Learning Complete 3D Morphable Face Models from Images and Videos

Mallikarjun B R Ayush Tewari Hans-Peter Seidel Mohamed Elgharib Christian Theobalt
Max Planck Institute for Informatics, Saarland Informatics Campus

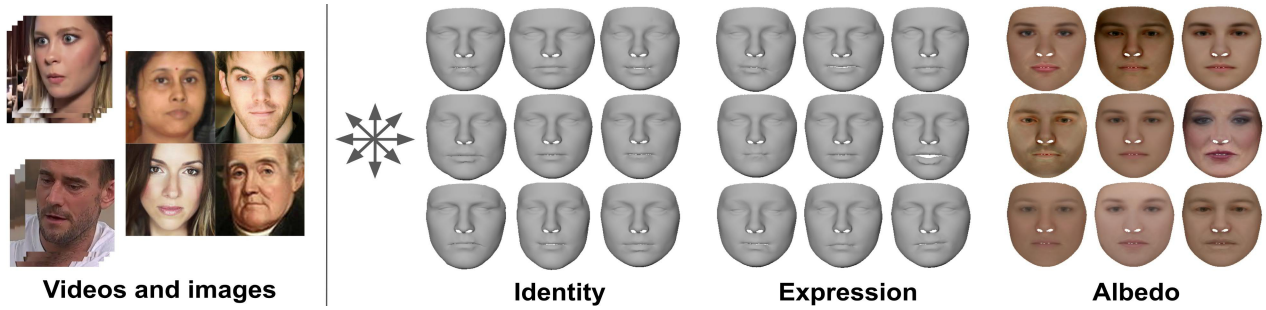

Figure 1. We present a method for learning complete 3D morphable models of faces from videos and images. We show visualizations of the learned models on the right. Faces in each direction of indicated arrows is obtained by linearly scaling individual component of respective models. Identity geometry captures variations in the face shape (second column), lips (top left to bottom right) and jaw (top right to bottom left), while expressions capture variations due to mouth opening (second row), smile (second column) and eye movement (top right to bottom left). Albedo/Reflectance spans a variety of skin color (second column), eye color (top right to bottom left) and gender specific features such as facial hair and make-up (second row).

Abstract

Most 3D face reconstruction methods rely on 3D morphable models, which disentangle the space of facial deformations into identity and expression geometry, and skin reflectance. These models are typically learned from a limited number of 3D scans and thus do not generalize well across different identities and expressions. We present the first approach to learn complete 3D models of face identity and expression geometry, and reflectance, just from images and videos. The virtually endless collection of such data, in combination with our self-supervised learning-based approach allows for learning face models that generalize beyond the span of existing approaches. Our network design and loss functions ensure a disentangled parameterization of not only identity and albedo, but also, for the first time, an expression basis. Our method also allows for in-the-wild monocular reconstruction at test time. We show that our learned models better generalize and lead to higher quality image-based reconstructions than existing approaches. We show that the learned model can also be personalized to a video, for a better capture of the geometry and albedo.

1. Introduction

Monocular 3D face reconstruction is defined as recovering the dense 3D facial geometry and skin reflectance of a face from a monocular image. It has applications in several domains such as VR/AR, entertainment, medicine, and human computer interaction [65, 16]. We are concerned with in-the-wild images which can include faces of many different identities with varied expressions and poses, in unconstrained environments with widely different illumniation. This problem has been well-studied, where a lot of success can be owed to the emergence of *3D Morphable Models* [5]. These models define the space of deformations for faces as separate disentangled models such as facial identity, expression and reflectance. They are widely used in the literature to limit the search space for reconstruction [65, 16]. However, these models are often learned from a limited number of 3D scans, which constrains their generalizability to subjects and expressions outside the space of the scans.

Recent efforts have proposed to learn face models with better generalizability from internet images or videos [55, 56, 58, 59, 60]. However, learning from in-the-wild data is highly challenging, requiring solutions for handling the strong inherent ambiguities and for ensuring disentanglement between different components of the reconstruction.

Some approaches deal with a slightly easier problem of refining an initial morphable model pretrained on 3D data on in-the-wild imagery [56, 59, 61, 60, 58]. Our objective is to learn face models without using any pretrained models to start with. The closest approach to ours is Tewari *et al*. [55], which learns only the models of facial identity geometry and reflectance from in-the-wild videos. However, they still use a pretrained expression model to help disentangle the identity and expression variations in geometry. We present the first approach that learns the the complete face model of identity geometry, albedo and expression just from in-the-wild videos. We start just from a template face mesh without using any priors about deformations of the face, other than smoothness. This also makes ours the first approach to learn face expression models from 2D data.

