FML: Face Model Learning from Videos

Ayush Tewari[1]   Florian Bernard[1]   Pablo Garrido[2]   Gaurav Bharaj[2]   Mohamed Elgharib[1]
Hans-Peter Seidel[1]   Patrick Pérez[3]   Michael Zollhöfer[4]   Christian Theobalt[1]

[1]MPI Informatics, Saarland Informatics Campus   [2]Technicolor   [3]Valeo.ai   [4]Stanford University

Figure 1. We propose multi-frame self-supervised training of a deep network based on in-the-wild video data for jointly learning a face model and 3D face reconstruction. Our approach successfully disentangles facial shape, appearance, expression, and scene illumination.

Abstract

Monocular image-based 3D reconstruction of faces is a long-standing problem in computer vision. Since image data is a 2D projection of a 3D face, the resulting depth ambiguity makes the problem ill-posed. Most existing methods rely on data-driven priors that are built from limited 3D face scans. In contrast, we propose multi-frame video-based self-supervised training of a deep network that (i) learns a face identity model both in shape and appearance while (ii) jointly learning to reconstruct 3D faces. Our face model is learned using only corpora of in-the-wild video clips collected from the Internet. This virtually endless source of training data enables learning of a highly general 3D face model. In order to achieve this, we propose a novel multi-frame consistency loss that ensures consistent shape and appearance across multiple frames of a subject's face, thus minimizing depth ambiguity. At test time we can use an arbitrary number of frames, so that we can perform both monocular as well as multi-frame reconstruction.

1. Introduction

The reconstruction of faces from visual data has a wide range of applications in vision and graphics, including face tracking, emotion recognition, and interactive image/video editing tasks relevant in multimedia. Facial images and videos are ubiquitous, as smart devices as well as consumer and professional cameras provide a continuous and virtually endless source thereof. When such data is captured without controlled scene location, lighting, or intrusive equipment (e.g. egocentric cameras or markers on actors), one speaks of *"in-the-wild"* images. Usually *in-the-wild* data is of low resolution, noisy, or contains motion and focal blur,

making the reconstruction problem much harder than in a controlled setup. 3D face reconstruction from *in-the-wild* monocular 2D image and video data [71] deals with disentangling facial shape identity (neutral geometry), skin appearance (or albedo) and expression, as well as estimating the scene lighting and camera parameters. Some of these attributes, *e.g.* albedo and lighting, are not easily separable in monocular images. Besides, poor scene lighting, depth ambiguity, and occlusions due to facial hair, sunglasses and large head rotations complicates 3D face reconstruction.

In order to tackle the difficult monocular 3D face reconstruction problem, most existing methods rely on the availability of strong prior models that serve as regularizers for an otherwise ill-posed problem [6, 20, 68]. Although such approaches achieve impressive facial shape and albedo reconstruction, they introduce an inherent bias due to the used face model. For instance, the 3D Morphable Model (3DMM) by Blanz et al. [6] is based on a comparably small set of 3D laser scans of Caucasian actors, thus limiting generalization to general real-world identities and ethnicities. With the rise of CNN-based deep learning, various techniques have been proposed, which in addition to 3D reconstruction also perform face model learning from monocular images [63, 62, 59, 55]. However, these methods heavily rely on a pre-existing 3DMM to resolve the inherent depth ambiguities of the monocular reconstruction setting. Another line of work, where 3DMM-like face models are not required, are based on photo-collections [30, 37, 57]. However, these methods need a very large number (*e.g.* ≈100) of facial images of the same subject, and thus they impose strong demands on the training corpus.

In this paper, we introduce an approach that learns a comprehensive face identity model using clips crawled from *in-the-wild* Internet videos [19]. This face identity model

comprises two components: One component to represent the geometry of the facial identity (modulo expressions), and another to represent the facial appearance in terms of the albedo. As we have only weak requirements on the training data (cf. Sec. 3.1), our approach can employ a virtually endless amount of community data and thus obtain a model with better generalization; laser scanning a similarly large group of people for model building would be nearly impossible. Unlike most previous approaches, we do not require a pre-existing shape identity and albedo model as initialization, but instead learn their variations from scratch. As such, our methodology is applicable in scenarios when no existing model is available, or if it is difficult to create such a model from 3D scans (*e.g.* for faces of babies).

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

1. A deep neural network that learns a facial shape and appearance space from a big dataset of unconstrained images that contains multiple images of each subject, *e.g.* multi-view sequences, or even monocular videos.

2. Explicit blendshape and identity separation by a projection onto the blendshapes' nullspace that enables a multi-frame consistency loss.

3. A novel multi-frame identity consistency loss based on a Siamese network [67], with the ability to handle monocular and multi-frame reconstruction.

2. Related Work

The literature on 3D model learning is quite vast and we mainly review methods for reconstructing 3D face models from scanner data, monocular video data, photo-collections and a single 2D image. An overview of the state-of-the-art in model-based face reconstruction is given in [71].

Morphable Models from High-quality Scans: 3DMMs represent deformations in a low-dimensional subspace and are often built from scanner data [7, 8, 36]. Traditional 3DMMs model geometry/appearance variation from limited data via PCA [7, 6, 26]. Recently, richer PCA models have been obtained from large-scale datasets [13, 44]. Multilinear models generalize statistical models by capturing a set of mutually orthogonal variation modes (e.g., global and local deformations) via a tensor decompo-

sition [68, 9, 10]. However, unstructured subspaces or even tensor generalizations are incapable of modeling localized deformations from limited data. In this respect, Neumann et al. [41] and Bernard et al. [5] devise methods for computing sparse localized deformation components directly from mesh data. Lüthi et al. [38] propose the so-called Gaussian Process morphable models (GPMMs), which are modeled with arbitrary non-linear kernels, to handle strong non-linear shape deformations. Ranjan et al. [46] learn a non-linear model using a deep mesh autoencoder with fast spectral convolution kernels. Garrido et al. [25] train radial basis functions networks to learn a corrective 3D lip model from multiview data. In an orthogonal direction, Li et al. [36] learn a hybrid model that combines a linear shape space with articulated motions and semantic blendshapes. All these methods mainly model shape deformations and are limited to the availability of scanner data.

