborate model where the albedo and the lighting is separately modeled [Tran et al. 2019; Tran and Liu 2018b]. Moreover, a range of generative methods that synthesize facial textures have been proposed, e.g., by Saito et al. [2017], Slossberg et al. [2018], Deng et al. [2018], Lombardi et al. [2018], Nagano et al. [2018] and Yamaguchi et al. [2018]. Gecer et al. [2019b] use GAN-based texture model for the task of 3D face reconstruction, and Nagano et al. [2019] use GAN-based texture models for the task of face normalization.

6.2 Deep Face Reconstruction

In the following, we discuss dense monocular face reconstruction approaches that are based on deep neural networks. We discuss requirements on the used training data, as well as different training strategies. Let us first have a closer look at the reconstruction problem, Blanz and Vetter [1999] tackle monocular face reconstruction by fitting a parametric model based on an optimization approach, i.e., gradient descent. Deep learning approaches follow a similar optimization strategy, but instead of solving the optimization problem at 'test' time, they for example train a parameter regressor based on a large dataset of training images, see Figure 7. The regressor can be interpreted as an encoder network that takes a 2D image as input and outputs the low-dimensional face representation. Learned encoders can be combined with decoders based on classical face models to give rise to end-to-end encoder-decoder architectures. This methodology is widely-used and enables the fusion of classical model-based and deep learning approaches.

6.2.1 Supervised Reconstruction. Supervised regression approaches are trained based on paired training data, i.e., a set of monocular images and the corresponding ground truth 3DMM parameters. One of the essential questions here is how to efficiently obtain the ground truth for such a supervised learning task. In the following, we will categorize the approaches based on the type of employed ground truth training data.

One option would be to let users annotate the ground truth. While this is a popular strategy, which is often employed for sparse reconstruction problems [Saragih et al. 2011], the accurate annotation of dense geometry, appearance, and scene illumination is almost intractable. A related approach is for example employed in the work of Olszewski et al. [2016], where three professional animators manually created the blendshape animation to match a video clip.

For dense reconstruction tasks, some approaches [Laine et al. 2017] are trained based on images captured in a controlled multi-view capture setup. Thus, ground truth can be obtained by a multi-view reconstruction approach followed by fitting a 3DMM to the resulting 3D data. Normally, the ground truth is of very high quality, but the distribution of the captured monocular images does not match in-the-wild data, which can lead to generalization problems at test time.

The approach of Tran et al. [2017] performs monocular reconstruction for multiple images of the same person and computes a consolidated face identity based on simple averaging of the 3DMM parameters.

Currently, many approaches [Feng et al. 2018; Kim et al. 2018b; Klaudiny et al. 2017; McDonagh et al. 2016; Richardson et al. 2016; Sela et al. 2017; Yu et al. 2017] in the research community are trained on synthetic training data, since it is easy to acquire and comes by design with perfect annotations. Given a face 3DMM, random identities and expressions can be sampled in parameter space. Afterward, the models can be rendered under randomized illumination conditions and from different viewpoints to create the monocular images. Often, background augmentation is employed by rendering the generated faces on top of a large variety of real-world background images. Since all the parameters are controlled, they are explicitly known and can be used as ground truth. While it is easy to get access to synthetic training data, there is often a large domain gap between synthetic and real-world images, which severely impacts generalization to real images. For example, hair, facial hair, torsos, or mouth interiors are often not modeled at all. One possibility to counteract this problem in the future would be better models that include all these components.

To leverage the advantages of both real as well as synthetic training data, many current approaches [Kim et al. 2018b; Richardson et al. 2017] are trained on a mixture of data from these two domains. The hope here is that the approach learns to deal with real-world images, while the perfect ground truth of the synthetic training data can be used to stabilize training. One interesting variant of this is self-supervised bootstrapping [Kim et al. 2018b] of the training corpus. Other approaches that can be trained without requiring ground truth data are presented in the next sections.

6.2.2 Self-Supervised Reconstruction. Supervised training of a convolutional neural network requires an annotated dataset. Most of the methods we have discussed so far use such datasets, either synthetic or real. Recently, some approaches explored self-supervised learning i.e., training on real image datasets without any 3D labels.

This was made possible by a combination of analysis-by-synthesis (Sec. 4) and deep learning techniques. Tewari et al. [2017] introduced a model-based encoder-decoder architecture, which replaces the trainable decoder with an expert-designed fixed decoder. This expert-designed decoder takes the 3DMM parameters (latent code) predicted by an encoder as input and transforms it into a 3D reconstruction using the 3DMM. It further renders a synthetic image of the reconstruction using a differentiable renderer. Extrinsic parameters required for rendering are also predicted by the encoder. The loss function used is very similar to those used in analysis-by-synthesis (Sec. 5.2), consisting of photometric alignment and statistical regularization. We can think of such a technique as a joint analysis-by-synthesis optimization problem over a large training dataset, instead of a single image, see Figure 7. This allows for training a parameter regressor without any 3D supervision. This concept, usually in combination with supervised synthetic data has also been explored using higher-level loss functions like identity preservation [Genova et al. 2018; Sanyal et al. 2019], or perceptual and adversarial losses [Tran et al. 2017]. Gecer et al. [2019b] employ GANs in combination with differentiable rendering to learn a powerful generator of facial texture. [Richardson et al. 2017; Sengupta et al. 2018] refine 3DMM predictions for higher quality or more detailed results. [Deng et al. 2019; Sanyal et al. 2019] extend the network architecture to allow for training using multiple images of a person as constraint. Bas et al. [2017a] use a 3DMM as a spatial transformer network such that model fitting is learned as a by-product of solving a downstream task.

6.3 Joint Learning of Model and Reconstruction

Model-based encoder-decoder networks consist of a trainable encoder and a fixed decoder, where the decoder implements a 3DMM. However, the 3DMM itself could be trainable. We could simply update its values using the gradients from the loss function. This would allow face model learning using only 2D supervision. Learning 3D models entirely from 2D data was first shown in [Cashman and Fitzgibbon 2012] without the use of deep learning. Several deep learning approaches have explored refining an existing 3DMM using large image datasets [Lin et al. 2020; Tewari et al. 2018; Tran et al. 2019; Tran and Liu 2018b]. Nonlinear convolutional decoders have also be used to build nonlinear face models [Tran et al. 2019; Tran and Liu 2018b]. Models learned from 2D data are more generalizable to different identities, as the image datasets contain significantly more identities compared to the 3D datasets used to compute 3DMMs. Recently, an extension of the model-based encoder-decoder architecture was used to learn the identity component of a face model from videos [Tewari et al. 2019].

6.4 Open Challenges

Applying deep learning to the analysis of 3D face data is an active research topic that the community has only started to explore during the past few years, with many ongoing advances. Hence, many challenges currently remain to be solved. The most pressing ones include analyzing the limitations of current methods and providing comprehensive comparisons. This includes a clear analysis of the methods' tendency to overfit, especially when mostly synthetic data is used for training and the interpretability of the learned representations. It also includes a clear analysis of whether training in the 2D or 3D domain offers clear benefits for different applications.

It is interesting that deep learning methods are learning from essentially the same energy functions as classical methods using similar optimization approaches (e.g., stochastic gradient descent). The difference is that backpropagation updates are averaged over batches and whole datasets, seemingly alleviating problems of local minima or overfitting to a single sample. The problem then becomes overfitting to the *distribution* of faces in the training set. The training data used in these learning-based methods are often biased (e.g., Liu et al. [2015] includes mostly smiling faces). This leads to biases in the reconstruction methods. A practical question that requires to be solved is to determine the minimum amount of data required to apply deep learning methods. This is important when high-quality data is used for supervised training.

As learning-based and analysis-by-synthesis methods come together through self-supervised reconstruction methods, there are many shared challenges such as perspective face shape ambiguities and dealing with occlusions (e.g., Tran et al. [2018] already did a first step in this direction). Learning-based methods typically are very fast and robust to initialization but achieve lower quality results compared to analysis-by-synthesis methods. One way to combine the desirable properties of these different paradigms is to use the learning-based solution as initialization for analysis-by-synthesis optimization [Tewari et al. 2018].

While some recent methods have tried to build 3DMMs just from 2D data for better generalization, the resulting models are not as high-quality and lack details due to the low resolution of faces in currently available in-the-wild images. Bridging the gap in terms of details between models trained using high-quality data, and those built using only 2D data is an important open challenge.

Other challenges include extending recently developed methods to new applications. For instance, while monocular face reconstruction has started being explored, there is not yet much work on reconstructing a coherently deforming facial geometry from 2D video data.

7 APPLICATIONS

Parametric face models enable many compelling applications. In the following, we will discuss applications in the domains of face recognition, entertainment, medical applications, forensics, cognitive science, neuroscience, and psychology. All these applications have been pushed by the availability of publicly shared models and code (see Tab. 2), as well as other resources [Community 2019].

7.1 Face Recognition

In the context of face recognition, 3DMMs have a manifold of potential applications. Blanz et al. [2002] proposed to perform face recognition using the cosine angle on the shape and color coefficients estimated from a pair of 2D images as a distance metric for identification and recognition. This distance metric exploits the natural disentanglement of 3DMMs separating identity (shape and color) from camera and illumination variation. It was shown that this 3DMM-based distance metric enables to recognize faces across

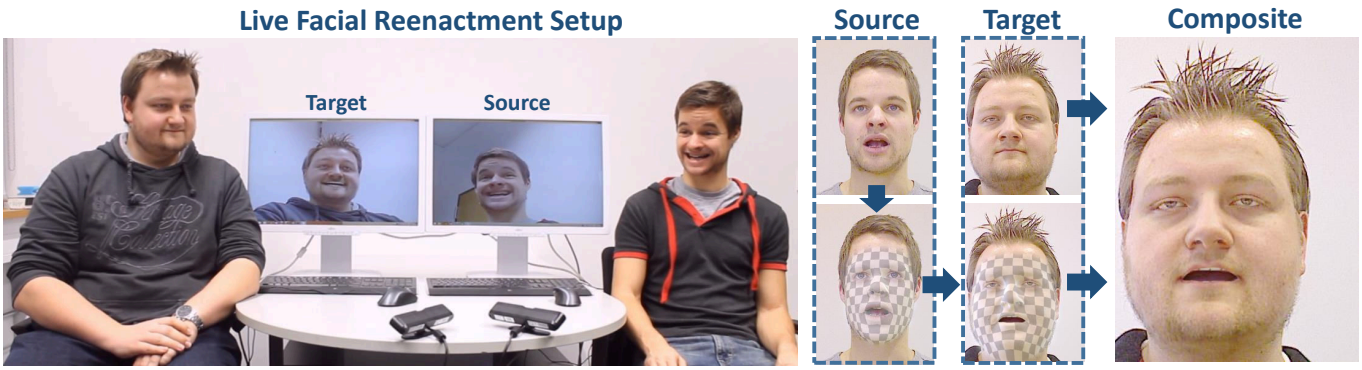


Fig. 8. The first real-time facial reenactment approach [Thies et al. 2015] was based on RGB-D sensors. The approach tracks the facial expressions of a source and target actor, transfers the expression from source to target, and re-renders the target actor with the new expression on top of the input video stream.

large pose and illumination variations [Blanz et al. 2002; Blanz and Vetter 2003; Paysan et al. 2009a], while being robust to facial expressions [Gerig et al. 2018], as well as being able to do recognition from features in the facial texture [Pierrard and Vetter 2007]. Recently, Tran et al. [2017] have shown that the performance of face recognition with 3DMM parameters can be enhanced by specifically taking the face recognition task into account when regressing the 3DMM parameters. Whilst most work focuses on face recognition from 2D images, the 3DMM was also applied to the 3D face recognition task focusing on the shape coefficients and robust recognition with respect to facial expressions [Amberg et al. 2008; Paysan et al. 2009a; ter Haar and Veltkamp 2008].