We achieve this through several technical contributions. We design a neural network architecture which, in combination with specially tailored self-supervised loss functions, enables (1) learning of face identity, expression and skin reflectance models, as well as (2) joint 3D reconstruction of faces from monocular images at state-of-the-art accuracy. We use a siamese network architecture which can process multiple frames of video during training, enabling consistent identity reconstructions along with per-frame expressions and scene parameters. We use a differentiable renderer to render synthetic images of the network's reconstructions. To compare reconstructions to the input, we use a new combination of appearance-based and face segmentation losses that permit learning of the face geometry and appearance, as well as a high-quality expression basis of detailed mouth and lip motion. Our novel lip segmentation consistency loss aligns the lip region in 3D with 2D segmentations. Our loss is robust to noisy outliers, leading to qualitatively better lip segmentations than the ground truth used. We also introduce a disentanglement loss which ensures that the expression component of a reconstructed mesh is small when the input image contains a neutral face. We show that the combination of these innovations is crucial to learn a full face model with proper component disentanglement from in-the-wild imagery. Our monocular reconstruction outperforms the state-of-the-art image-based face reconstruction methods.

In summary we make the following contributions: 1) the first approach for learning all components - identity, albedo and expression bases - of a morphable face model, trained on in-the-wild 2D data, 2) the first approach to learn 3D expression models of faces in a self-supervised manner, 3) a lip segmentation consistency loss to enforce accurate mouth modeling and reconstruction, 4) enforcing disentanglement of identity and expression geometry by utilizing a dataset of neutral images.

2. Related Work

2.1. Face Modeling

Faces are typically modeled as a combination of several components. 3D parametric identity [5, 3] and blendshape [40, 31, 54] models are used to represent the identity (geometry and reflectance) and facial expressions. This generalizes active appearance models [13] from 2D to 3D space. PCA is commonly used to independently learn the different models from a dataset of 3D scans [5, 3, 34, 7]. Multi-linear face models extend this concept by using tensor-based representations to better model the correlations between the identity and expression components [14, 6, 17]. Recent efforts have focused on learning models from large scale 3D data [8, 34, 30, 33]. Physics-based face models [24, 53] have also been proposed, however their complexity makes their use in real-time rendering or efficient reconstruction difficult. Animation artists can also manually create face rigs, with custom-designed control parameters. They often use blendshapes, linear combinations of designed base expressions, to control face expressions [31].

2.2. Face Reconstruction

Image-based reconstruction methods [65] estimate the face reflectance and geometry from images and videos. 3DMMs [5, 3] are often used as priors for this task. Methods differ in the type of inputs they use, such as monocular [44], multi-frame [55] or unstructured photo collection [45]. Current methods can be classified into 1) optimization-based and 2) learning-based. Optimization-based techniques rely on a personalized model [10, 18, 19, 63, 23] or a general parametric prior [1, 9, 32, 52, 48, 48] to estimate 3D geometry, often combined with texture and illumination, from a 2D video or image. Learning-based approaches regress the 3D reconstruction from a single image by learning an image-to-parameter or image-to-geometry mapping [38, 43, 57, 56, 49, 62, 26]. Most methods require ground truth face geometry [62, 28], are trained on synthetic data [42, 43, 49, 26], or a mixture of both [37, 27, 61].

Tewari *et al*. [57] proposed a differentiable rendering-based loss which allows for self-supervised training from 2D images. Other approaches have proposed using a facial recognition network and perceptual losses for higher quality reconstructions [21, 15]. Using multiple images of a person during traing has shown to be effective for high-quality reconstruction in challenging conditions [46, 51]. While these techniques are fast and produce good results, reconstructions are limited to the pre-dedfined 3DMM space.

2.3. Joint Modeling and Reconstruction

Recent learning-based methods for monocular face reconstruction [56, 60, 59, 7, 50, 55] allow for capturing variations outside of the 3DMM space, by learning models from