Parametric Models from Monocular Data: Here, we distinguish between personalized, corrective, and morphable model learning. Personalized face models have been extracted from monocular video by first refining a parametric model in a coarse-to-fine manner (e.g., as in [49]) and then learning a mapping from coarse semantic deformations to finer non-semantic detail layers [28, 24]. Corrective models represent out-of-space deformations (e.g., in shape or appearance) which are not modeled by the underlying parametric model. Examples are adaptive linear models customized over a video sequence [15, 27] or non-linear models learned from a training corpus [48, 59]. A number of works have been proposed for in-the-wild 3DMM learning [53, 63, 4, 12]. Such solutions decompose the face into its intrinsic components through encoder-decoder architectures that exploit weak supervision. Tran et al. [63] employ two separate convolutional decoders to learn a non-linear model that disentangles shape from appearance. Similarly, Sengupta et al. [53] propose residual blocks to produce a complete separation of surface normal and albedo features. There also exist approaches that learn 3DMMs of rigid [65] or articulated objects [29] by leveraging image collections. These methods predict an instance of a 3DMM directly from an image [29] or use additional cues (e.g., segmentation and shading) to fit and refine a 3DMM [65].

Monocular 3D Reconstruction: Optimization-based reconstruction algorithms rely on a personalized model [18, 21, 23, 69] or a parametric prior [2, 15, 35, 24, 54] to estimate 3D geometry from a 2D video. Learning-based approaches regress 3D face geometry from a single image by learning an image-to-parameter or image-to-geometry mapping [42, 48, 60, 59, 52, 64, 32]. These methods require ground truth face geometry [64, 34], a morphable model from which synthetic training images are generated [47, 48, 52, 32], or a mixture of both [39, 33]. Recently, Tewari et al. [60] trained fully unsupervised through an

inverse rendering-based loss. However, color and shape variations lie in the subspace of a parametric face prior. Only very recent methods for monocular face reconstruction [59, 63, 62, 12] allow for out-of-space model generalization while training from in-the-wild data.

3D Reconstruction via Photo-collections: Face reconstruction is also possible by fitting a template model to photo-collections. In [31], an average shape and appearance model is reconstructed from a person-specific photo-collection via low-rank matrix factorization. Suwajanakorn et al. [57] use this model to track detailed facial motion from unconstrained video. Kemelmacher-Shlizerman [30] learns a 3DMM from a large photo-collection of people, grouped into a fixed set of semantic labels. Also, Liang et al. [37] leverage multi-view person-specific photo-collections to reconstruct the full head. In a different line of research, Thies et al. [61] fit a coarse parametric model to user-selected views to recover personalized face shape and albedo. Roth et al. [49] personalize an existing morphable model to an image collection by using a coarse-to-fine photometric stereo formulation. Note that most of these methods do not learn a general face model, e.g. a shape basis that spans the range of facial shapes of an entire population, but instead they obtain a single person-specific 3D face instance. Besides, these methods require curated photo-collections. We, on the contrary, build a 3DMM representation that generalizes across multiple face identities and impose weaker assumptions on the training data.

Multi-frame 3D Reconstruction: Multi-frame reconstruction techniques exploit either temporal information or multiple views to better estimate 3D geometry. Shi et al. [54] globally fit a multilinear model to 3D landmarks at multiple keyframes and enforce temporal consistency of in-between frames via interpolation. In [24], person-specific facial shape is obtained by averaging per-frame estimates of a parametric face model. Ichim et al. [28] employ a multi-view bundle adjustment approach to reconstruct facial shape and refine expressions using actor-specific sequences. Piotraschke et al. [43] combine region-wise reconstructions of a 3DMM from many images using a normal distance function. Garg et al. [22] propose a model-free approach that globally optimizes for dense 3D geometry in a non-rigid structure from motion framework. Beyond faces, Tulsian et al. [66] train a CNN to predict single-view 3D shape (represented as voxels) using multi-view ray consistency.

3. Face Model Learning

Our novel face model learning approach solves two tasks: it jointly learns (i) a parametric face geometry and appearance model, and (ii) an estimator for facial shape, expression, albedo, rigid pose and incident illumination parameters. An overview of our approach is shown in Fig. 2.

Training: Our network is trained in a self-supervised fashion based on a training set of multi-frame images, i.e., multiple images of the same person sampled from a video clip, see Section 3.1. The network jointly learns an appearance and shape identity model (Section 3.2). It also estimates per-frame parameters for the rigid head pose, illumination, and expression parameters, as well as shape and appearance identity parameters that are shared among all frames. We train the network based on a differentiable renderer that incorporates a per-vertex appearance model and a graph-based shape deformation model (Section 3.3). To this end, we propose a set of training losses that account for geometry smoothness, photo-consistency, sparse feature alignment and appearance sparsity, see Section 3.4.