Although 3DMMs have shown promising results at face recognition in controlled settings, they did not achieve a convincing performance on in the wild data. This arises from the ill-posed problem of estimating shape and color parameters from a 2D image, at the same time a high precision of this estimation is needed for face recognition. Therefore, purely data-driven approaches have remained the dominant approach to face recognition, in particular since the advancement of deep learning technology [Parkhi et al. 2015; Schroff et al. 2015; Taigman et al. 2014]. However, data-driven approaches have fundamental problems such as their dependence on large-scale training data and their lack of generalization to out-of-distribution samples [Klare et al. 2012]. One of the main issue for face recognition is the alignment of images. Careful alignment of face images has a big impact on face recognition accuracy and even for state of the art deep learning systems. 3DMMs are particularly useful for tackling these limitations, e.g., by using 3DMMs as a tool for face frontalization [Blanz et al. 2005; Hassner et al. 2015; Tena et al. 2007]. In this context, it was shown that 9 out of 10 2D algorithms in the Face Recognition Vendor Test 2002 [Phillips et al. 2003] improved considerably when combined with a 3DMM for face frontalization [Blanz et al. 2005]. Other applications of 3DMMs include augmenting real-world data in 3D [Masi et al. 2016] and the generation of synthetic data for training [Kortylewski et al. 2018b; Sela et al. 2017] and for analyzing the effects of dataset bias on face recognition systems [Kortylewski et al. 2018a, 2019].

Almost all applications of 3DMMs in the context of face recognition would benefit from improvements of the parametric model as well as the fitting process. A more realistic texture model including textural details and modeling hair would enhance the quality of synthetic data, possibly further reducing the amount of real-world data needed to train data-driven models. A more accurate fitting process would enhance the model's performance on face frontalization and face recognition from 3DMM parameters.

7.2 Entertainment

3DMMs are an integral building block for many compelling applications in the entertainment sector. Such applications normally have to work in the wild and based on a low number of sensors, e.g., only the images captured by a single color camera are accessible. In such underconstrained scenarios, the statistical prior that is encapsulated in the face model is a powerful tool to better constrain the underlying reconstruction problems. In the following, we discuss several entertainment applications in detail. These applications are also covered in more depth in the state of the art report of Zollhoefer et al. [2018].

7.2.1 Controlling 3D Avatars for Games and VR. Realistic 3D face avatars can be reconstructed based on multi-view video [Lombardi et al. 2018], a few images [Cao et al. 2016b; Ichim et al. 2015] or given only a single image [Hu et al. 2017b; Wang et al. 2019a]. Such avatars or even artist-designed characters can be controlled in gaming scenarios based on dense trackers that employ a parametric face model. Such vision-based control was first demonstrated in an off-line setting [Chai et al. 2003; Chuang and Bregler 2002; Pighin and Lewis 2006; Wang et al. 2004; Weise et al. 2009]. Nowadays, dense facial performance capture is feasible at real-time rates based on RGB-D [Li et al. 2013; Thies et al. 2015; Weise et al. 2011b] and color [Bouaziz et al. 2013; Thies et al. 2016] cameras. Besides vision-based animation, there is extended work on audio-based control [Cudeiro et al. 2019; Karras et al. 2017; Kshirsagar and Magnenat-Thalmann 2003; Taylor et al. 2017]. Face tracking can also be used to enable face-to-face communication [Li et al. 2015a; Lombardi et al. 2018; Olszewski et al. 2016; Thies et al. 2018c] in virtual reality.

7.2.2 Virtual Try-On and Make-Up. Face reconstruction and tracking based on a parametric face model can also be employed to build virtual mirrors that enable the try-on of accessories or make-up.

To this end, first, a personalized model of the face is recovered and tracked across the video resulting in a dense set of correspondences. These enable spatio-temporal re-texturing, e.g., to virtually place tattoos [Garrido et al. 2014] and can be used to add facial make-up [Bronstein et al. 2007] and try out different suggestions [Scherbaum et al. 2011]. Virtual make-up can be applied based on a reflectance/shading decomposition [Li et al. 2014b, 2015b]. Similar techniques enable the try-on of accessories, e.g., eyeglasses [Azevedo et al. 2016; Niswar et al. 2011].

7.2.3 Face Replacement a.k.a. Face Swap. Face replacement enables the replacement of the inner face region in a target video with that from a source video. To this end, both persons are reconstructed based on the same parametric model resulting in dense inter-person correspondences. First approaches enabled face replacement between images [Bitouk et al. 2008; Blanz et al. 2004b; Jones et al. 2008; Kemelmacher-Shlizerman 2016]. Later works extended those ideas including skin and hair segmentation to deal with glasses and occlusion by hair [Pierrard 2008]. Other techniques focus on swapping faces between video sequences [Dale et al. 2011; Garrido et al. 2014]. Today the effect is mostly known under the term 'face swap' and has been popularized by a SnapChat[14] filter.

7.2.4 Face Reenactment and Visual Dubbing. Facial reenactment is the process of transferring the facial expressions from a source to a target video. First, off-line techniques have been proposed [Blanz et al. 2003; Bregler et al. 1997; Kemelmacher-Shlizerman et al. 2010; Li et al. 2014a, 2012; Theobald et al. 2009; Vlasic et al. 2005b]. The first real-time facial reenactment approach [Thies et al. 2015] was based on an RGB-D sensor, see Fig. 8. Afterward, also real-time techniques for reenacting standard video have been proposed [Thies et al. 2016]. Other approaches enable to take control of a single image [Averbuch-Elor et al. 2017; Saragih et al. 2011]. Follow-up work focused on controlling more than just the face region, e.g., the complete upper body [Thies et al. 2018b]. Nowadays, many reenactment approaches are based on deep generative models [Kim et al. 2018a; Pumarola et al. 2018].

Facial reenactment [Kim et al. 2018a; Thies et al. 2016] can also be applied to the problem of visual dubbing, i.e., the task of adapting the mouth motion of a target actor to match a new audio track. More sophisticated visual dubbing approaches [Garrido et al. 2015] directly take the new audio track into account for better audio-visual alignment. There is also some work on audio-based animation of video [Brand 1999; Suwajanakorn et al. 2017].

7.3 Medical Applications

The clinical applications of the 3DMM cover both, analysis as well as synthesis. The dominant applications lie in analysis, where diseases can be recognized by facial shape. One example of such an effect is the classification and early diagnosis of fetal alcohol spectrum disorder [Suttie et al. 2013] or epilepsy [Ahmedt Aristizabal 2019]. Similarly, Hammond et al. [2004] demonstrated both visualization and recognition of congenital craniofacial growth disorders. Both these works used 3D data. However, the capability of 3D reconstruction from 2D images was explored for the screening of acromegaly [Learned-Miller et al. 2006] and genetic disorders [Tu et al. 2018].

In the direction of synthesis, the 3D shape model was explored to perform reconstruction of missing face parts based on the model statistics [Basso and Vetter 2005; Mueller et al. 2011]. Such a reconstruction can be applied for personalized implant design. Another work explored the synthesis capabilities for analysis and generated controlled stimuli to study responses in the fusiform face area and correlates them with autism spectrum disorder [Jiang et al. 2013].

3DMM and statistical shape models, in general, are a popular standard framework in the field of medical imaging for segmentation and as models of variations in anatomical structures [Zheng et al. 2017]. A lot of those applications deal with pathologies in young or elderly people which are underrepresented even in the biggest face models [Ploumpis et al. 2019]. Those applications would profit from models built from a wider population or models than can better generalize beyond the data they are trained on.

7.4 Forensics

Applications in forensics range from identikit pictures over virtual aging to face reconstruction from dry skulls and recently also the detection of manipulated videos.

Describing faces from vague mental images is a challenging task. A tool based on a 3DMM [Blanz et al. 2006] allows exploring correlations within the face to generate indentikit pictures when providing descriptions based on vague features.

Virtual aging is a challenging task and can be helpful to later find missing children or victims of sexual abuse. The 3DMM helps to reduce the subjectivity of age progression methods. Several works in this direction are modeling age trajectories on 3DMM shape [Hutton et al. 2003; Koudelová et al. 2015; Shen et al. 2014] and at least two attempts have been made to do so for both 3DMM shape and texture [Hunter and Tiddeman 2009; Scherbaum et al. 2007]. Most methods focus on children and neglect textural details or wrinkles which are modeled in Pascal [2010]; Schneider et al. [2019].

Face reconstruction from dry skulls is an ill-posed problem. The mapping from the skull to face is not a one-to-one, but a one-to-many mapping. Models allow to control attributes for this reconstruction [Paysan et al. 2009b], explicitly estimate the posterior solution of possible faces per skull [Madsen et al. 2018] or model soft tissue thickness directly grounded by a 3DMM [Gietzen et al. 2019].

Recently 3DMMs were used successfully to generate or manipulate images and videos as discussed in Chapter 7.2.4. At the same time 3DMMs are also helpful to detect those manipulations from state of the art methods with high accuracy [Rossler et al. 2019].

7.5 Cognitive Science, Neuroscience, and Psychology

The ability to generate faces that can be controlled via parameters is very popular when studying how the human and non-human primate brain process faces. Studies with generated stimuli from a 3DMM can be found in Cognitive Science, Neuroscience, Psychology, and Social Science.

One of the earliest works using 3DMMs presented high-level aftereffects that indicate a model related to a statistical face model

[14] https://www.snapchat.com/

in the human brain. Those aftereffects were demonstrated using caricaturized faces and antifaces [Leopold et al. 2001]. Later it was shown that those results can not only be observed as aftereffects but also as responses of single neurons across caricaturization in macaque monkey to principal axes of a 3DMM [Leopold et al. 2006]. Later those aftereffects were shown to incorporate 3D information [Jiang et al. 2009a]. The effects based on caricatures for recognition were recently also investigated with 3DMMs in artificial neural networks trained on face recognition [Hill et al. 2018].

A topic that was heavily researched over the past decades and is still under investigation is how much the 3D shape contributes to face perception and if the face representation in our brain is built as a 3D model. Early studies based on functional MRI and behavioral techniques evaluated a shape-based model of human face discrimination [Jiang et al. 2006]. Later studies investigated the importance of 3D shape and surface reflectance [Jiang et al. 2009b] and event-related potentials to 3D shape are faster than to surface reflection [Caharel et al. 2009]. Other work explored how well humans can estimate a profile picture from a frontal view [Schumacher and Blanz 2012].

Recently it was shown that a face-processing system based on stepwise inverse rendering correlates better to neural measurements in macaque monkey than state of the art artificial neural networks [Yildirim et al. 2020].

Face image manipulation is another key application of the 3DMM to generate stimuli [Walker and Vetter 2009] to e.g., investigate social judgments based on facial appearance. Again the ability to control exactly what is manipulated is key for those research results sometimes measuring subtle effects [Walker et al. 2011]. Recently a dataset of controlled manipulated images was released to perform such experiments [Walker et al. 2018].