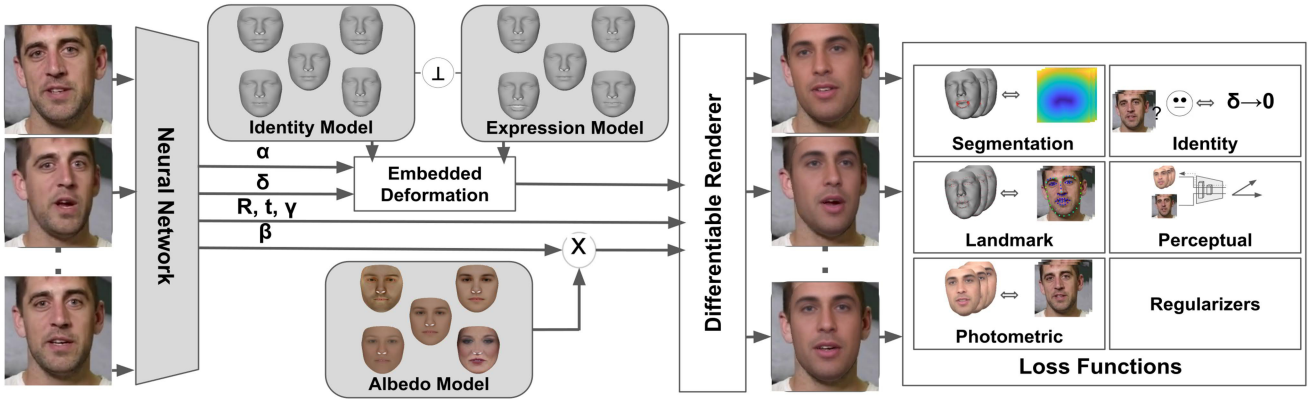

Figure 2. Our approach jointly learns identity, expression and albedo models along with the input-dependent parameters for these models. The network is trained in a siamese manner using differentiable renderer to compute self-supervised loss.

in-the-wild 2D data. Most approaches either initialize the learned model with an existing 3DMM [60, 58, 36], or learn a corrective space which acts in addition to a fixed 3DMM [56, 11]. Learning morphable models from scratch is a relatively less studied problem. Tewari *et al*. [55] learned the identity (shape and reflecance) model from community videos in a self-supervised manner. The learning starts from a neutral reflectance and coarse deformation graph, which are refined during training. It however relies on a learned expression model. Our method is the first to learn all dimensions–reflectance, identity geometry, and expression from in-the-wild data.

3. Method

We present the first method to learn a deformable face model that jointly learns all three of the following dimensions - identity geometry, expression and reflectance - from unlabelled community videos, without using a pre-defined 3DMM to start with. The starting point for our deformation models is a mesh which defines the topology of reconstructions, as well as the initial geometry and reflectance values for our networks. We design a multi-frame siamese network which processes the videos at training time. The training is self-supervised, without any 3D supervision. We use a differentiable renderer to define our loss functions in the image space. Our network design, in addition to the loss functions enable disentangled learning of the face model subspaces. Our network also jointly learns to predict parameters of the models, thus enabling 3D reconstruction at test time, even from monocular images.

3.1. Model Representation

We learn linear face models, similar to many existing face models [5, 56, 55]. (Stacked) Mesh vertex positions and reflectances are represented as V and R, $|V| = |R| = 3N$, where N is the number of vertices. We use the mesh topology of Tewari *et al*. [56] with $N = 60,000$ vertices.

Geometry Models 3D face deformations due to identity and expression can be represented using linear geometry models.

$$V(\mathbf{M}_{id}, \mathbf{M}_{exp}, \alpha, \delta) = \bar{V} + \mathbf{M}_{id}\alpha + \mathbf{M}_{exp}\delta\,. \quad (1)$$

Here, $\mathbf{M}_{id} \in \mathbb{R}^{3N \times m_i}$ and $\mathbf{M}_{exp} \in \mathbb{R}^{3N \times m_e}$ are the learnable linear identity and expression models. We use the mean face from [4] as $\bar{V}$. $\alpha \in \mathbb{R}^{m_i}$ and $\delta \in \mathbb{R}^{m_e}$ are the identity and expression parameters for the corresponding models.

We use a low-dimensional embedded deformation graph to represent the linear models $\mathbf{M}_{id}$ and $\mathbf{M}_{exp}$,

$$\mathbf{M}_{id} = \mathbf{U}\mathbf{M}_{gid}, \mathbf{M}_{exp} = \mathbf{U}\mathbf{M}_{gexp}\,. \quad (2)$$

Here, $\mathbf{M}_{gid} \in \mathbb{R}^{3G \times m_i}$ and $\mathbf{M}_{gexp} \in \mathbb{R}^{3G \times m_e}$ are linear models defined on a lower dimensional graph with $G = 521$ nodes. The fixed upsampling matrix $\mathbf{U} \in \mathbb{R}^{3N \times 3G}$ couples the deformation graph to the full face mesh and is precomputed before training. Learning the shape models in the graph-space reduces the number of learnable parameters in the model, and makes it easier to formulate smoothness constraints over the reconstructions.

Reflectance Model We employ a linear model of diffuse face reflectance.

$$R(\mathbf{M}_R, \beta) = \bar{R} + \mathbf{M}_R\beta \quad (3)$$

Here, $\mathbf{M}_R \in \mathbb{R}^{3N \times m_r}$ is the learnable reflectance model, and $\beta \in \mathbb{R}^{m_r}$ are the estimated parameters. We use the mean face reflectance from [4] as $\bar{R}$. Unlike geometry, we learn a per-vertex reflectance model on the full mesh resolution. This allows us to preserve photorealistic details of the face in the reconstructions.