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

3.1. Dataset

We train our approach using the VoxCeleb2 multi-frame video dataset [19]. This dataset contains over 140k videos of over 6000 celebrities crawled from Youtube. We sample a total of $N = 404$k multi-frame images $\mathcal{F}_1, \ldots, \mathcal{F}_N$ from this dataset. The ℓ-th multi-frame image $\mathcal{F}_\ell = \{F_\ell^{[f]}\}_{f=1}^M$ comprises $M = 4$ frames $F_\ell^{[1]}, \ldots, F_\ell^{[M]}$ of the same person ℓ extracted from the same video clip to avoid unwanted variations, e.g., due to aging or accessories. The same person can appear multiple times in the dataset. To obtain these images, we perform several sequential steps. First, the face region is cropped based on automatically detected facial landmarks [50, 51]. Afterwards, we discard images whose cropped region is smaller than a threshold (i.e., 200 pixels) and that have low landmark detection confidence, as provided by the landmark tracker [50, 51]. The remaining crops are re-scaled to 240×240 pixels. When sampling the M frames in $\mathcal{F}_\ell$, we ensure sufficient diversity in head pose based on the head orientation obtained by the landmark tracker. We split our multi-frame dataset $\mathcal{F}_1, \ldots, \mathcal{F}_N$ into a training (383k images) and test set (21k images).

3.2. Graph-based Face Representation

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

Learnable Graph-based Identity Model: Rather than learning the identity model on the high-res mesh $\mathcal{V}$ with $|\mathcal{V}| = 60$k vertices, we simplify this task by considering a lower-dimensional parametrization based on deformation
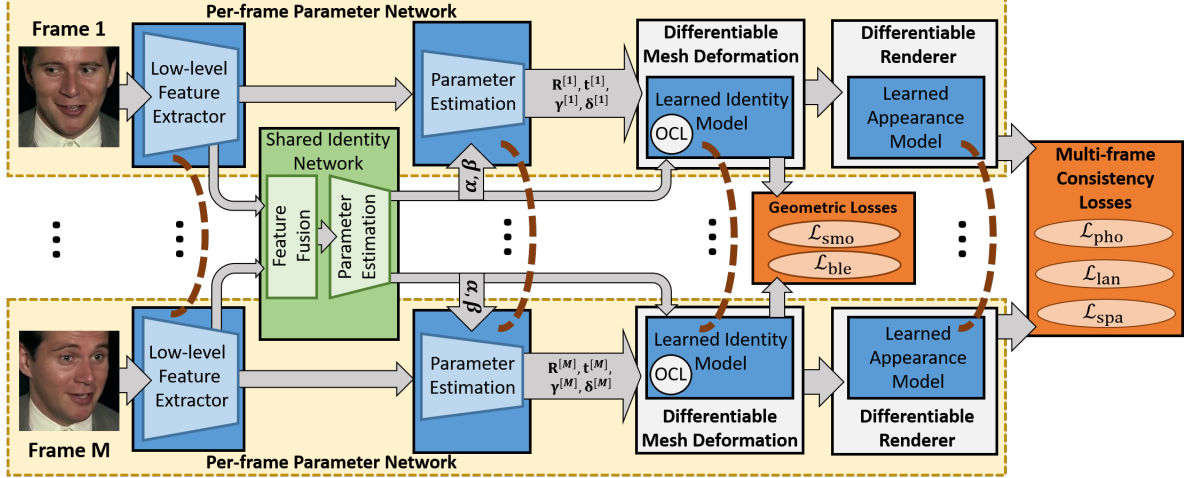
Figure 2. Pipeline overview. Given multi-frame input that shows a person under different facial expression, head pose, and illumination, our approach first estimates these parameters per frame. In addition, it jointly obtains the shared identity parameters that control facial shape and appearance, while at the same time learning a graph-based geometry and a per-vertex appearance model. We use a differentiable mesh deformation layer in combination with a differentiable face renderer to implement a model-based face autoencoder.

graphs [56]. We obtain our (coarse) deformation graph $\mathcal{G}$ by downsampling the mesh to $|\mathcal{G}| = 521$ nodes, see Fig. 3. The network now learns a deformation on $\mathcal{G}$ that is then transferred to the mesh $\mathcal{V}$ via linear blend skinning. The vector $\mathbf{g} \in \mathbb{R}^{3|\mathcal{G}|}$ of the $|\mathcal{G}|$ stacked node positions of the 3D graph is defined as

$$\mathbf{g} = \bar{\mathbf{g}} + \boldsymbol{\Theta}_s \boldsymbol{\alpha} \,, \qquad (1)$$

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

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

$$\mathbf{v}(\boldsymbol{\Theta}_s, \boldsymbol{\alpha}) = \bar{\mathbf{v}} + \mathbf{S} \boldsymbol{\Theta}_s \boldsymbol{\alpha} \,. \qquad (2)$$

Here, $\bar{\mathbf{v}} \in \mathbb{R}^{3|\mathcal{V}|}$ is fixed to the neutral mean face shape as defined in the 3DMM [7]. The skinning matrix $\mathbf{S} \in \mathbb{R}^{3|\mathcal{V}| \times 3|\mathcal{G}|}$ is obtained based on the mean shape $\bar{\mathbf{v}}$ and mean graph nodes $\bar{\mathbf{g}}$.

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

Blendshape Expression Model: For capturing facial expressions, we use a linear blendshape model that combines the facial expression models from [3] and [16]. This model is fixed, i.e. not learned. Hence, the expression deformations are directly applied to the high-res mesh. The

vertex positions of the high-res mesh that account for shape identity as well as the facial expression are given by

$$\mathbf{v}(\boldsymbol{\Theta}_s, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \bar{\mathbf{v}} + \mathbf{S} \cdot \mathrm{OCL}(\boldsymbol{\Theta}_s) \boldsymbol{\alpha} + \mathbf{B} \boldsymbol{\delta} \,, \qquad (3)$$

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

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

$$\mathrm{OCL}(\boldsymbol{\Theta}_s) = \mathrm{proj}_{\mathbf{B}_{\mathcal{G}}^\perp}(\boldsymbol{\Theta}_s) = \boldsymbol{\Theta}_s - \mathrm{proj}_{\mathbf{B}_{\mathcal{G}}}(\boldsymbol{\Theta}_s) \qquad (4)$$

$$= \boldsymbol{\Theta}_s - \mathbf{B}_{\mathcal{G}}(\mathbf{B}_{\mathcal{G}}^T \mathbf{B}_{\mathcal{G}})^{-1} \mathbf{B}_{\mathcal{G}}^T \boldsymbol{\Theta}_s \,. \qquad (5)$$

The property $\mathbf{B}_{\mathcal{G}}^T \mathrm{OCL}(\boldsymbol{\Theta}_s) = \mathbf{0}$ can easily be verified.