One of the major limitations compared to 2D based methods is that 3DMMs do not include hair. In a lot of studies faces and hair are not separated since faces without hair appear less face-like. For those models, it plays a substantial role to have controls over the parameters and that parameters can be interpreted which secures the future of 3DMMs in those fields.

8 PERSPECTIVE

In this last section, we want to look beyond the state of the art. We explicitly highlight the unsolved challenges in the field. In addition to focusing on face models, we look further and share our thoughts about the scalability of 3DMMs beyond faces. We also share our thoughts of the applicability of models including data, model and algorithm sharing also with its potential of misuse. We close with an outlook on how a 3DMM could look like in 10 or 20 years.

8.1 Global Challenges

In this section, we summarize the major open challenges that are shared across the different parts of 3DMMs. Local challenges that are specific to capturing, modeling, image formation or analysis-by-synthesis are mentioned in the respective sections.

One of the leading challenges is the balance between a low-dimensional parametric model and the degree of detail we are capable of modeling. Parametric models for eyes, teeth, hairs, skin details, soft tissue or even anatomical grounded muscles are not available. Additional complexity also renders analysis-by-synthesis even more challenging. Building faces with all those details is currently possible for a single face with a lot of manual labor, but automatic methods to extract those details or build models on top of them are in their beginnings. Current state of the art methods from capture, to modeling over image formation to analysis-by-synthesis use a lot of oversimplifying assumptions. Besides including more facial details there are also models that exploit the knowledge that a face is part of the body. Whilst faces and bodies are mostly analyzed separately, there exist first models that include faces and bodies jointly [Joo et al. 2018; Pavlakos et al. 2019]. Pavlakos et al. [2019] presented first results indicating that fitting the whole body is also beneficial for the quality measured in the face region only.

Another major challenge is the comparability of all the components of a 3DMM. Already the modeling itself can only be evaluated on specific tasks and different models have a different focus and might perform better on a specific task. For analysis-by-synthesis, comparing the performance of a model and also of the model adaptation algorithm is an unsolved problem. Current state of the art research frequently focuses on task-specific qualitative results and those results can barely be compared across models and algorithms. The current trend in the community to share source-code and models helps to compare and reproduce results, however, there is a lack of useful benchmarks. A first step in this direction is a new dataset providing natural images in combination with a 3D scan of the same individual [Sanyal et al. 2019]. However this is focused on shape reconstruction only, there is no single benchmark for 3D reconstruction from 2D images including illumination and albedo estimation.

The last challenges are of an ethical nature. Concerns around image analysis and synthesis, especially for faces is currently discussed within the scientific community as well as in the media and the broad public. The current algorithmic development in computer vision and graphics allows to recognize faces and to generate or manipulate images and video. In addition most methods around 3DMMs elicit some dataset bias. Saito et al. [2017] approached this using the Chicago Face Databse[Ma et al. 2015] to build a face model with balanced ethnicities. Those challenges are not a purely scientific one, but also a political one. We start to see regulations of those technologies and there will be likely more regulations across the world in the near future. As a community, we can choose on what projects we focus to work on and there are plenty of meaningful and valuable applications of 3DMMs, face analysis, and face synthesis as we presented in the Section 7 which could be explored less with restrictive regulations.

8.2 Scalability

Research on parametric models of human faces has seen a lot of progress in recent years. This raises the question of how scalable the found solutions are to other types of real-world entities beyond humans. On the one hand, human faces are highly challenging as we are attuned to noticing even slightest inaccuracies in their modeling. At the same time, they are also more amenable to statistical modeling as their structure is relatively regular and correspondence

across faces is quite well-defined. Other types of real-world entities, or even humans in clothing or the human head with full hair, are exhibiting much stronger appearance, structure, and shape variation that may require additional methodical innovations to empower proper modeling. The vision and graphics communities have begun to build and learn statistical models of other types of shape categories. Researchers also increasingly attempt to learn such models in an unsupervised or weakly supervised way for better real-world scalability. These approaches partially build on many concepts learned from the models described in this article but introduce additional representation innovations, like learned implicit representations [Cole et al. 2017; Eslami et al. 2018; Sitzmann et al. 2019], to handle their specific structural properties. Future research will certainly see more work in this direction that answers the question of what is the right shape, appearance and deformation representations for a wider range of real-world object classes.

8.3 Application

An additional challenge for our research community will be to agree on efficient ways to share and combine research efforts performed by different research groups. We should agree on common data formats and dissemination channels for available scan databases, which would simplify building integrated models, and enable us to better test and compare them. In that context, ever more pressing questions of privacy and security will also need to be addressed. On the one hand, it is needless to say that we have to adhere to highest standards of privacy protection in data sets we share, so not to reveal personal data or identities beyond what is needed and permitted by law or by the captured individuals. For handling this, community-wide procedures for providing consent on the use of data that are compatible with legal regulations could be agreed on and shared.

However, beyond this, increasingly powerful methods to build and reconstruct such face models from image and video will in the future enable us to build highly believable 3D human avatars from casually captured imagery. These avatars will enable us to create virtual renditions of real people at unseen accuracy to populate computer graphically generated virtual spaces at high visual fidelity. However, algorithmic tools should be investigated as well that prevent the reconstruction or use of such avatars in undesired or questionable applications that a reconstructed person did not provide consent on. Advanced reconstruction algorithms on the basis of parametric models may also make it possible to extract semantic information of people from imagery that they may not want to reveal (e.g., about emotional state, health, and physical condition, etc.). Therefore, algorithmic strategies to balance personal privacy and reconstruction ability shall be investigated and provided by our research community.

Also, the continuously improving performance of algorithms to reconstruct detailed human models from single images or videos enables advanced new ways to synthesize new face imagery or even modify existing face images and videos at very high visual fidelity. As an example, some recent combinations of model-based reconstruction algorithms and adversarially trained neural networks have shown impressive results in that respect. Such advanced synthesis algorithms will simplify many applications and open up entirely new applications, for instance in content creation for animation and visual effects, in content creation for virtual and augmented reality, in telepresence, visual dubbing or advanced video editing. However, they might also be used to create or modify media content with malicious intent. Therefore, as a community focusing on basic research, we will continue our efforts to objectively inform the general public about the great possibilities opened up by advanced parametric models of face, body and other real-world entities to build the next generation of intelligent, interactive and creative computing systems. At the same time, we will use our essential basic expertise about the underlying algorithmic principles to develop new ways to detect unwanted media synthesis and modification and to prevent such unwanted modifications algorithmically.

8.4 Outlook

The big question we ask is how will a generative face model look like in 10 or 20 years? What will be the representation and will it be a complete model of the human face with all its variation and details? Currently, we experience a divergence of 3DMMs. Different research teams put a different focus and model some parts in more detail but lack other details or statistical variation. Recent modeling advances are focusing on building task-specific representations rather than a more general face model to be applicable for multiple tasks. For some applications, the model itself is the limiting factor, whilst other applications profit from a simple model based on PCA. The requirements in terms of quality, realism, generalization, and performance are very different e.g., content creation vs. computer vision. The gap between state of the art computer-generated renderings for a single face including expressions versus generative and parametric face models based on statistics is dramatic.

Current advances in the field of machine learning will contribute to build more general and at the same time more realistic models. The core of the face model was always interpreted as a learning problem, recent advances lifted the analysis-by-synthesis task from a per image optimization task to a learning challenge. However, this loop is not yet closed - why not learn or improve the model itself? There are already first works in the direction of model learning (compare Section 6.3), but they are limited by very similar modeling assumptions as traditional 3DMMs. First steps to overcome those were recently performed in the direction of neural rendering [Eslami et al. 2018; Thies et al. 2019], 3D representation learning [Sitzmann et al. 2019] and unsupervised shape model learning [Szabó et al. 2019]. Other modeling approaches like generative adversarial networks [Goodfellow et al. 2014; Karras et al. 2019b,a] are currently operating in 2D image space. Such parametric models can be used to embed faces of real people in a latent space Abdal et al. [2019a,b], but the resulting embedding is hard for humans to interpret.

20 years ago 3DMMs were part of a revolution in computer graphics and computer vision to go away from 2D image processing to 3D modeling. The computer vision community is currently focusing again on mainly 2D based approaches and we have to propose the missing key to again move the community to 3D. Additionally one of the leading benefits of 3DMM is the natural disentanglement of

shape, color, illumination and camera parameters. Such a disentanglement is very hard to be derived purely from data [Locatello et al. 2019] and for faces, 3DMMs build it manually based on the image formation process. According to "Pattern Theory" [Grenander 1996; Mumford and Desolneux 2010] it is a prerequisite for any high-performance image analysis system to find and separate conditional independent parameters that describe the image to analyze. The discovery and separation of such parameters purely from 2D data is still an unsolved challenge. 3DMMs directly implement models using the parameter also used by physics and geometry to model light and three-dimensional objects.

One direction which might be particularly interesting is to break out of the common modeling assumptions and oversimplification but at the same time automate the tedious manual work behind the photo-realistic generation of faces. We expect some kind of living 3DMM to evolve from the community. Automation will be the leading modeling idea. A living 3DMM should be able to learn from 3D data as well as 2D data, both still and in motion. We imagine the model to be learned from a minimal seed like a mean face, a sphere or just a rough prototype based on the first few data points. The optimal living model would not be task-specific but should be able to generalize to various tasks. The face model must, therefore, be hierarchical in some form to represent multiple degrees of detail but share statistics across those levels. Such an optimal face model would be general enough to be applicable for real-time computer vision tasks, analysis-by-synthesis from currently challenging images as well as photorealistic rendering with a high level of facial details. Last but not least some tasks rely on an interpretable parametrization and not just a black box learning machine. Basic knowledge of geometry and physics would not only ease the learning but also at least disentangle pose and illumination variation from the facial shape and appearance. Building such a general face model might remain a challenge for the next 10 or 20 years but would align with the original idea behind 3DMMs.

ACKNOWLEDGMENTS

This survey paper was initiated at the Dagstuhl Seminar 19102 on 3D Morphable Models [Egger et al. 2019] and contains ideas resulting from discussions at this seminar. This survey paper was partially funded by Early PostDoc Mobility Grant, Swiss National Science Foundation P2BSP2_178643, ERC Consolidator Grant 4DRepLy and the Max Planck Center for Visual Computing and Communications (MPC-VCC). We thank Barış Geçer for his help on the teaser figure, and Haiwen Feng for providing the FLAME texture space. We thank the anonymous reviewers whose comments have greatly improved this manuscript.

REFERENCES

2005. CASIA-3D FaceV1. (2005). http://biometrics.idealtest.org/

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019a. Image2StyleGAN++: How to Edit the Embedded Images?

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

Jascha Achenbach, Robert Brylka, Thomas Gietzen, Katja zum Hebel, Elmar Schömer, Ralf Schulze, Mario Botsch, and Ulrich Schwanecke. 2018. A multilinear model for bidirectional craniofacial reconstruction. In *Proc. Eurographics Workshops*. Eurographics Association, 67–76.

Jens Ackermann, Michael Goesele, et al. 2015. A survey of photometric stereo techniques. *Foundations and Trendsin Computer Graphics and Vision* 9, 3-4 (2015), 149–254.

David Esteban Ahmedt Aristizabal. 2019. *Multi-modal analysis for the automatic evaluation of epilepsy*. Ph.D. Dissertation. Queensland University of Technology.

Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, and Stefano Berretti. 2014. A 3D Dynamic Database for Unconstrained Face Recognition. In *Proc. International Conference and Exhibition on 3D Body Scanning Technologies*.

Oswald Aldrian and WA Smith. 2010. A linear approach of 3d face shape and texture recovery using a 3d morphable model. In *Proc. British Machine Vision Conference (BMVC)*.

Oswald Aldrian and William AP Smith. 2011a. Inverse rendering in suv space with a linear texture model. In *Proc. International Conference on Computer Vision (ICCV) Workshops*. IEEE, 822–829.

Oswald Aldrian and William AP Smith. 2012. Inverse rendering of faces on a cloudy day. In *Proc. European Conference on Computer Vision (ECCV)*. 201–214.

Oswald Aldrian and William AP Smith. 2013. Inverse rendering of faces with a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 5 (2013), 1080–1093.

Oswald Aldrian and William A. P. Smith. 2011b. Inverse Rendering with a Morphable Model: A Multilinear Approach. In *Proc. British Machine Vision Conference (BMVC)*.

Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. 2003. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM Transactions on Graphics*, Vol. 22. ACM, 587–594.

Sarah Alotaibi and William AP Smith. 2017. A Biophysical 3D Morphable Model of Face Appearance. In *Proc. International Conference on Computer Vision (ICCV) Workshops*. IEEE, 824–832.

Brian Amberg, Andrew Blake, Andrew Fitzgibbon, Sami Romdhani, and Thomas Vetter. 2007. Reconstructing high quality face-surfaces using model based stereo. In *Proc. International Conference on Computer Vision (ICCV)*. IEEE, 1–8.

Brian Amberg, Reinhard Knothe, and Thomas Vetter. 2008. Expression invariant 3D face recognition with a morphable model. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE, 1–6.

Brian Amberg and Thomas Vetter. 2011. Optimal landmark detection using shape models and branch and bound. In *Proc. International Conference on Computer Vision (ICCV)*. IEEE, 455–462.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. 408–416.

Joseph J Atick, Paul A Griffin, and A Norman Redlich. 1996. Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images. *Neural computation* 8, 6 (1996), 1321–1340.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Transactions on Graphics* 36, 6 (2017), 196:1–196:13.

Pedro Azevedo, Thiago Oliveira-Santos, and Edilson De Aguiar. 2016. An Augmented Reality Virtual Glasses Try-On System. In *Symposium on Virtual Reality*. 1–9. https://doi.org/10.1109/SVR.2016.12

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. 2018. Modeling Facial Geometry Using Compositional VAEs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. 2011. The Florence 2D/3D Hybrid Face Dataset. In *Joint ACM Workshop on Human Gesture and Behavior Understanding (J-HGBU '11)*. ACM, New York, NY, USA, 79–80. https://doi.org/10.1145/2072572.2072597

Anil Bas, Patrik Huber, William AP Smith, Muhammad Awais, and Josef Kittler. 2017a. 3D Morphable Models as Spatial Transformer Networks. In *Proc. International Conference on Computer Vision (ICCV) Workshops*. IEEE, 895–903.

Anil Bas and William A. P. Smith. 2019. What does 2D geometric information really tell us about 3D face shape? *International Journal of Computer Vision* 127 (2019).

Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. 2017b. Fitting a 3D Morphable Model to Edges: A Comparison Between Hard and Soft Correspondences. In *Asian Conference on Computer Vision Workshops*, Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma (Eds.). Springer International Publishing, Cham, 377–391.

Curzio Basso and Alessandro Verri. 2007. Fitting 3D morphable models using implicit representations. In *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. (VISIGRAPP)*. 45–52.

Curzio Basso and Thomas Vetter. 2005. Statistically motivated 3D faces reconstruction. In *Proc. International Conference on Reconstruction of Soft Facial Parts*, Vol. 31. Citeseer.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, Vol. 29. 40.

Thabo Beeler, Bernd Bickel, Gioacchino Noris, Paul Beardsley, Steve Marschner, Robert W Sumner, and Markus Gross. 2012. Coupled 3D reconstruction of sparse facial hair and skin. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 31, 4 (2012), 117.

Thabo Beeler and Derek Bradley. 2014. Rigid stabilization of facial expressions. *ACM Transactions on Graphics* 33, 4 (2014), 44.

Thabo Beeler, 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 Transactions on Graphics (Proceedings of SIGGRAPH)*. ACM, New York, NY, USA, Article 75, 10 pages. https://doi.org/10.1145/1964921.1964970

Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus Gross. 2014. High-quality capture of eyes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 33, 6 (2014), 223.

Amit Bermano, Thabo Beeler, Yeara Kozlov, Derek Bradley, Bernd Bickel, and Markus Gross. 2015. Detailed spatio-temporal reconstruction of eyelids. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 34, 4 (2015), 44.

Stefano Berretti, Boulbaba Ben Amor, Mohamed Daoudi, and Alberto del Bimbo. 2011. 3D facial expression recognition using SIFT descriptors of automatically detected keypoints. *The Visual Computer* 27, 11 (2011), 1021–1036.

Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. 2008. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Transactions on Graphics* 27, 3 (2008), 39:1–39:8.

Volker Blanz, Irene Albrecht, Jörg Haber, and H-P Seidel. 2006. Creating face models from vague mental images. In *Compututer Graphics Forum*, Vol. 25. Wiley Online Library, 645–654.

Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. 2003. Reanimating faces in images and video. In *Compututer Graphics Forum*, Vol. 22. Wiley Online Library, 641–650.

Volker Blanz, Patrick Grother, P Jonathon Phillips, and Thomas Vetter. 2005. Face recognition based on frontal views generated from non-frontal images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. IEEE, 454–461.

Volker Blanz, Albert Mehl, Thomas Vetter, and Hans-Peter Seidel. 2004a. A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data. In *Proc. 3D Data Processing Visualization and Transmission*. 293–300.

Volker Blanz, Sami Romdhani, and Thomas Vetter. 2002. Face identification across different poses and illuminations with a 3d morphable model. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE, 202–207.

Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. 2004b. Exchanging Faces in Images. *Compututer Graphics Forum* 23, 3 (2004), 669–676.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. 187–194.

Volker Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 9 (2003), 1063–1074.

Timo Bolkart and Stefanie Wuhrer. 2015a. 3D Faces in Motion: Fully Automatic Registration and Statistical Analysis. *Computer Vision and Image Understanding* 131 (2015), 100–115.

Timo Bolkart and Stefanie Wuhrer. 2015b. A Groupwise Multilinear Correspondence Optimization for 3D Faces. In *Proc. International Conference on Computer Vision (ICCV)*. 3604–3612.

Timo Bolkart and Stefanie Wuhrer. 2016. A Robust Multilinear Model Learning Framework for 3D Faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4911–4919.

James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 2017. 3D face morphable models "In-The-Wild". In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5464–5473.

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

James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou. 2018b. 3D Reconstruction of In-the-Wild Faces in Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 11 (2018), 2638–2652.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniahy, and David Dunaway. 2016. A 3D Morphable Model Learnt from 10,000 Faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5543–5552.

James Booth and Stefanos Zafeiriou. 2014. Optimal uv spaces for facial morphable model construction. In *Proc. IEEE International Conference on Image Processing*.

Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013. Online Modeling for Realtime Facial Animation. *ACM Transactions on Graphics* 32, 4 (2013), 40:1–40:10.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High resolution passive facial performance capture. In *ACM Transactions on Graphics*, Vol. 29. ACM, 41.

Matthew Brand. 1999. Voice Puppetry. In *ACM Transactions on Graphics*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 21–28.

Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: Driving Visual Speech with Audio. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 353–360.

Pia Breuer, Kwang-In Kim, Wolf Kienzle, Bernhard Scholkopf, and Volker Blanz. 2008. Automatic 3D face reconstruction from single images or video. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE, 1–8.

Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel. 2007. Calculus of non-rigid surfaces for geometry and texture manipulation. *Transactions on Visualization and Computer Graphics* 13, 5 (2007), 902–913.

Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. 2014a. Multilinear wavelets: A statistical shape space for human faces. In *Proc. European Conference on Computer Vision (ECCV)*. 297–312.

Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. 2014b. Review of statistical shape spaces for 3D data with comparative analysis for human faces. *Computer Vision and Image Understanding* 128, 0 (2014), 1 – 17.

Alan Brunton, Chang Shu, Jochen Lang, and Eric Dubois. 2011. Wavelet Model-based Stereo for Fast, Robust Face Reconstruction. In *Proc. Canadian Conference on Computer and Robot Vision*.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How Far Are We From Solving the 2D & 3D Face Alignment Problem? (And a Dataset of 230,000 3D Facial Landmarks). In *Proc. International Conference on Computer Vision (ICCV)*.

Stéphanie Caharel, Fang Jiang, Volker Blanz, and Bruno Rossion. 2009. Recognizing an individual face: 3D shape contributes earlier than 2D surface reflectance information. *Neuroimage* 47, 4 (2009), 1809–1818.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time High-fidelity Facial Performance Capture. *ACM Transactions on Graphics* 34, 4, Article 46 (July 2015), 9 pages. https://doi.org/10.1145/2766943

Chen Cao, Qiming Hou, and Kun Zhou. 2014a. Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation. *ACM Transactions on Graphics* 33, 4, Article 43 (July 2014), 10 pages. https://doi.org/10.1145/2601097.2601204

Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D Shape Regression for Real-time Facial Animation. *ACM Transactions on Graphics* 32, 4, Article 41 (July 2013), 10 pages. https://doi.org/10.1145/2461912.2462012

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014b. FaceWarehouse: A 3D facial expression database for visual computing. *Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425.

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016a. Real-time Facial Animation with Image-based Dynamic Avatars. *ACM Transactions on Graphics* 35, 4, Article 126 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925873

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016b. Real-time Facial Animation with Image-based Dynamic Avatars. *ACM Transactions on Graphics* 35, 4 (2016), 126:1–126:12.

Thomas J Cashman and Andrew W Fitzgibbon. 2012. What shape are dolphins? building 3d morphable models from 2d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2012), 232–244.

Jin-xiang Chai, Jing Xiao, and Jessica Hodgins. 2003. Vision-based Control of 3D Facial Animation. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 193–206.

Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. 2018. ExpNet: Landmark-free, deep, 3D facial expressions. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE, 122–129.

Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. 2017. Faceposenet: Making a case for landmark-free face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*. 1599–1608.

Anpei Chen, Zhang Chen, Guli Zhang, Ziheng Zhang, Kenny Mitchell, and Jingyi Yu. 2019. Photo-Realistic Facial Details Synthesis from Single Image. *Proc. International Conference on Computer Vision (ICCV)* (2019).

Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 2019. MeshGAN: Non-linear 3D Morphable Models of Faces. *arXiv preprint arXiv:1903.10384* (2019).

Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 2018. 4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Erika Chuang and chris. Bregler. 2002. *Performance-driven Facial Animation using Blend Shape Interpolation.* Technical Report CS-TR-2002-02. Stanford University.

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. 2017. Synthesizing normalized faces from facial identity features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3703–3712.

3DMM Community. 2019. Curated List of 3D Morphable Model Software and Data. https://github.com/3d-morphable-models/curated-list-of-awesome-3D-Morphable-Model-software-and-data.

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. 1998. Active Appearance Models. In *Proc. European Conference on Computer Vision (ECCV)*.

Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. 1995. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding* 61, 1 (1995), 38 – 59. https://doi.org/10.1006/cviu.1995.1004

Darren Cosker, Eva Krumhuber, and Adrian Hilton. 2011. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *Proc. International Conference on Computer Vision (ICCV)*. 2296–2303.

Ian Craw and Peter Cameron. 1991. Parameterising images for recognition and reconstruction. In *Proc. British Machine Vision Conference (BMVC)*. Springer, 367–370.

Clement Creusot, Nick Pears, and Jim Austin. 2013. A Machine-Learning Approach to Keypoint Detection and Landmarking on 3D Meshes. *International Journal of Computer Vision* 102, 1-3 (2013), 146–179.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hang Dai, Nick Pears, and William Smith. 2018. A Data-augmented 3D Morphable Model of the Ear. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE, 404–408.

Hang Dai, Nick Pears, William A. P. Smith, and Christian Duncan. 2017. A 3D Morphable Model of Craniofacial Shape and Texture Variation. In *Proc. International Conference on Computer Vision (ICCV)*.

Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video Face Replacement. *ACM Transactions on Graphics* 30, 6 (2011), 130:1–130:10.

Michael De Smet, Rik Fransens, and Luc Van Gool. 2006. A generalized EM approach for 3D model based face recognition under occlusions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. IEEE, 1423–1430.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proc. Conference on Computer graphics and Interactive Techniques*. ACM Press/Addison-Wesley Publishing Co., 145–156.

Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. 2018. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7093–7102.

Yu Deng, 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Arnaud Dessein, William A.P. Smith, Richard C. Wilson, and Edwin R. Hancock. 2015. Example-Based Modeling of Facial Texture from Deficient Data. In *Proc. International Conference on Computer Vision (ICCV)*. 3898–3906.

Roman Dovgard and Ronen Basri. 2004. Statistical symmetric shape from shading for 3D structure recovery of faces. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 99–113.

Ian Dryden and Kanti Mardia. 2002. *Statistical Shape Analysis*. Wiley.

Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. 2019. Deep Appearance Models: A Deep Boltzmann Machine Approach for Face Modeling. *International Journal of Computer Vision* 127, 5 (2019), 437–455.

Jose I Echevarria, Derek Bradley, Diego Gutierrez, and Thabo Beeler. 2014. Capturing and stylizing hair for 3D fabrication. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 33, 4 (2014), 125.

Bernhard Egger. 2018. *Semantic Morphable Models*. Ph.D. Dissertation. University of Basel.

Bernhard Egger, Dinu Kaufmann, Sandro Schönborn, Volker Roth, and Thomas Vetter. 2016a. Copula eigenfaces. In *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. (GRAPP)*. 50–58.

Bernhard Egger, Dinu Kaufmann, Sandro Schönborn, Volker Roth, and Thomas Vetter. 2016b. Copula Eigenfaces with Attributes: Semiparametric Principal Component Analysis for a Combined Color, Shape and Attribute Model. In *Communications in Computer and Information Science*. Springer, 95–112.

Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. 2018. Occlusion-Aware 3D Morphable Models and an Illumination Prior for Face Image Analysis. *International Journal of Computer Vision* 126, 12 (01 Dec 2018), 1269–1287. https://doi.org/10.1007/s11263-018-1064-8

Bernhard Egger, William Smith, Christian Theobalt, and Thomas Vetter. 2019. 3D Morphable Models (Dagstuhl Seminar 19102). *Dagstuhl Reports* 9, 3 (2019), 16–38. https://doi.org/10.4230/DagRep.9.3.16

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. 2018. Neural scene representation and rendering. 360, 6394 (2018), 1204–1210.

Tony Ezzat, Gadi Geiger, and Tomaso Poggio. 2002. Trainable Videorealistic Speech Animation. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. ACM, New York, NY, USA, 388–398. https://doi.org/10.1145/566570.566594

Gabrielle Fanelli, Jürgen Gall, Harald Romsdorfer, Thibaut Weise, and Luc van Gool. 2010. A 3D Audio-Visual Corpus of Affective Communication. *IEEE MultiMedia* 12, 6 (2010), 591 – 598.

Tianhong Fang, Xi Zhao, Omar Ocegueda, Shishir K. Shah, and Ioannis A. Kakadiaris. 2012. 3D/4D facial expression analysis: An advanced annotated face model approach. *Image and Vision Computing* 30, 10 (2012), 738–749.

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *Proc. European Conference on Computer Vision (ECCV)*.

Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. 2019. A Generative 3D Facial Model by Adversarial Training. In *Proc. International Conference on Computer Vision (ICCV)*.

Victoria Fernández Abrevaya, Stefanie Wuhrer, and Edmond Boyer. 2018. Multilinear Autoencoder for 3D Face Model Learning. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Victoria Fernández Abrevaya, Stefanie Wuhrer, and Edmond Boyer. 2018. Spatiotemporal Modeling for Efficient Registration of Dynamic 3D Faces. In *Proc. IEEE International Conference on 3D Vision (3DV)*. 371–380.

Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. 2015. Dictionary Learning based 3D Morphable Model Construction for Face Recognition with Varying Expression and Pose. In *Proc. IEEE International Conference on 3D Vision (3DV)*. 509–517.

Rik Fransens, Christoph Strecha, and Luc Van Gool. 2005. Parametric stereo for multi-pose face recognition and 3D-face modeling. In *Proc. International Conference on Automatic Face and Gesture Recognition*. Springer, 109–124.

Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. 2016. Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics* 35, 4 (2016), 128.

Yasutaka Furukawa and Jean Ponce. 2009. Dense 3D motion capture for human faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1674–1681.

Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormaehlen, Patrick Pérez, and Christian Theobalt. 2014. Automatic Face Reenactment. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 4217–4224.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2015. VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Compututer Graphics Forum* 34, 2 (2015), 193–204.

Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics* 32, 6 (2013), 158–1.

Pablo Garrido, Michael Zollhoefer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016a. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics* 35, 3 (2016), 28.

Pablo Garrido, Michael Zollhoefer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt. 2016b. Corrective 3D reconstruction of lips from monocular video. *ACM Transactions on Graphics* 35, 6 (2016), 219–1.

Baris Gecer, Alexander Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. 2019a. Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. *arXiv preprint arXiv:1909.02215* (2019).

Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019b. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jason Geng. 2011. Structured-light 3D surface imaging: a tutorial. *Advances in Optics and Photonics* 3, 2 (Jun 2011), 128–160.

Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. 2018. Unsupervised training for 3d morphable model regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8377–8386.

Athinodoros S Georghiades. 2003. Incorporating the Torrance and Sparrow Model of Reflectance in Uncalibrated Photometric Stereo. In *Proc. International Conference on Computer Vision (ICCV)*. 816.

Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. 2018. Morphable Face Models - An Open Framework. In *Proc. International Conference on Automatic Face and Gesture Recognition*. 75–82.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, Vol. 30. 129.

Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. 2008. Practical modeling and acquisition of layered facial reflectance. In *ACM Transactions on Graphics*, Vol. 27. ACM, 139.

Thomas Gietzen, 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. *PloS one* 14, 1 (2019), e0210257.

Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser. 2006. A statistical model for synthesis of detailed facial geometry. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 25, 3 (2006), 1025–1034.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. Advances in neural information processing systems (NeurIPS)*. 2672–2680.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical Dynamic Facial Appearance Modeling and Acquisition. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 37, 6 (2018), 232:1–232:13.

Paulo F. U. Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. 2015. Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction. In *Proc. International Conference on Computer Vision (ICCV)*.

Ulf Grenander. 1996. *Elements of pattern theory.* JHU Press.

Jianya Guo, Xi Mei, and Kun Tang. 2013. Automatic landmark annotation and dense correspondence registration for 3D human facial images. *BMC Bioinformatics* 14, 1 (2013).

Peter L. Hallinan, Gaile G. Gordon, Alan L. Yuille, Peter Giblin, and David Mumford. 1999. *Two- and Three-Dimensional Patterns of the Face.* A K Peters/CRC Press.

Peter Hammond, Tim J Hutton, Judith E Allanson, Linda E Campbell, Raoul CM Hennekam, Sean Holden, Michael A Patton, Adam Shaw, I Karen Temple, Matthew Trotter, et al. 2004. 3D analysis of facial morphology. *American Journal of Medical Genetics Part A* 126, 4 (2004), 339–348.

Fang Han and Han Liu. 2012. Semiparametric principal component analysis. In *Proc. Advances in neural information processing systems (NeurIPS)*. 171–179.

Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective face frontalization in unconstrained images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4295–4304.

Behrend Heeren, Chao Zhang, Martin Rumpf, and William Smith. 2018. Principal Geodesic Analysis in the Space of Discrete Shells. In *Compututer Graphics Forum*, Vol. 37. 173–184.

Carlos Hernández, George Vogiatzis, Gabriel J Brostow, Bjorn Stenger, and Roberto Cipolla. 2007. Non-rigid photometric stereo with colored lights. In *Proc. International Conference on Computer Vision (ICCV)*. IEEE, 1–8.

Thomas Heseltine, Nick Pears, and Jim Austin. 2008. Three-dimensional face recognition using combinations of surface feature map subspace components. *Image and Vision Computing* 26, 3 (2008), 382–396.

Alexander Hewer, Stefanie Wuhrer, Ingmar Steiner, and Korin Richmond. 2018. A multilinear tongue model derived from speech related MRI data of the human vocal tract. *Computer Speech and Language* 51 (2018), 68–92.

Matthew Q Hill, Connor J Parde, Carlos D Castillo, Y Ivette Colon, Rajeev Ranjan, Jun-Cheng Chen, Volker Blanz, and Alice J O'Toole. 2018. Deep Convolutional Neural Networks in the Face of Caricature: Identity and Image Revealed. (2018).

Yoshitaka Ushiku Hiroharu Kato and Tatsuya Harada. 2018. Neural 3D Mesh Renderer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pei-Lu Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. 2015. Unconstrained Realtime Facial Performance Capture. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1675–1683.

Guosheng Hu, Pouria Mortazavian, Josef Kittler, and William Christmas. 2013. A facial symmetry prior for improved illumination fitting of 3D morphable model. In *Proc. International Conference on Biometrics (ICB)*. IEEE, 1–6.

Guosheng Hu, Fei Yan, Josef Kittler, William Christmas, Chi Ho Chan, Zhenhua Feng, and Patrik Huber. 2017c. Efficient 3D morphable face model fitting. *Pattern Recognition* 67 (2017), 366–379.

Liwen Hu, Derek Bradley, Hao Li, and Thabo Beeler. 2017a. Simulation-ready hair capture. In *Compututer Graphics Forum*, Vol. 36. 281–294.

Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2014a. Robust hair capture using simulated examples. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 33, 4 (2014), 126.

Liwen Hu, Chongyang Ma, Linjie Luo, Li-Yi Wei, and Hao Li. 2014b. Capturing braided hairstyles. *ACM Transactions on Graphics* 33, 6 (2014), 225.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017b. Avatar Digitization from a Single Image for Real-time Rendering. *ACM Transactions on Graphics* 36, 6, Article 195 (Nov. 2017), 14 pages. https://doi.org/10.1145/3130800.31310887

Patrik Huber. 2017. *Real-time 3D morphable shape model fitting to monocular in-the-wild videos.* Ph.D. Dissertation. University of Surrey.