3.2. Image Formation

Given a face mesh with positions V and reflectance values R, we additionally need the extrinsic camera parameters in order to render synthetic images. Rigid face pose is

represented as $\phi(v) = \mathbf{R}v + t$, where t includes 3 translation parameters, and rotation $\mathbf{R} \in SO(3)$ is represented with 3 Euler angles. We use a perspective camera model, with projection function $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. For any point $v \in \mathbb{R}^3$, the corresponding projection $p(v) \in \mathbb{R}^2$ is defined as $p(v) = \pi(\phi(v))$.

To define the color, we need to model the scene illumination. We assume a lambertian surface, and use spherical harmonics (SH) coefficients γ to represent the illumination [41]. The color c of a point with reflectance r and position v can be computed as

$$c = r \cdot \sum_{b=1}^{B^2} \gamma_b \cdot \mathbf{H}_b(n) \tag{4}$$

$\mathbf{H}_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ are the SH basis functions, $\gamma \in \mathbb{R}^{B^2}$ are the SH coefficients, n are the normals at point v and $B = 3$.

Differentiable Rendering We implement a differentiable rasterizer to render 2D images from 3D face meshes. For each pixel, we first compute the 3D face points which project into the pixel. We use a z-buffering algorithm to select the visible triangles. Pixel color is computed by linearly interpolating between vertex colors using barycentric coordinates. We implement the renderer in a data-parallel fashion as a custom TensorFlow layer.

This implementation also allows for gradients to back propagate through the rendering step. The gradients computed at any pixel location can be distributed across the vertices of the relevant triangle according to the barycentric coordinates. While such an implementation cannot differentiate through the visibility check, it works well in practice.

3.3. Network Architecture

Our network consists of siamese towers which take as input different frames of a video $F_i, \forall i \in \{0..N_f - 1\}$, where N_f is the number of frames. Each such set of N_f frames of one person identity is called a multi-frame image. The output of the siamese towers are the face parameters which are independent per-frame, i.e. expressions (δ_i), illumination (γ_i) and rigid pose (ϕ_i). We formulate multi-frame constraints for the identity component of the model. By design, the network only produces one output per multi-frame input for the identity shape (α) and reflectance (β) parameters. This is done through a multi-frame pooling of features from the siamese towers, followed by a small network. Thus, the network produces per-frame parameters, $\mathbf{p}_i = (\alpha, \beta, \delta_i, \gamma_i, \phi_i)$

In addition to the face parameters, we also learn the face models for expression ($\mathbf{M}_{exp}$), identity shape ($\mathbf{M}_{id}$) and reflectance ($\mathbf{M}_R$). These models are implemented as weights of the learnable network. More specifically, the position and reflectance of the face mesh, represented as $V_i(\mathbf{M}_{id}, \mathbf{M}_{exp}, \alpha_i, \delta_i)$ and $R(\mathbf{M}_R, \beta)$ are computed by applying the learnable models to the predicted parameters as explained in Eqs. 1 and 3. The computed reconstructions are then rendered using the differentiable renderer to produce synthetic images $S_i \in \mathbb{R}^{240X240X3}$. We enforce orthogonality between the geometry and expression models such that $\mathbf{M}_{id}\mathbf{M}_{exp} = 0$. This is done by dynamically constructing $\mathbf{M}_{id}$ in a forward pass by projecting itself onto the orthogonal complement of $\mathbf{M}_{exp}$ [55]. Please see Fig. 2 for a visualization of the architecture.

3.4. Dataset

We use two datasets to train our approach: *VoxCeleb* [12] and *EmotioNet* [2]. VoxCeleb consists of over 140k videos covering 6000 different identities crawled from YouTube. We sample N_f = 4 frames per video clip for training. This gives us a variety of head pose, expressions and illumination per identity. All our images are cropped around the face, and we discard images containing less than 200 pixels. We resize the crops to 240x240 pixels.

EmotionNet is a large-scale image dataset of in-the-wild faces, covering a wide variety of expressions, automatically annotated with Action Units (AU) intensities. We use a subset of 7,000 images of neutral faces by selecting images with no active AU. We use these neutral images to enforce model disentanglement between the identity and expression geometry components (Sec. 3.5.1).

3.5. Loss Functions

We perform self-supervised training, without using any 3D supervision. Let $\mathbf{x}$ be the learnable variables in the network, which includes all trainable weights in the neural network, as well as the learnable face models $\mathbf{M}_{id}$, $\mathbf{M}_{exp}$ and $\mathbf{M}_R$. All the estimated parameters $\mathbf{p}_i$ can be parametrized using these learnable variables. Our loss function consists of:

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = \mathcal{L}_{land}(\mathbf{x}) + \lambda_{seg} \cdot \mathcal{L}_{seg}(\mathbf{x}) + \\ \lambda_{pho} \cdot \mathcal{L}_{pho}(\mathbf{x}) + \lambda_{per} \cdot \mathcal{L}_{per}(\mathbf{x}) + \\ \lambda_{smo} \cdot \mathcal{L}_{smo}(\mathbf{x}) + \lambda_{dis} \cdot \mathcal{L}_{dis}(\mathbf{x}) \ ,\end{aligned} \tag{5}$$

The last two terms are regularizers and the first four are data terms. We used fixed $\lambda_\bullet$ values to weigh the losses.