Learnable Per-vertex Appearance Model: The facial appearance is encoded in the $3|\mathcal{V}|$-dimensional vector

$$\mathbf{r}(\boldsymbol{\beta}) = \bar{\mathbf{r}} + \boldsymbol{\Theta}_a \boldsymbol{\beta} \qquad (6)$$

that stacks all $|\mathcal{V}|$ per-vertex colors represented as RGB triplets. The mean facial appearance $\bar{\mathbf{r}} \in \mathbb{R}^{3|\mathcal{V}|}$ and the
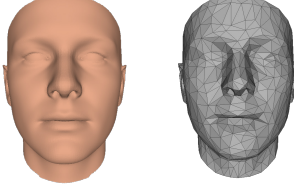
Figure 3. Neutral face shape and appearance (left), and the coarse deformation graph of the face mesh (right).

appearance basis $\mathbf{\Theta}_a \in \mathbb{R}^{3|\mathcal{V}| \times |\boldsymbol{\beta}|}$ are learnable, while the facial appearance parameters $\boldsymbol{\beta}$ are regressed. Note that we initialize the mean appearance $\bar{\mathbf{r}}$ to a constant skin tone and define the reflectance directly on the high-res mesh $\mathcal{V}$.

3.3. Differentiable Image Formation

To enable end-to-end self-supervised training, we employ a differentiable image formation model that maps 3D model space coordinates $\mathbf{v} \in \mathbb{R}^3$ onto 2D screen space coordinates $\mathbf{u} \in \mathbb{R}^2$. The mapping is implemented as $\mathbf{u} = \Pi(\Phi(\mathbf{v}))$, where Φ and Π denote the rigid head pose and camera projection, respectively. We also apply a differentiable illumination model that transforms illumination parameters $\boldsymbol{\gamma}$ as well as per-vertex appearance $\mathbf{r}_i$ and normal $\mathbf{n}_i$ into shaded per-vertex color $\mathbf{c}_i(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma})$. We explain these two models in the following.

Camera Model: We assume w.l.o.g. that the camera space corresponds to world space. We model the head pose via a rigid mapping $\Phi(\mathbf{v}) = \mathbf{R}\mathbf{v} + \mathbf{t}$, defined by the global rotation $\mathbf{R} \in SO(3)$ and the translation $\mathbf{t} \in \mathbb{R}^3$. After mapping a vertex from model space $\mathbf{v}$ onto camera space $\hat{\mathbf{v}} = \Phi(\mathbf{v})$, the full perspective camera model $\Pi : \mathbb{R}^3 \to \mathbb{R}^2$ projects the points $\hat{\mathbf{v}}$ into screen space $\mathbf{u} = \Pi(\hat{\mathbf{v}}) \in \mathbb{R}^2$.

Illumination Model: Under the assumption of distant smooth illumination and purely *Lambertian* surface properties, we employ Spherical Harmonics (SH) [45] to represent the incident radiance at a vertex $\mathbf{v}_i$ with normal $\mathbf{n}_i$ and appearance $\mathbf{r}_i$ as

$$\mathbf{c}_i(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma}) = \mathbf{r}_i \cdot \sum_{b=1}^{B^2} \boldsymbol{\gamma}_b \cdot H_b(\mathbf{n}_i) \ . \tag{7}$$

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

3.4. Multi-frame Consistent Face Model Learning

We propose a novel network for consistent multi-frame face model learning. It consists of M Siamese towers that simultaneously process M frames of the multi-frame image in different streams, see Fig. 2. Each tower consists of an encoder that estimates frame-specific parameters and identity feature maps. Note that the jointly learned geometric

identity $\mathbf{\Theta}_s$ and appearance model $(\mathbf{\Theta}_a, \bar{\mathbf{r}})$, which are common to all faces, are shared across streams.

Regressed Parameters: We train our network in a self-supervised manner based on the multi-frame images $\{\mathcal{F}_\ell\}_{\ell=1}^N$. For each frame $F_\ell^{[f]}, \forall f = 1 : M$ of the multi-frame image $\mathcal{F}_\ell$, we stack the frame-specific parameters regressed by a Siamese tower (see *Parameter Estimation* in Fig. 2) in a vector $\mathbf{p}^{[f]} = (\mathbf{R}^{[f]}, \mathbf{t}^{[f]}, \boldsymbol{\gamma}^{[f]}, \boldsymbol{\delta}^{[f]})$ that parametrizes rigid pose, illumination and expression. The frame-independent person-specific identity parameters $\hat{\mathbf{p}} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ for the multi-frame image $\mathcal{F}_\ell$ are pooled from all the towers. We use $\mathbf{p} = (\hat{\mathbf{p}}, \mathbf{p}^{[1]}, \dots, \mathbf{p}^{[M]})$ to denote all regressed frame-independent and frame-specific parameters of $\mathcal{F}_\ell$.

Per-frame Parameter Estimation Network: We employ a convolutional network to extract low-level features. We then apply a series of convolutions, ReLU, and fully connected layers to regress the per-frame parameters $\mathbf{p}^{[f]}$. We refer to the supplemental document for further details.