Patrik Huber, Guosheng Hu, Jose Rafael Tena, Pouria Mortazavian, Willem P. Koppen, William J. Christmas, Matthias Rätsch, and Josef Kittler. 2016. A Multiresolution 3D Morphable Face Model and Fitting Framework. In *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. (VISIGRAPP)*.

David William Hunter and Bernard Paul Tiddeman. 2009. Visual ageing of human faces in three dimensions using morphable models and projection to latent structures. In *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. (VISAPP)*.

Tim J. Hutton, Bernard F. Buxton, and Peter Hammond. 2001. Dense Surface Point Distribution Models of the Human Face. In *Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA)*. 153–.

Tim J. Hutton, Bernard F. Buxton, Peter Hammond, and Henry WW Potts. 2003. Estimating average growth trajectories in shape-space using kernel smoothing. *IEEE transactions on medical imaging* 22, 6 (2003), 747–753.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Transactions on Graphics* 34, 4, Article 45 (July 2015), 14 pages. https://doi.org/10.1145/2766974

Alexandru-Eugen Ichim, Petr Kadlecek, Ladislav Kavan, and Mark Pauly. 2017. Phace: Physics-based Face Modeling and Animation. 36, 4 (2017), 153:1–14.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. In *Proc. Advances in neural information processing systems (NeurIPS)*. 2017–2025.

Fang Jiang, Volker Blanz, and Alice J O'Toole. 2009a. Three-dimensional information in face representations revealed by identity aftereffects. *Psychological Science* 20, 3 (2009), 318–325.

Fang Jiang, Laurence Dricot, Volker Blanz, Rainer Goebel, and Bruno Rossion. 2009b. Neural correlates of shape and surface reflectance information in individual faces. *Neuroscience* 163, 4 (2009), 1078–1091.

Xiong Jiang, Angela Bollich, Patrick Cox, Eric Hyder, Joette James, Saqib Ali Gowani, Nouchine Hadjikhani, Volker Blanz, Dara S Manoach, Jason JS Barton, et al. 2013. A quantitative link between face discrimination deficits and neuronal selectivity for faces in autism. *NeuroImage: Clinical* 2 (2013), 320–331.

Xiong Jiang, Ezra Rosen, Thomas Zeffiro, John VanMeter, Volker Blanz, and Maximilian Riesenhuber. 2006. Evaluation of a shape-based model of human face discrimination using FMRI and behavioral techniques. *Neuron* 50, 1 (2006), 159–172.

Andrew Jones, Jen-Yuan Chiang, Abhijeet Ghosh, Magnus Lang, Matthias Hullin, Jay Busch, and Paul Debevec. 2008. *Real-time Geometry and Reflectance Capture for Digital Face Replacement.* Technical Report 4s. University of Southern California.

Michael J Jones and Tomaso Poggio. 1998. Multidimensional morphable models. In *IJCV*. IEEE, 683–688.

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ioannis A. Kakadiaris, Georgios Passalis, Theoharis Theoharis, George Toderici, I. Konstantinidis, and N. Murtuza. 2005. Multimodal face recognition: combination of geometry with physiological information. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. 1022–1029.

Ioannis A. Kakadiaris, Georgios Passalis, George Toderici, Mohammed N Murtuza, Yunliang Lu, Nikos Karamelpatzis, and Theoharis Theoharis. 2007. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 4 (2007), 640–649.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven Facial Animation by Joint End-to-end Learning of Pose and Emotion. *ACM Transactions on Graphics* 36, 4, Article 94 (July 2017), 12 pages.

Tero Karras, Samuli Laine, and Timo Aila. 2019b. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019a. Analyzing and Improving the Image Quality of StyleGAN. *arXiv preprint arXiv:11912.04958* (2019).

Michael Keller, Reinhard Knothe, and Thomas Vetter. 2007. 3D reconstruction of human faces from occluding contours. In *MIRAGE*. Springer, 261–273.

Ira Kemelmacher-Shlizerman. 2016. Transfiguring Portraits. *ACM Transactions on Graphics* 35, 4, Article 94 (July 2016), 8 pages.

Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M. Seitz. 2010. Being John Malkovich. In *Proc. European Conference on Computer Vision (ECCV)*, Vol. 6311. Springer, 341–353.

Ira Kemelmacher-Shlizerman and Steven M. Seitz. 2011. Face Reconstruction in the Wild. In *Proc. International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Washington, DC, USA, 1746–1753. https://doi.org/10.1109/ICCV.2011.6126439

Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. 2015. Learning an efficient model of hand shape variation from depth images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2540–2548.

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

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

Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE*

Transactions on Information Forensics and Security 7, 6 (2012), 1789–1801.

Martin Klaudiny, Steven McDonagh, Derek Bradley, Thabo Beeler, and Kenny Mitchell. 2017. Real-Time Multi-View Facial Capture with Synthetic Training. In *Compututer Graphics Forum*, Vol. 36. 325–336.

Reinhard Knothe, Sami Romdhani, and Thomas Vetter. 2006. Combining PCA and LFA for Surface Reconstruction from a Sparse Set of Control Points. In *Proc. International Conference on Automatic Face and Gesture Recognition*. 637–644.

Paul Koppen, Zhen-Hua Feng, Josef Kittler, Muhammad Awais, William Christmas, Xiao-Jun Wu, and He-Feng Yin. 2018. Gaussian mixture 3D morphable face model. *Pattern Recognition* 74 (2018), 617–628.

Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2018a. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2093–2102.

Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2019. Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Adam Kortylewski, Andreas Schneider, Thomas Gerig, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. 2018b. Training deep face recognition systems with synthetic data. *arXiv preprint arXiv:1802.05891* (2018).

Adam Kortylewski, Mario Wieser, Andreas Morel-Forster, Aleksander Wieczorek, Sonali Parbhoo, Volker Roth, and Thomas Vetter. 2018c. Informed MCMC with Bayesian Neural Networks for Facial Image Analysis. *Proc. Advances in neural information processing systems workshops (NeurIPS)* (2018).

Jana Koudelová, Ján Dupej, Jaroslav Bružek, Petr Sedlak, and Jana Velemínská. 2015. Modelling of facial growth in Czech children based on longitudinal data: Age progression from 12 to 15 years using 3D surface models. *Forensic Science International* 248 (2015), 33–40.

Aravind Krishnaswamy and Gladimir VG Baranoski. 2004. A biophysically-based spectral model of light interaction with human skin. *Compututer Graphics Forum* 23, 3 (2004), 331–340.

Sumedha Kshirsagar and Nadia Magnenat-Thalmann. 2003. Visyllable Based Speech Animation. *Compututer Graphics Forum* 22, 3 (2003), 632–640.

Samuli Laine, 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 *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*. ACM, New York, NY, USA, Article 10, 10 pages.

Erik Learned-Miller, Qifeng Lu, Angela Paisley, Peter Trainer, Volker Blanz, Katrin Dedden, and Ralph Miller. 2006. Detecting acromegaly: screening for disease with a morphable model. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 495–503.

Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (2005), 684–698.

David A Leopold, Igor V Bondar, and Martin A Giese. 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442, 7102 (2006), 572.

David A Leopold, Alice J O'Toole, Thomas Vetter, and Volker Blanz. 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Machine Intelligence* 4, 1 (2001), 89.

Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. 2000. The digital Michelangelo project: 3D scanning of large statues. In *Proc. Conference on Computer graphics and Interactive Techniques*. ACM Press/Addison-Wesley Publishing Co., 131–144.

JP Lewis, Zhenyao Mo, Ken Anjyo, and Taehyun Rhee. 2014b. Probable and improbable faces. In *Mathematical Progress in Expressive Image Synthesis I*. 21–30.

J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. 2014a. Practice and Theory of Blendshape Facial Models. In *Computer Graphics Forum (Eurographics State of the Art Reports)*.

Chen Li, Kun Zhou, and Stephen Lin. 2014b. Intrinsic Face Image Decomposition with Human Face Priors. In *Proc. European Conference on Computer Vision (ECCV)*, Vol. 8693. Springer, 218–233.

Chen Li, Kun Zhou, and Stephen Lin. 2015b. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 4621–4629.

Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015a. Facial Performance Sensing Head-Mounted Display. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 34, 4 (July 2015).

Hao Li, Thibaut Weise, and Mark Pauly. 2010. Example-based facial rigging. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 29, 4 (2010), 32:1–32:6.

Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime Facial Animation with On-the-fly Correctives. *ACM Transactions on Graphics* 32, 4 (2013), 42:1–42:10.

Kai Li, Qionghai Dai, Ruiping Wang, Yebin Liu, Feng Xu, and Jue Wang. 2014a. A Data-Driven Approach for Facial Expression Retargeting in Video. *IEEE Transactions on Multimedia* 16, 2 (2014), 299–310.

Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. 2012. A data-driven approach for facial expression synthesis in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 57–64.

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

Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. 2018. Differentiable monte carlo ray tracing through edge sampling. (2018), 222.

Shu Liang, Ira Kemelmacher-Shlizerman, and Linda G. Shapiro. 2014. 3D Face Hallucination from a Single Depth Frame. In *Proc. IEEE International Conference on 3D Vision (3DV)*, Vol. 1. 31–38. https://doi.org/10.1109/ThreeDV.2014.67

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

Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Towards High-Fidelity 3D Face Reconstruction from In-the-Wild Images Using Graph Convolutional Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Feng Liu, Luan Tran, and Xiaoming Liu. 2019b. 3D Face Modeling from Diverse Raw Scan Data. In *Proc. International Conference on Computer Vision (ICCV)*.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019a. Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning. In *Proc. International Conference on Computer Vision (ICCV)*.

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. International Conference on Machine Learning*.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics* 37, 4 (2018), 68.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Linjie Luo, Hao Li, and Szymon Rusinkiewicz. 2013. Structure-aware hair capture. *ACM Transactions on Graphics* 32, 4 (2013), 76.

Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. 2018. Gaussian process morphable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 8 (2018), 1860–1873.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. 2007. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proc. Eurographics Workshops*. 183–194.

Dennis Madsen, Marcel Lüthi, Andreas Schneider, and Thomas Vetter. 2018. Probabilistic Joint Face-Skull Modelling for Facial Reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5295–5303.

Stephen R Marschner, Henrik Wann Jensen, Mike Cammarano, Steve Worley, and Pat Hanrahan. 2003. Light scattering from human hair fibers. *ACM Transactions on Graphics* 22, 3 (2003), 780–791.

Stephen R Marschner, Stephen H Westin Eric PF Lafortune, and Kenneth E Torrance Donald P Greenberg. 1999. Image-Based BRDF Measurement Including Human Skin. In *Proc. Eurographics Workshops*. 131.

Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. 2016. Do we really need to collect millions of faces for effective face recognition?. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 579–596.

Bogdan J. Matuszewski, Wei Quan, Lik-Kwan Shark, Alison S. McLoughlin, Catherine E. Lightbody, Hedley C.A. Emsley, and Caroline L. Watkins. 2012. Hi4D-ADSIP 3-D dynamic facial articulation database. *Image and Vision Computing* 30, 10 (2012), 713 – 727. 3D Facial Behaviour Analysis and Understanding.

Steven McDonagh, Martin Klaudiny, Derek Bradley, Thabo Beeler, Iain Matthews, and Kenny Mitchell. 2016. Synthetic Prior Design for Real-Time Face Tracking. In *Proc. IEEE International Conference on 3D Vision (3DV)*. 639–648.