Landmark Consistency For each frame F_i, we automatically annotate 66 sparse 2D keypoints [47] $l_i \in \mathbb{R}^2, i \in \{0..65\}$. We compare these 2D landmarks with sparse vertices of the reconstruction which corresponds to these landmarks.

$$\mathcal{L}_{land}(\mathbf{x}) = \sum_{i=0}^{N_f-1} \sum_{k=0}^{65} ||l_k - p(v_k(\mathbf{x}))||^2 \ . \tag{6}$$

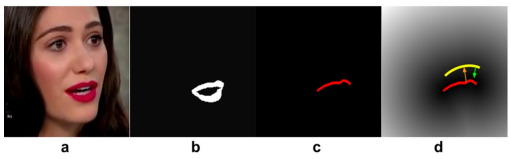

Figure 3. For a given image [a], we obtain the segmentation masks [b], its boundary [c] and distance transform (DT) image [d] of [c]. We employ a segmentation loss which tries to move the vertices on the projected mesh contour (yellow) to a lower energy position in DT. In addition, each pixel in the boundary (red) attracts the nearest vertex on the mesh contour.

Here, $v_k(\mathbf{x}) \in \mathbb{R}^3$ indicates the position of the kth landmark vertex, and $p(v_k(\mathbf{x}))$ is its 2D projection (Sec. 3.2). While most face landmarks can be manually annotated on the template mesh, the face contour is not fixed and thus has to be calculated dynamically (see supplemental for details).

Segmentation Consistency The estimated keypoints are ambiguous in the inner lip regions, due to rolling lip contours. In addition, the accuracy of sparse keypoint prediction is inadequate to learn expressive expression models. We use a dense contour loss for the lip region, guided by automatic segmentation mask prediction [29]. The lip segmentation contours are converted into distance transform images $\mathbf{D}_a^b$, where $a \in \{upper, lower\}$ and $b \in \{outer, inner\}$ corresponding to the outer and inner contours of both lips. We also compute the contours of both lips projected by the predicted reconstruction, where each element of set $\mathbf{C}_a^b(\mathbf{x})$ stores a 2D pixel location on the contour. For a given distance transform image and the corresponding contour of the predicted mesh, the loss function minimizes the distance between the mesh contours and segmentation contours, see Fig. 3.

$$\mathcal{L}_{seg}(\mathbf{x}) = \sum_{i=0}^{N_f-1} \sum_{\forall(a,b)} \Big(\sum_{\forall(x,y)\in\mathbf{C}_a^b(\mathbf{x})} \mathbf{D}_a^b(x,y) + \sum_{\{(x,y)\,|\,\mathbf{D}_a^b(x,y)=0\}} ||(x,y) - closest(\mathbf{C}_a^b(\mathbf{x}),(x,y))||^2 \Big). \tag{7}$$

Here, the first term minimizes the distance from every pixel in the mesh contour to the image contour. The second term is a symmetric term minimizing the distance between every pixel in the image contour to the closest mesh contour. $closest(\mathbf{C}_a^b(\mathbf{x}),(x,y))$ is a function which gives the position of the closest pixel in $\mathbf{C}_a^b(\mathbf{x})$ to (x,y). We use our differentiable renderer to compute the rolling inner contours on the mesh. The outer contours are computed as the projection of some manually annotated vertices on the template mesh. In practice, we ignore this loss term at pixels where the distance between the image and mesh contours is greater than a threshold. This helps in training with noisy segmentation labels.

Photometric Consistency We evaluate the dense photometric consistency between the reconstructions and the input. For each pixel, we minimize the color difference between the input images F_i and the rendered images $S_i(\mathbf{x})$.

$$\mathcal{L}_{pho}(\mathbf{x}) = \sum_{i=0}^{N_f-1} ||M_i \odot (F_i - S_i(\mathbf{x}))||^2 \ . \tag{8}$$

M_i is a mask computed using the renderer, and $\odot$ is an element-wise multiplication operator.