Multi-frame Identity Estimation Network: As explained in Section 3.1, each frame of our multi-frame input exhibits the same face identity under different head poses and expression. We exploit this information and use a single identity estimation network (see Fig. 2) to impose the estimation of common identity parameters $\hat{\mathbf{p}}$ (shape $\boldsymbol{\alpha}$, appearance $\boldsymbol{\beta}$) for all M frames. This way, we model a hard constraint on $\hat{\mathbf{p}}$ by design. More precisely, given the frame-specific low-level features obtained by the Siamese networks we apply two additional convolution layers to extract medium-level features. The resulting M medium-level feature maps are fused into a single multi-frame feature map via average pooling. Note that the average pooling operation allows us to handle a variable number of inputs. As such, we can perform monocular or multi-view reconstruction at test time, as demonstrated in Sec. 4. This pooled feature map is then fed to an identity parameter estimation network that is based on convolution layers, ReLU, and fully connected layers. For details, we refer to the supplemental.

3.5. Loss Functions

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

$$\mathcal{L}(\mathbf{x}) = \lambda_{\text{pho}} \cdot \mathcal{L}_{\text{pho}}(\mathbf{x}) + \lambda_{\text{lan}} \cdot \mathcal{L}_{\text{lan}}(\mathbf{x}) + \tag{8}$$
$$\lambda_{\text{smo}} \cdot \mathcal{L}_{\text{smo}}(\mathbf{x}) + \lambda_{\text{spa}} \cdot \mathcal{L}_{\text{spa}}(\mathbf{x}) + \lambda_{\text{ble}} \cdot \mathcal{L}_{\text{ble}}(\mathbf{x}) \ , \tag{9}$$

which is based on two data terms (8) and three regularization terms (9). We found the weights $\lambda_\bullet$ empirically and

kept them fixed in all experiments, see supplemental document for details.

Multi-frame Photometric Consistency: One of the key contributions of our approach is to enforce multi-frame consistency of the shared identity parameters $\hat{\mathbf{p}}$. This can be thought of as solving model-based non-rigid structure-from-motion (NSfM) on each of the multi-frame inputs during training. We do this by imposing the following photometric consistency loss with respect to the frame $F^{[f]}$:

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

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

Multi-frame Landmark Consistency: To better constrain the problem, we also employ a sparse 2D landmark alignment constraint. This is based on a set of 66 automatically detected 2D feature points $\mathbf{s}_i^{[f]} \in \mathbb{R}^2$ [50, 51] in each frame $F^{[f]}$. Each feature point $\mathbf{s}_i^{[f]}$ comes with a confidence $\mathbf{c}_i^{[f]}$, so that we use the loss

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

Here, $\mathbf{u}_{\mathbf{s}_i} \in \mathbb{R}^2$ is the 2D position of the i-th mesh feature point in screen space. We use sliding correspondences, akin to [59]. Note, the position of the mesh landmarks depends both on the predicted per-frame parameters $\mathbf{p}^{[f]}$ and the shared identity parameters $\hat{\mathbf{p}}$.

Geometry Smoothness on Graph-level: We employ a linearized membrane energy [14] to define a first-order geometric smoothness prior on the displacements $\mathbf{t}_i(\hat{\mathbf{p}}) = \mathbf{g}_i(\hat{\mathbf{p}}) - \bar{\mathbf{g}}_i$ of the deformation graph nodes

$$\mathcal{L}_{\text{smo}}(\mathbf{x}) = \sum_{i=1}^{|\mathcal{G}|} \sum_{j \in \mathcal{N}_i} \left|\left|\mathbf{t}_i(\hat{\mathbf{p}}) - \mathbf{t}_j(\hat{\mathbf{p}})\right|\right|_2^2 , \qquad (10)$$

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

Appearance Sparsity: In our learned face model, skin appearance is parameterized on a per-vertex basis. To fur-

Figure 4. Our approach produces high-quality monocular reconstructions of facial geometry, reflectance and illumination by learning an optimal model from in-the-wild data. This enables us to also reconstruct facial hair and makeup.

ther constrain the underlying intrinsic decomposition problem, we employ a local per-vertex spatial reflectance sparsity prior as in [40, 11], defined as follows

$$\mathcal{L}_{\text{spa}}(\mathbf{x}) = \sum_{i=1}^{|\mathcal{V}|} \sum_{j \in \mathcal{N}_i} w_{ij} \cdot \left|\left|\mathbf{r}_i(\hat{\mathbf{p}}) - \mathbf{r}_j(\hat{\mathbf{p}})\right|\right|_2^p . \qquad (11)$$

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

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

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

Expression Regularization: To prevent over-fitting and enable a better learning of the identity basis, we regularize the magnitude of the expression parameters $\boldsymbol{\delta}$:

$$\mathcal{L}_{\text{ble}}(\mathbf{x}) = \sum_{f=1}^{M} \sum_{u=1}^{|\boldsymbol{\delta}^{[f]}|} \left(\frac{\boldsymbol{\delta}_u^{[f]}}{\sigma_{\boldsymbol{\delta} u}}\right)^2 . \qquad (12)$$