Baback Moghaddam, Jinho Lee, Hanspeter Pfister, and Raghu Machiraju. 2003. Model-Based 3D Face Capture with Shape-from-Silhouettes. In *Proc. International Conference on Computer Vision (ICCV) Workshops*. IEEE Computer Society, 20.

Andreas Morel-Forster. 2016. *Generative shape and image analysis by combining Gaussian processes and MCMC sampling*. Ph.D. Dissertation. University of Basel.

Stylianos Moschoglou, Evangelos Ververas, Yannis Panagakisand Mihalis A. Nicolaou, , and Stefanos Zafeiriou. 2018. Multi-attribute robust component analysis for facial uv maps. *IEEE Journal of Selected Topics in Signal Processing* 12, 6 (2018), 1324–1337.

Iordanis Mpiperis, Sotiris Malassiotis, and Michael G. Strintzis. 2008. Bilinear Models for 3-D Face and Facial Expression Recognition. *IEEE Transactions on Information Forensics and Security* 3, 3 (2008), 498–511.

Andreas Mueller, Pascal Paysan, Ralf Schumacher, Hans-Florian Zeilhofer, Isabelle Berg-Boerner, Juerg Maurer, Thomas Vetter, Erik Schkommodau, Philipp Juergens, and Katja Schwenzer-Zimmerer. 2011. Missing facial parts computed by a morphable model and transferred directly to a polyamide laser-sintered prosthesis: an innovation study. *British Journal of Oral and Maxillofacial Surgery* 49, 8 (2011), e67–e71.

David Mumford and Agnès Desolneux. 2010. *Pattern theory: the stochastic analysis of real-world signals.* AK Peters/CRC Press.

Koki Nagano, Huiwen Luo, Zejian Wang, Jeawoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. 2019. Deep Face Normalization. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 38, 6 (2019), 183:1–16.

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. 2018. paGAN: real-time avatars using dynamic textures. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*. ACM, 258.

Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. 2005. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 24, 3 (2005), 536–543.

Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. 2013. Sparse Localized Deformation Components. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 32, 6 (2013), 179:1–179:10.

Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time dense surface mapping and tracking. In *Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 127–136.

Arthur Niswar, Ishtiaq Khan, and Farzam Farbiz. 2011. Virtual try-on of eyeglasses using 3D model of the head. *Proc. International Conference on Virtual Reality Continuum and Its Applications in Industry* (12 2011). https://doi.org/10.1145/2087756.2087838

Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. 2016. High-Fidelity Facial and Speech Animation for VR HMDs. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 35, 6 (December 2016).

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proc. British Machine Vision Conference (BMVC)*.

Paysan Pascal. 2010. *Statistical modeling of facial aging based on 3D scans.* Ph.D. Dissertation. University of Basel.

Georgios Passalis, Panagiotis Perakis, Theoharis Theoharis, and Ioannis A. Kakadiaris. 2011. Using Facial Symmetry to Handle Pose Variations in Real-World 3D Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 10 (2011), 1938–1951.

Ankur Patel and William A.P. Smith. 2009. 3D morphable face models revisited. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1327–1334.

Ankur Patel and William A.P. Smith. 2011. Simplification of 3D morphable models. In *Proc. International Conference on Computer Vision (ICCV)*. 271–278.

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

Ankur Patel and William AP Smith. 2016. Manifold-based constraints for operations in face space. *Pattern Recognition* 52 (2016), 206–217.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009a. A 3D face model for pose and illumination invariant face recognition. In *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*. Ieee, 296–301.

Pascal Paysan, Marcel Lüthi, Thomas Albrecht, Anita Lerch, Brian Amberg, Francesco Santini, and Thomas Vetter. 2009b. Face reconstruction from skull shapes and physical attributes. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposum (DAGM)*. Springer, 232–241.

P Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M Blackburn, Elham Tabassi, and Mike Bone. 2003. Face recognition vendor test 2002. In *Proc. International SOI Conference*. IEEE.

Jean-Sébastien Pierrard. 2008. *Skin segmentation for robust face image analysis.* Ph.D. Dissertation. University of Basel.

Jean-Sébastien Pierrard and Thomas Vetter. 2007. Skin detail analysis for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fred Pighin and J.P. Lewis. 2006. Performance-Driven Facial Animation. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*.

Marcel Piotraschke and Volker Blanz. 2016. Automated 3D Face Reconstruction from Multiple Images Using Quality Measures. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3418–3427. https://doi.org/10.1109/CVPR.2016.372

Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A. P. Smith, and Stefanos Zafeiriou. 2019. Combining 3D Morphable Models: A Large Scale Face-And-Head Model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-aware Facial Animation from a Single Image. In *Proc. European Conference on Computer Vision (ECCV)*.

Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. ACM, 497–500.

Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. 2018. Generating 3D faces using Convolutional Mesh Autoencoders. In *Proc. European Conference on Computer Vision (ECCV)*. 725–741.

Elad Richardson, Matan Sela, and Ron Kimmel. 2016. 3D face reconstruction by learning from synthetic data. In *Proc. IEEE International Conference on 3D Vision (3DV)*. 460–469.

Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. 2017. Learning detailed face reconstruction from a single image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1259–1268.

Sami Romdhani, Volker Blanz, and Thomas Vetter. 2002. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 3–19.

Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. 986–993 vol. 2. https://doi.org/10.1109/CVPR.2005.145

Fabiano Romeiro and Todd Zickler. 2007. Model-based stereo with occlusions. In *Proc. International Conference on Automatic Face and Gesture Recognition*. Springer, 31–45.

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

Joseph Roth, Yiying Tong, and Xiaoming Liu. 2015. Unconstrained 3D Face Reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA.

Joseph Roth, Yiying Tong, and Xiaoming Liu. 2016. Adaptive 3D Face Reconstruction from Unconstrained Photo Collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 11 (December 2016), 2127–2141.

Shunsuke Saito, Liwen Hu, Chongyang Ma, Hikaru Ibayashi, Linjie Luo, and Hao Li. 2018. 3D hair synthesis using volumetric variational autoencoders. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 37, 6 (2018).

Shunsuke Saito, Tianye Li, and Hao Li. 2016. Real-Time Facial Segmentation and Performance Capture from RGB Input. In *Proc. European Conference on Computer Vision (ECCV)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 244–261.

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2017. Photorealistic facial texture inference using deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5144–5153.

Augusto Salazar, Stefanie Wuhrer, Chang Shu, and Flavio Prieto. 2014. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications* 25, 4 (2014), 859–879.

Dalila Sánchez-Escobedo, Mario Castelán, and William AP Smith. 2016. Statistical 3D face shape estimation from occluding contours. *Computer Vision and Image Understanding* 142 (2016), 111–124.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. 2019. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011. Real-time avatar animation from a single image. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, 117–124.

Arman Savran, Nese Alyuöz, Hamdi Dibeklioglu, Oya Celiktutan, Berk Gökberk, Bülent. Sankur, and Lale Akarun. 2008. Bosphorus database for 3D face analysis. In *Proc. European Workshop on Biometrics and Identity Management*. 47–56.

Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. 2011. Computer-Suggested Facial Makeup. *Compututer Graphics Forum* (2011).

Kristina Scherbaum, Martin Sunkel, H-P Seidel, and Volker Blanz. 2007. Prediction of individual non-linear aging trajectories of faces. In *Compututer Graphics Forum*, Vol. 26. Wiley Online Library, 285–294.

Andreas Schneider, Ghazi Bouabene, Ayet Shaiek, Sandro Schönborn, and Thomas Vetter. 2019. Photo-Realisitc Exemplar-Based Face Aging. *Proc. International Conference on Automatic Face and Gesture Recognition* (2019).

Andreas Schneider, Bernhard Egger, and Thomas Vetter. 2018. A Parametric Freckle Model for Faces. In *Proc. International Conference on Automatic Face and Gesture Recognition*.

Andreas Schneider, Sandro Schönborn, Lavrenti Frobeen, Bernhard Egger, and Thomas Vetter. 2017. Efficient global illumination for morphable models. In *Proc. International Conference on Computer Vision (ICCV)*. 3865–3873.

Sandro Schönborn, Bernhard Egger, Andreas Forster, and Thomas Vetter. 2015. Background modeling for generative image models. *Computer Vision and Image Understanding* 136 (2015), 117–127.

Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. 2017. Markov Chain Monte Carlo for Automated Face Image Analysis. *International Journal of Computer Vision* 123, 2 (01 Jun 2017), 160–183. https://doi.org/10.1007/s11263-016-0967-5

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823.

Matthaeus Schumacher and Volker Blanz. 2012. Which facial profile do humans expect after seeing a frontal view? a comparison with a linear face model. *ACM Transactions on Applied Perception* 9, 3 (2012), 11.

Matthaeus Schumacher and Volker Blanz. 2015. Exploration of the correlations of attributes and features in faces. In *Proc. International Conference on Automatic Face and Gesture Recognition*, Vol. 1. IEEE, 1–8.

Alassane Seck, William AP Smith, Arnaud Dessein, Bernard Tiddeman, Hannah Dee, and Abhishek Dutta. 2016. Ear-to-ear capture of facial intrinsics. *arXiv preprint arXiv:1609.02368* (2016).

Matan Sela, Elad Richardson, and Ron Kimmel. 2017. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *Proc. International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 1585–1594.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. SfSNet: Learning Shape, Reflectance and Illuminance of Facesin the Wild'. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6296–6305.

Gil Shamai, Ron Slossberg, and Ron Kimmel. 2020. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 3 (2020), #87:1–24.

Christian R Shelton. 2000. Morphable surface models. *International Journal of Computer Vision* 38, 1 (2000), 75–91.

Cheng-Ta Shen, Fay Huang, Wan-Hua Lu, Sheng-Wen Shih, and Hong-Yuan Mark Liao. 2014. 3D Age Progression Prediction in Children's Faces with a Small Exemplar-Image Set. *Journal of Information Science & Engineering* 30, 4 (2014).

Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics* 33, 6 (2014), 222.

Il-Kyu Shin, A Cengiz Öztireli, Hyeon-Joong Kim, Thabo Beeler, Markus Gross, and Soo-Mi Choi. 2014. Extraction and transfer of facial expression wrinkles for facial performance enhancement. In *Proc. of The Pacific Conference on Computer Graphics and Applications*. 113–118.

Lawrence Sirovich and Michael Kirby. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A* 4, 3 (1987), 519–524.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *Proc. Advances in neural information processing systems (NeurIPS)*.

Peter-Pike Sloan, Jan Kautz, and John Snyder. 2002. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *ACM Transactions on Graphics* 21, 3 (2002), 527–536.

Ron Slossberg, Gil Shamai, and Ron Kimmel. 2018. High quality facial surface and texture synthesis via generative adversarial networks. In *Proc. European Conference on Computer Vision (ECCV)*. 0–0.

Michael De Smet and Luc Van Gool. 2010. Optimal regions for linear model-based 3D face reconstruction. In *Asian Conference on Computer Vision (ACCV)*. 276–289.

William AP Smith. 2016. The perspective face shape ambiguity. In *Perspectives in Shape Analysis*. Springer, 299–319.

William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. 2020. A Morphable Face Albedo Model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. 2011. Effect of illumination on automatic expression recognition: a novel 3D relightable facial database. In *Proc. International Conference on Automatic Face and Gesture Recognition*. IEEE, 611–618.