Perceptual Loss We additionally employ a dense perceptual loss to help our networks learn higher quality models, including high-frequency reflectance details. In particular, we use a VGG network pretrained on ImageNet [25] to get the intermediate features for both input frames and the output synthetic frames. We then minimize the cosine distance between these features.

$$\mathcal{L}_{per}(\mathbf{x}) = \sum_{i=0}^{N_f-1} \sum_{l=0}^{4} 1 - \frac{< f_l(S_i(\mathbf{x})), f_l(F_i) >}{||f_l(S_i(\mathbf{x}))|| \cdot ||f_l(F_i)||}, \tag{9}$$

where $f_l(\cdot)$ denotes the output of the lth intermediate layer for input x and $< \cdot, \cdot >$ denotes the inner product.

Geometry Smoothness To ensure smoothness of the final geometry, we use a smoothness loss at the graph level. Let $G_i(\mathbf{x}) \in \mathbb{R}^{N_g \times 3}$ with $N_g = 521$ nodes denote the geometry reconstruction for frame F_i at the graph level. We employ an ℓ_2 loss to constrain the difference between the deformation of adjacent nodes.

$$\mathcal{L}_{smo}(\mathbf{x}) = \sum_{i=0}^{N_f-1} \sum_{g\in G_i(\mathbf{x})} \sum_{n\in\mathcal{N}(g)} ||g-n||^2 \ , \tag{10}$$

where $\mathcal{N}(g)$ is the neighbourhood of node g.

3.5.1 Model Disentanglement

Our goal is to learn deformation models for facial geometry, expression and reflectance. Disentangling these deformations in the absence of an initial 3DMM is challenging. We use a combination of network design choices and loss functions to enable simultaneous learning of these models.

Siamese Networks: Our siamese network design ensures that the identity components of our reconstructions are consistent across all frames of the batch. Such a network architecture allows us to disentangle illumination from reflectance in addition to helping with the disentanglement of expressions from identity geometry.

Disentanglement Loss: Our method can still lead to some failure modes. For example, $\mathbf{M}_{id}$ can collapse to a zero matrix, and all geometric deformations including those due

Figure 4. Our approach reconstructs all facial components with high fidelity and good disentanglement.

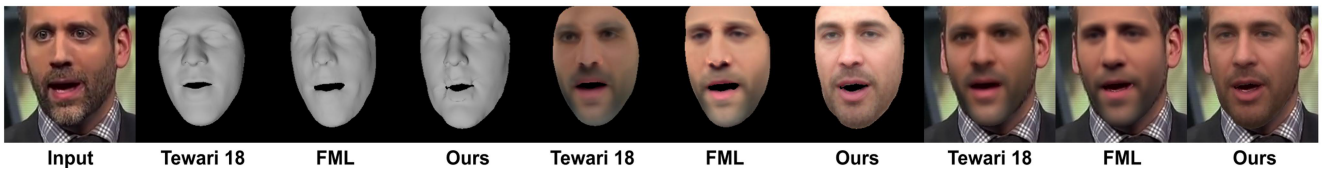

Figure 5. Our approach produces better geometry, including detailed mouth shapes compared to Tewari *et al*. [56] and FML [55]. Our albedo is also more detailed and better disentangled from the illumination component. Results are visualized in the order of geometry, albedo, and full reconstruction for all methods.

to identity can be learned by the expression model without any penalty from any loss function. To prevent such failure modes, we design a loss function to disentangle these components. As mentioned in Sec. 3.4, a subset of our dataset includes images of faces with neutral expression. For these images, we employ a loss function which minimizes the deformations due to expressions.

$$\mathcal{L}_{dis}(\mathbf{x}) = \sum_{i=0}^{N_f-1} ||\delta_i(\mathbf{x})||^2 \ . \tag{11}$$

Since we do not have videos for these images, we simply duplicate the same image as input to the siamese towers. Finally, our training strategy further helps with disentanglement. Please refer to the supplemental for details.

3.6. Personalized Model

While the expression model we have described is generic, describing the deformations for any identity, we can also personalize the model by finetuning it on a video at test time. We can also update the identity geometry and reflectance models for higher quality reconstructions. The loss function for finetuning is the same as the training loss. The rest of the network is kept fixed, such that the parameter estimation is not affected. We show that this leads to high quality reconstructions, without changing the semantics of the models.

4. Results

Training Details We implement our approach in *Tensorflow* and train it over three stages: 1) pose pretraining 2) identity pretraining and 3) combined training. We empirically found this curriculum learning to help with stable training and disentanglement of the identity and expression models. *Pose Pretraining:* We first train only for the rigid head pose. All other parameters are kept fixed to their initial value. *Identity pretraining:* Next, we train for the identity model. This step is only trained on the EmotionNet data with neutral expressions. We enforce the expression parameters to be zero, enforcing all deformations to be induced by the identity model. *Combined Training:* Last, we train for the complete model with the loss functions as explained in (5). Similar to the first stage, we continue to impose the landmark loss term on the mean mesh throughout model learning. This helps in avoiding the geometric models learning the head pose. Our training data now consists of mini-batches sampled from EmotionNet and VoxCeleb with 1:3 ratio. We train for 650k iterations with a batch size of 1. This results in a training time of 117 hours on a TitanV. We use 80 basis vectors for identity geometry and albedo, and 64 for expression.