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

4. Results

We show qualitative results reconstructing geometry, reflectance and scene illumination from monocular images in Fig. 4. As our model is trained on a large corpus of multi-view images, it generalizes well to different ethnicities, even in the presence of facial hair and makeup. We
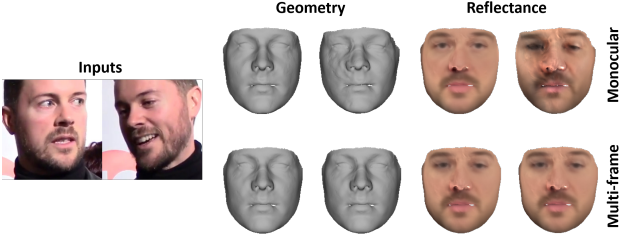
Figure 5. Monocular vs. multi-frame reconstruction. For clarity, all results are shown with a frontal pose and neutral expression. Multi-view reconstruction improves consistency and quality especially in regions which are occluded in one of the images.
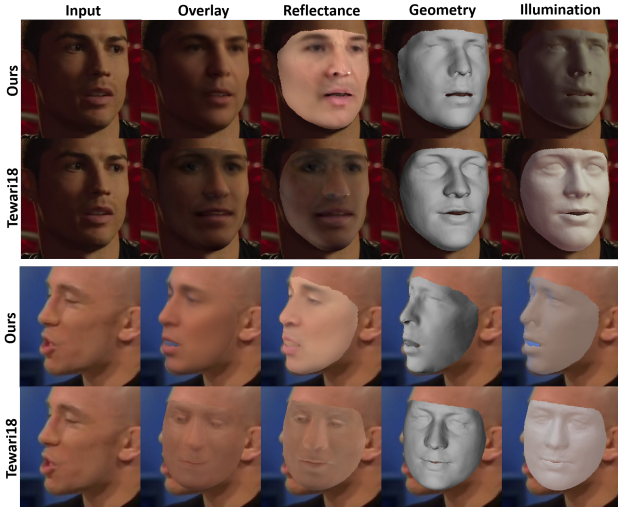

Figure 6. Comparison to Tewari et al. [59]. Multi-frame based training improves illumination estimation. Our approach also outperforms that of Tewari et al. under large poses.

implement and train our networks in TensorFlow [1]. We pre-train the expression model and then train the full network end-to-end. After convergence, the network is fine-tuned using a larger learning rate for reflectance. We empirically found that this training strategy improves the capture of facial hair, makeup and eyelids, and thus the model's generalization. Our method can also be applied to multi-frame reconstruction at test time. Fig. 5 shows that feeding two images simultaneously improves the consistency and quality of the obtained 3D reconstructions when compared to the monocular case. Please note that we can successfully separate identity and reflectance due to our novel Orthogonal Complement Layer (OCL). For the experiments shown in the following sections, we trained our network on $M = 4$ multi-frame images and used only one input image at test time, unless stated otherwise. Our networks take around 30 hours to train. Inference takes only 5.2 ms on a Titan Xp. More details, results, and experiments can also be found in the supplemental document and video[1].
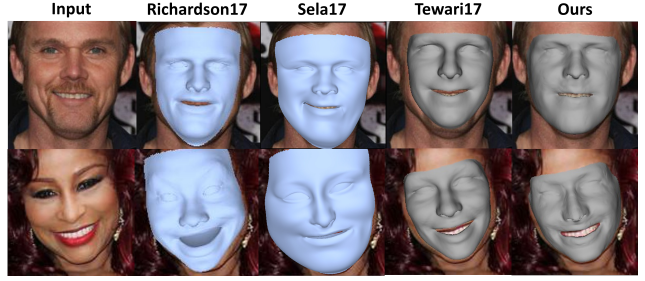
Figure 7. Comparison to [48, 52, 60]. These approaches are constrained by the (synthetic) training corpus and/or underlying 3D face model. Our optimal learned model produces more accurate results, since it is learned from a large corpus of real images.
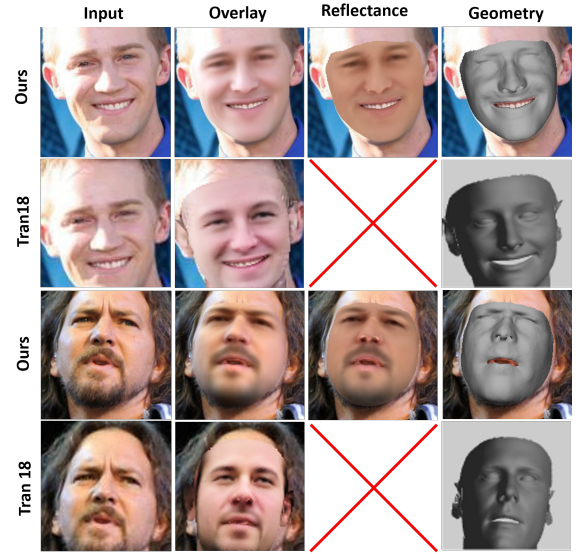

Figure 8. In contrast to Tran et al. [62], we estimate better geometry and separate reflectance from illumination. Note, the approach of Tran et al. does not disentangle reflectance and shading.
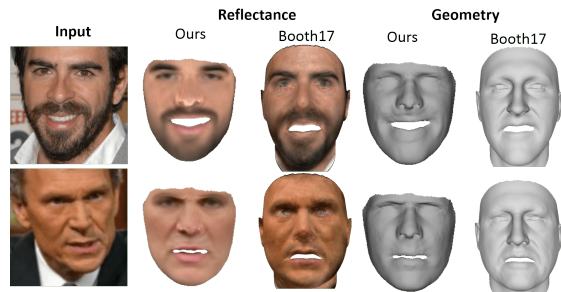

Figure 9. In contrast to the texture model of Booth et al. [12] that contains shading, our approach estimates a reflectance model.

4.1. Comparisons to Monocular Approaches

State-of-the-art monocular reconstruction approaches that rely on an existing face model [60] or synthetically generated data [52, 48] during training, do not generalize well to faces outside the span of the model. As such, they can not handle facial hair, makeup, and unmodeled expressions, see Fig. 7. Since we train our models on in-the-wild videos,

Table 1. Geometric reconstruction error on the BU-3DFE dataset [70]. Our approach produces higher quality results than the current state of the art. The approach of Tewari et al. [60] does not generalize to the ± 45 degree head poses contained in this dataset.