Martin A. Styner, Kumar T. Rajamani, Lutz-Peter Nolte, Gabriel Zsemlye, Gábor Székely, Christopher J. Taylor, and Rhodri H. Davies. 2003. Evaluation of 3D Correspondence Methods for Model Building. In *Proc. International conference on Information Processing in Medical Imaging (IPMI)*, Chris Taylor and J. Alison Noble (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 63–75.

Yi Sun, Xiaochen Chen, Matthew Rosato, and Lijun Yin. 2010. Tracking Vertex Flow and Model Adaptation for Three-Dimensional Spatiotemporal Face Analysis. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 3 (2010), 461–474.

Yifan Sun and Noboru Murata. 2020. CAFM: A 3D Morphable Model for Animals. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Michael Suttie, Tatiana Foroud, Leah Wetherill, Joseph L Jacobson, Christopher D Molteno, Ernesta M Meintjes, H Eugene Hoyme, Nathaniel Khaole, Luther K Robinson, Edward P Riley, et al. 2013. Facial dysmorphism across the fetal alcohol spectrum. *Pediatrics* 131, 3 (2013), e779.

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. 2014. Total Moving Face Reconstruction. In *Proc. European Conference on Computer Vision (ECCV)*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 796–812.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2015. What makes tom hanks look like tom hanks. In *Proc. International Conference on Computer Vision (ICCV)*. 3952–3960.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics* 36, 4, Article 95 (July 2017), 13 pages.

Attila Szabó, Givi Meishvili, and Paolo Favaro. 2019. Unsupervised Generative 3D Shape Learning from Natural Images. *arXiv preprint arXiv:1910.00287* (2019).

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1701–1708.

Gary K.L. Tam, Zhi-Quan Cheng, Yu-Kun Lai, Frank C. Langbein, Yonghuai Liu, David Marshall, Ralph R. Martin, Xian-Fang Sun, and Paul L. Rosin. 2013. Registration of 3D point clouds and meshes: A survey from rigid to Nonrigid. *Transactions on Visualization and Computer Graphics* 19, 7 (2013), 1199–1217.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A Deep Learning Approach for Generalized Speech Animation. *ACM Transactions on Graphics* 36, 4 (2017), 93:1–93:11.

Jose Rafael Tena, Fernando De la Torre, and Iain Matthews. 2011. Interactive Region-based Linear 3D Face Models. *ACM Transactions on Graphics* 30, 4 (July 2011), 76:1–76:10.

Jose Rafael Tena, Raymond S Smith, Miroslav Hamouz, Josef Kittler, Adrian Hilton, and John Illingworth. 2007. 2d face pose normalisation using a 3d morphable model. In *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 51–56.

Frank B ter Haar and Remco C Veltkamp. 2008. 3D face model fitting for recognition. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 652–664.

Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhoefer, and Christian Theobalt. 2019. FML: Face Model Learning from Videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. 2018. High-Fidelity Monocular Face Reconstruction based on an Unsupervised Model-based Face Autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 1–1. https://doi.org/10.1109/TPAMI.2018.2876842

Ayush Tewari, Michael Zollhoefer, 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2549–2559.

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

Barry-John Theobald, Iain Matthews, Michael Mangini, Jeffrey R Spies, Timothy R Brick, Jeffrey F Cohn, and Steven M Boker. 2009. Mapping and manipulating facial expression. *Language and Speech* 52, 2–3 (2009), 369–386.

Justus Thies, Michael Zollhoefer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time Expression Transfer for Facial Reenactment. *ACM Transactions on Graphics* 34, 6 (2015), 183:1–183:14.

Justus Thies, Michael Zollhoefer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

Justus Thies, Michael Zollhoefer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2018a. FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality. *ACM Transactions on Graphics* 37, 2, Article 25 (June 2018), 15 pages. https://doi.org/10.1145/3182644

Justus Thies, Michael Zollhoefer, Marc Stamminger, Christian Theobalt, and Marc Nießner. 2018b. HeadOn: Real-time Reenactment of Human Portrait Videos. *ACM Transactions on Graphics* (2018).

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 38, 4 (2019), 1–12.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2018c. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in

Virtual Reality. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (2018).
Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5163–5172.
Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. 2018. Extreme 3D Face Reconstruction: Seeing Through Occlusions.. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3935–3944.
Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards High-fidelity Nonlinear 3D Face Morphable Model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Luan Tran and Xiaoming Liu. 2018a. Nonlinear 3D Face Morphable Model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT.
Luan Tran and Xiaoming Liu. 2018b. On learning 3d face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).
Liyun Tu, Antonio R Porras, Alec Boyle, and Marius George Linguraru. 2018. Analysis of 3D Facial Dysmorphology in Genetic Syndromes from Unconstrained 2D Photographs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 347–355.
Matthew Turk and Alex Pentland. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3, 1 (1991), 71–86.
Oliver van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. 2011. A Survey on Shape Correspondence. *Compututer Graphics Forum* 30, 6 (2011), 1681–1707.
Zdravko Velinov, Marios Papas, Derek Bradley, Paulo Gotardo, Parsa Mirdehghan, Steve Marschner, Jan Novák, and Thabo Beeler. 2018. Appearance Capture and Modeling of Human Teeth. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 37, 6 (2018), 207:1–207:13.
Thomas Vetter and Tomaso Poggio. 1997. Linear Object Classes and Image Synthesis from a Single Example Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1997), 733–742.
Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005a. Face transfer with multilinear models. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 24, 3 (2005), 426–433.
Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005b. Face Transfer with Multilinear Models. *ACM Transactions on Graphics* 24, 3 (2005), 426–433.
Mirella Walker, Fang Jiang, Thomas Vetter, and Sabine Sczesny. 2011. Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science* 2, 6 (2011), 609–617.
Mirella Walker, Sandro Schönborn, Rainer Greifeneder, and Thomas Vetter. 2018. The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PloS one* 13, 3 (2018), e0193190.
Mirella Walker and Thomas Vetter. 2009. Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision* 9, 11 (2009), 12–12.
Christian Wallraven, Volker Blanz, and Thomas Vetter. 1999. 3D-Reconstruction of Faces: Combining Stereo with Class-Based Knowledge. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposum (DAGM)*, Wolfgang Förstner, Joachim M. Buhmann, Annett Faber, and Petko Faber (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 405–412.
Mengjiao Wang, Yannis Panagakis, Patrick Snape, and Stefanos Zafeiriou. 2017. Learning the Multilinear Structure of Visual Data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. 2019b. An Adversarial Neuro-Tensorial Approach for Learning Disentangled Representations. *International Journal of Computer Vision* 127 (2019), 743–762.
Ruizhe Wang, Chih-Fan Chen, Hao Peng, Xudong Liu, Oliver Liu, and Xin Li. 2019a. Digital Twin: Acquiring High-Fidelity 3D Avatar from a Single Image. *arXiv preprint arXiv:1912.03455* (2019).
Yang Wang, Xiaolei Huang, Chan-Su Lee, Song Zhang, Zhiguo Li, Dimitris Samaras, Dimitris Metaxas, Ahmed Elgammal, and Peisen Huang. 2004. High Resolution Acquisition, Learning and Transfer of Dynamic 3-D Facial Expressions. *Compututer Graphics Forum* 23, 3 (2004), 677–686.
Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2009. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (2009), 1968–1984.
Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011a. Realtime performance-based facial animation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 30, 4 (2011), 77:1–77:10.
Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011b. Realtime Performance-based Facial Animation. *ACM Transactions on Graphics* 30, 4 (2011), 77:1–77:10.
Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. 2009. Face/Off: Live Facial Puppetry. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*. ACM, 7–16.
Cyrus A Wilson, Abhijeet Ghosh, Pieter Peers, Jen-Yuan Chiang, Jay Busch, and Paul Debevec. 2010. Temporal upsampling of performance geometry using photometric alignment. *ACM Transactions on Graphics* 29, 2 (2010), 17.
Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhoefer, Christian Theobalt, Markus Gross, and Thabo Beeler. 2016a. Model-Based Teeth Reconstruction. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 35, 6 (2016).
Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016b. An anatomically-constrained local deformation model for monocular face capture. 35, 4 (2016), 115:1–12.
Zexiang Xu, Hsiang-Tao Wu, Lvdi Wang, Changxi Zheng, Xin Tong, and Yue Qi. 2014. Dynamic hair capture using spacetime optimization. *ACM Transactions on Graphics* 33 (2014), 6.
Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics* 37, 4 (2018), 162.
Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas. 2012. Facial expression editing in video using a temporally-smooth factorization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 861–868.
Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. 2020. Efficient inverse graphics in biological face processing. *Science Advances* 6, 10 (2020).
Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. 2008. A high-resolution 3D dynamic facial expression database. In *Proc. International Conference on Automatic Face and Gesture Recognition*. 1–6.
Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. 2006. A 3D Facial Expression Database for Facial Behavior Research. In *Proc. International Conference on Automatic Face and Gesture Recognition*. 211–216.
Ronald Yu, Shunsuke Saito, Haoxiang Li, Duygu Ceylan, and Hao Li. 2017. Learning Dense Facial Correspondences in Unconstrained Images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4723–4732.
Stefanos Zafeiriou, Gary A Atkinson, Mark F Hansen, William AP Smith, Vasileios Argyriou, Maria Petrou, Melvyn L Smith, and Lyndon N Smith. 2013. Face recognition and verification using photometric stereo: The photoface database and a comprehensive evaluation. *IEEE Transactions on Information Forensics and Security* 8, 1 (2013), 121–135.
Chao Zhang, William Smith, Arnaud Dessein, Nick Pears, and Hang Dai. 2016b. Functional Faces: Groupwise Dense Correspondence using Functional Maps. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Lei Zhang and Dimitris Samaras. 2006. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 3 (2006), 351–363.
Li Zhang, Noah Snavely, Brian Curless, and Steven M Seitz. 2004. Spacetime faces: high resolution capture for modeling and animation. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, Vol. 23. 548–558.
Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. 2014. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692 – 706.
Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michaele Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. 2016a. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3438–3446.
Guoyan Zheng, Shuo Li, and Gabor Szekely. 2017. *Statistical shape and deformation analysis: methods, implementation and applications*. Academic Press.
Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. 2019. Dense 3D Face Decoding over 2500FPS: Joint Texture and Shape Convolutional Mesh Decoders. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. 2015. High-fidelity pose and expression normalization for face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 787–796.
Jasenko Zivanov, Andreas Forster, Sandro Schönborn, and Thomas Vetter. 2013. Human face shape analysis under spherical harmonics illumination considering self occlusion. In *Proc. International Conference on Biometrics (ICB)*. IEEE, 1–8.
Jasenko Zivanov, Pascal Paysan, and Thomas Vetter. 2009. Facial normal map capture using four lights–an effective and inexpensive method of capturing the fine scale detail of human faces using four point lights. In *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. (GRAPP)*.
Michael Zollhoefer , Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Marc Nießner, and Christian Theobalt. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum (Eurographics State of the Art Reports)* 37, 2 (2018).
Gaspard Zoss, Thabo Beeler, Markus Gross, and Derek Bradley. 2019. Accurate markerless jaw tracking for facial performance capture. *ACM Transactions on Graphics* 38, 4 (2019), 50.

Gaspard Zoss, Derek Bradley, Pascal Bérard, and Thabo Beeler. 2018. An empirical rig for jaw animation. *ACM Transactions on Graphics* 37, 4 (2018), 59.

Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. 2018. Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.