4.1. Qualitative Evaluation

Fig. 1 visualizes the different modes of the learned model. Our method disentangles the various facial components of identity geometry, expressions and albedo. The identity model correctly captures a variety of face shapes, mouth and eye structure. The expression model captures a variety of deformations produced by the mouth and eyes, while the reflectance captures different skin color, and gender specific features such as facial hair and make-up. Fig. 4 shows all components of our reconstruction for several images. Our approach can handle different ethnicities, genders and scene conditions, and produces high-quality reconstruction, both in geometry and reflectance.

Comparisons: Figs. 5 - 8 compare our approach to several state-of-the-art face reconstruction techniques. Tran *et al*. [60, 58] learn a combined geometry model for identity and expressions, while we learn separate models (Fig. 6). Like other 3DMM based approaches, RingNet [46], which estimates the parameters of a pre-trained face model [35], struggles with out-of-space variations especially in the mouth region (Fig. 8). Reconstrutions of MoFA [57] and

Figure 6. Both approaches of Tran *et al*. [60, 58] do not disentangle the identity geometry from expressions. In contrast, our method estimates and disentagles all the facial components. It also produces more accurate mouth shapes. [O] refers to our method.

Figure 7. MoFA [57] and GANFIT [20] produce less accurate mouth shapes, compared to our method. GANFIT albedo reconstructions can often include artifacts, especially around the eyes.

Figure 8. Our method better captures mouth shapes and eye geometry, compared to RingNet [46]. It can also additionally estimate the appearance of the face.

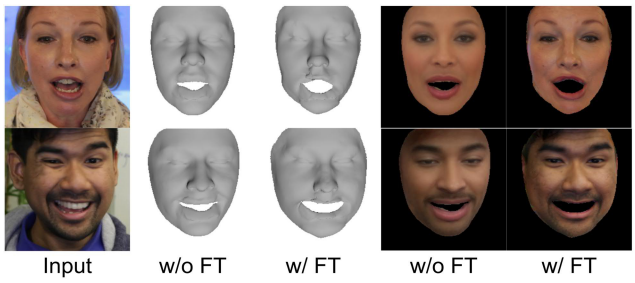

Figure 9. Finetuning (FT) allows us to better capture the personalized geometry and reflectance of the person.

GANFIT [20] are also limited by a pretrained 3DMM model and hence lead to less detailed shapes than ours (Fig. 7). While GANFIT produces detailed textures, it can often contain artifacts. Tewari *et al*. [56] refine a pretrained 3DMM model on an image dataset. We can better disentangle the reflectance and illumination components (Fig. 5). FML [55] is constrained by a pretrained expression model and thus produces lower quality shape reconstructions (Fig. 5 and 10). In addition, our reflectance estimates are more detailed compared to FML. Even though we start from just a template mesh without any deformation priors, we can produce high-quality results, better than the state of the art.

Our method can produce better lip segmentation than the approach used for generating the training data [29] in some cases. This is due to our segmentation loss function, $\mathcal{L}_{seg}$, where we selectively ignore unreliable segmentation estimates. Hence our final model is learned from only accurate segmentations in the training-set. We also provide ablative study of perceptual loss which helps in photorealism of the albedo and the final overlay in supplemental.

4.2. Quantitative Evaluation

Geometric Error: To evaluate the geometric accuracy of our 3D reconstructions, we compute the per-vertex root mean square error between the ground-truth geometry and the geometry estimated using different techniques. The GT and reconstructed meshes are first aligned such that they have the same scale, translation and orientation. We use the BU3DFE dataset [64] for evaluation, where the ground-truth geometry is obtained using 3D scans. The correspondences between the GT and reconstructed meshes are precomputed using non-rigid registration. Tab. 1 reports the results over 324 images. Our approach outperforms the approaches of MoFA [57], Tewari *et al*. [56] and FML [55]. Note that none of the approaches in Tab. 1 learn a complete face model from images and videos.

Segmentation Error: To specifically evaluate the quality of lip reconstructions, we use Intersection over Union (IoU) between our reconstructions and the input images over the lip regions. Since our approach learns an expression model from in-the-wild data, it can generalize better to different lip shapes and outperform FML [55] (see Tab. 2). Furthermore, Tab. 2 shows that removing the segmentation consistency term (Eq. 3.5) leads to lower quality results.