	Ours					[59] Fine	[59] Coarse	[60]
Train	M = 1	M = 2	M = 4	M = 2	M = 4			
Test	M = 1	M = 1	M = 1	M = 2	M = 2			
Mean	1.92 mm	1.82 mm	1.76 mm	1.80 mm	**1.74 mm**	1.83 mm	1.81 mm	3.22 mm
SD	0.48 mm	0.45 mm	0.44 mm	0.46 mm	**0.43 mm**	0.39 mm	0.47 mm	0.77 mm

Table 2. Geometric error on FaceWarehouse [17]. Our approach competes with [59] and [58], and outperforms [60] and [32]. Note, in contrast to these approaches, ours does not require a precomputed face model during training, but learns it from scratch. It comes close to the off-line high-quality approach of [24], while being orders of magnitude faster and not requiring feature detection.

	Ours	Others					
	Learning	Learning				Optimization	Hybrid
		[59] Fine	[59] Coarse	[60]	[32]	[24]	[58]
Mean	**1.90 mm**	1.84 mm	2.03 mm	2.19 mm	2.11 mm	**1.59 mm**	1.87 mm
SD	**0.40 mm**	0.38 mm	0.52 mm	0.54 mm	0.46 mm	**0.30 mm**	0.42 mm
Time	**5.2 ms**	**4 ms**	**4 ms**	**4 ms**	**4 ms**	120 s	110 ms

we can capture these variations and thus generalize better in such challenging cases. We also compare to the refinement based approaches of [59, 62]. Tran et al. [62] (see Fig 8) refine a 3DMM [7] based on in-the-wild data. Our approach produces better geometry without requiring a 3DMM and, contrary to [62], it also separates albedo from illumination. The approach of Tewari et al. [59] (see Fig 6) requires a 3DMM [7] as input and only learns shape and reflectance correctives. Since they learn from monocular data, their correctives are prone to artifacts, especially when occlusions or extreme head poses exist. In contrast, our approach learns a complete model from scratch based on multi-view supervision, thus improving robustness and reconstruction quality. We also compare to [12], which only learns a texture model, see Fig. 9. In contrast, our approach learns a model that separates albedo from illumination. Besides, their method needs a 3DMM [7] as initialization, while we start from a single constantly colored mesh and learn all variation modes (geometry and reflectance) from scratch.

4.2. Quantitative Results

We also evaluate our reconstructions quantitatively on a subset of the BU-3DFE dataset [70], see Tab. 1. This dataset contains images and corresponding ground truth geometry of multiple people performing a variety of expressions. It includes two different viewpoints. We evaluate the importance of multi-frame training in the case of monocular reconstruction using per-vertex root mean squared error based on a pre-computed dense correspondence map. The lowest error is achieved with multi-view supervision during training, in comparison to monocular input data. Multi-view supervision can better resolve depth ambiguity and thus learn a more accurate model. In addition, the multi-view supervision also leads to a better disentanglement of reflectance and shading. We also evaluate the advantage of multi-frame input at test time. When both images corresponding to a

shape are given, we consistently obtain better results. Further, our estimates are better than the state-of-the-art approach of [59]. Since [59] refine an existing 3DMM only using monocular images during training, it cannot resolve depth ambiguity well. Thus, it does not improve the performance compared to their coarse model on the ± 45 degree poses of BU-3DFE [70]. Similar to previous work, we also evaluate monocular reconstruction on 180 meshes of FaceWarehouse [17], see Tab. 2. We perform similar to the 3DMM-based state-of-the-art. Note that we do not use a precomputed 3DMM, but learn a model from scratch during training, unlike all other approaches in this comparison. For this test, we employ a model learned starting from an asian mean face, as FaceWarehouse mainly contains asians. Our approach is agnostic to the mean face chosen and thus allows us this freedom.

5. Conclusion & Discussion

We have proposed a self-supervised approach for joint multi-frame learning of a face model and a 3D face reconstruction network. Our model is learned from scratch based on a large corpus of in-the-wild video clips without available ground truth. Although we have demonstrated compelling results by learning from in-the-wild data, such data is often of low resolution, noisy, or blurred, which imposes a bound on the achievable quality. Nevertheless, our approach already matches or outperforms the state-of-the-art in learning-based face reconstruction. We hope that it will inspire follow-up work and that multi-view supervision for learning 3D face reconstruction will receive more attention.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 7

[2] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1558–1565. IEEE Computer Society, 2014. 2

[3] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. The Digital Emily Project: photoreal facial modeling and animation. In *ACM SIGGRAPH Courses*, pages 12:1–12:15. ACM, 2009. 4

[4] A. Bas and W. A. P. Smith. Statistical transformer networks: learning shape and appearance models via self supervision. arXiv:1804.02541, 2018. 2

[5] F. Bernard, P. Gemmar, F. Hertel, J. Goncalves, and J. Thunberg. Linear shape deformation models with local support using graph-based structured matrix factorisation. In *CVPR*, 2016. 2

[6] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Computer graphics forum*, pages 641–650. Wiley Online Library, 2003. 1, 2

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2, 4, 8

[8] F. Bogo, J. Romero, M. Loper, and M. J. Black. Faust: Dataset and evaluation for 3d mesh registration. In *CVPR '14*, pages 3794–3801. IEEE Computer Society, 2014. 2

[9] T. Bolkart and S. Wuhrer. A groupwise multilinear correspondence optimization for 3d faces. In *ICCV*, pages 3604–3612. IEEE Computer Society, 2015. 2

[10] T. Bolkart and S. Wuhrer. A robust multilinear model learning framework for 3d faces. In *CVPR*, pages 4911–4919. IEEE Computer Society, 2016. 2

[11] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2014)*, 33(6), 2014. 6

[12] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild". In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 3, 7, 8

[13] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 2

[14] M. Botsch and O. Sorkine. On linear variational surface deformation methods. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):213–230, Jan. 2008. 6

[15] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for realtime facial animation. *ACM Trans. Graph.*, 32(4):40:1–40:10, 2013. 2

[16] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, 2014. 4

[17] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, Mar. 2014. 8

[18] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph.*, 35(4):126:1–126:12, 2016. 2

[19] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 1, 3

[20] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 1

[21] G. Fyffe, A. Jones, O. Alexander, R. Ichikari, and P. Debevec. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph.*, 34(1):8:1–8:14, 2014. 2

[22] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 1272–1279. IEEE Computer Society, 2013. 3

[23] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, volume 32, pages 158:1–158:10, November 2013. 2

[24] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics*, 35(3):28:1–15, June 2016. 2, 3, 8

[25] P. Garrido, M. Zollhöfer, C. Wu, D. Bradley, P. Pérez, T. Beeler, and C. Theobalt. Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph.*, 35(6):219:1–219:11, 2016. 2

[26] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. Seidel. A statistical model of human pose and body shape. *Comput. Graph. Forum*, 28(2):337–346, 2009. 2

[27] P. Hsieh, C. Ma, J. Yu, and H. Li. Unconstrained realtime facial performance capture. In *CVPR*, pages 1675–1683. IEEE Computer Society, 2015. 2

[28] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph.*, 34(4):45:1–45:14, 2015. 2, 3

[29] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, volume 11219 of *Lecture Notes in Computer Science*, pages 386–402. Springer, 2018. 2

[30] I. Kemelmacher-Shlizerman. Internet-based morphable model. *ICCV*, pages 3256–3263, 2013. 1, 3

[31] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, pages 1746–1753, 2011. 3

[32] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. InverseFaceNet: Deep Single-Shot Inverse Face Rendering From A Single Image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 8

[33] M. Klaudiny, S. McDonagh, D. Bradley, T. Beeler, and K. Mitchell. Real-Time Multi-View Facial Capture with Synthetic Training. *Comput. Graph. Forum*, 2017. 2

[34] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *SCA*, pages 10:1–10:10. ACM, 2017. 2

[35] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42:1–42:10, 2013. 2

[36] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017. 2

[37] S. Liang, L. G. Shapiro, and I. Kemelmacher-Shlizerman. Head reconstruction from internet photos. In *European Conference on Computer Vision*, pages 360–374. Springer, 2016. 1, 3

[38] M. Lüthi, T. Gerig, C. Jud, and T. Vetter. Gaussian process morphable models. *PAMI*, 40(8):1860–1873, 2018. 2

[39] S. McDonagh, M. Klaudiny, D. Bradley, T. Beeler, I. Matthews, and K. Mitchell. Synthetic prior design for real-time face tracking. *3DV*, 00:639–648, 2016. 2

[40] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt. Live intrinsic video. *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, 35(4), 2016. 6

[41] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and C. Theobalt. Sparse localized deformation components. *ACM Trans. Graph.*, 32(6):179:1–179:10, 2013. 2

[42] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2016)*, 35(6), 2016. 2

[43] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *CVPR*, pages 3418–3427. IEEE Computer Society, 2016. 3

[44] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286, 2017. 2

[45] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Proc. SIGGRAPH*, pages 117–128. ACM, 2001. 5

[46] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3d faces using convolutional mesh autoencoders. In *ECCV '18*, volume 11207 of *Lecture Notes in Computer Science*, pages 725–741. Springer, 2018. 2

[47] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016. 2

[48] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 7, 8

[49] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2127–2141, 2017. 2, 3

[50] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. 91(2):200–215, 2011. 3, 6

[51] J. M. Saragih, S. Lucey, and J. F. Cohn. Real-time avatar animation from a single image. In *FG*, pages 213–220. IEEE, 2011. 3, 6

[52] M. Sela, E. Richardson, and R. Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *ICCV*, 2017. 2, 7, 8

[53] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 2

[54] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.*, 33(6):222:1–222:13, 2014. 2, 3

[55] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Conference on*, pages –. IEEE, 2017. 1

[56] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. In *ACM Transactions on Graphics (TOG)*, volume 26, page 80. ACM, 2007. 4

[57] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, pages 796–812, 2014. 1, 3

[58] A. Tewari, M. Zollhöfer, F. Bernard, P. Garrido, H. Kim, P. Pérez, and C. Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *PAMI*, 2018. 8

[59] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 6, 7, 8

[60] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017. 2, 7, 8

[61] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183:1–183:14, 2015. 3

[62] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 2018. 1, 3, 7, 8

[63] L. Tran and X. Liu. On learning 3d face morphable model from in-the-wild images. arXiv:1808.09560, 2018. 1, 2, 3

[64] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[65] S. Tulsiani, A. Kar, J. Carreira, and J. Malik. Learning category-specific deformable 3d models for object reconstruction. *PAMI*, 39(4):719–731, 2017. 2

[66] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 209–217. IEEE Computer Society, 2017. 3

[67] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016. 2

[68] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, July 2005. 1, 2

[69] C. Wu, D. Bradley, P. Garrido, M. Zollhöfer, C. Theobalt, M. Gross, and T. Beeler. Model-based teeth reconstruction. *ACM Trans. Graph.*, 35(6):220:1–220:13, 2016. 2

[70] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, 2006. 8

[71] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Comput. Graph. Forum (Eurographics State of the Art Reports 2018)*, 37(2), 2018. 1, 2