Disentanglement Error: One of our main objectives is to obtain a disentangled representation for faces. To evaluate the disentanglement between the reconstructed expression and identity geometry, we design a metric which measures the mean expression deformation for images with neutral faces. We test our approach on 1864 neutral faces mined

	Ours	MoFA	FML	Fine [56]	Coarse [56]
Mean	**1.75**	3.22	1.78	1.83	1.81
SD	**0.44**	0.77	0.45	0.39	0.47

Table 1. Geometric reconstruction error (in mm) on the BU-3DFE dataset [64]. Our technique outperforms MoFA [57], coarse and fine models of Tewari *et al*. [56] and FML [55].

	W/o $\mathcal{L}_{seg}$	With $\mathcal{L}_{seg}$	FML
UL IoU	0.49	**0.54**	0.51
LL IoU	0.52	**0.60**	0.58

Table 2. Intersection over Union (IoU) between the ground-truth and predicted masks for upper lip (UL) and lower lip (LL). Our segmentation consistency term produces better IoU and leads to noticeably better performance than FML [55].

	W/o $\mathcal{L}_{dis}$	With $\mathcal{L}_{dis}$	FML	MoFA
AE	4.0065	**0.0116**	2.0329	0.4056

Table 3. Our identity disentanglement term results in lesser leakage of the identity geometry into the expression component. It performs better than FML [55] and MoFA [57]. AE refers to the average expression deformation.

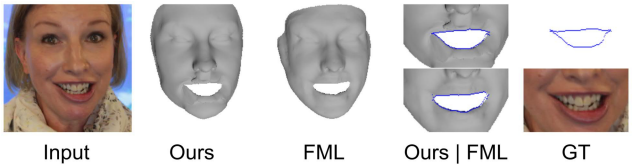

Figure 10. Our personalized model captures higher quality mouth geometry compared to FML, where only the identity models can be personalized. We show the inner contours of the meshes (ours-top, FML-bottom) in column 4. The ground truth inner contours and zoomed in image are visualized in column 5.

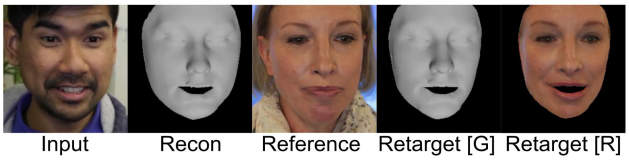

Figure 11. We can transfer expressions between personalized models. The expression parameters are transferred from the input to reference here. The personalized models preserve the semantics, which leads to correct transfer of expressions. Note the person-specific nature of the expressions. [R] refers to the full reconstruction while [G] is the geometry component.

using the same strategy described in Sec. 3.4. Tab. 3 reports the average length of the expression deformations for different approaches. Our approach achieves significantly better expression and identity disentanglement over FML [55] and MoFA [57]. This result also shows the importance of our disentanglement loss.

Verification Metric: To further evaluate disentanglement, we use the LFW dataset [22], which includes face image pairs of the same, as well as of different identities. We render the identity component of the reconstructions with the predicted pose and lighting parameters. Face embeddings are computed as the mean pooled version of the `conv5_3` output of VGG-Face [39]. We first compute the histogram distribution of cosine similarities between renderings of image pairs with the same identity in embedding space. Similarly, distribution of cosine similarities between rendering pairs of different identities is computed. The verification metric is then computed as the Earth Movers Distance (EMD) between these two distributions. Our method achieves an EMD of 0.15, compared to 0.09 for FML. A larger distance implies better representation of the differences between identities due to better disentanglement.

4.3. Personalized Model

We show results for personalizing the model by finetuning it on a video. We use one part of the video with 2000 frames for finetuning, and show qualitative improvements in the left out frames. Fig. 9 shows that the personalized model can represent the person specific mouth articulations, and can also improve the quality of reflectance. We also compare with FML [55] by finetuning its identity component. Note that the training strategy of FML does not allow for learning of the expression model. Thus, we obtain higher quality reconstructions, see Fig. 10

We show that personalizing the models does not change the semantics, by demonstrating expression transfer results in Fig. 11. Here, we take the estimated expression parameters of an input image, and the estimated identity parameters of a different reference image to obtain retargeting results using the personalized model of the reference identity. The retargeted results preserve the input expressions, which shows that the semantics of the expression model is preserved after personalization.

5. Conclusion

We presented the first approach for learning a full face model, including learned identity, reflectance and expression models from in-the-wild images and videos. Our method also learns to reconstruct faces on the basis of the learned model from monocular images. We introduced new training losses to enforce disentanglement between identity geometry and expressions, and to better capture detailed mouth shapes. Our approach outperforms existing methods, both in terms of the quality of image-based reconstruction, as well as disentanglement between the different model components. We hope that our work will inspire further research on building 3D models from 2D data.

Acknowledgements: This work was supported by the ERC Consolidator Grant 4DReply (770784). We also acknowledge support from InterDigital.

References

[1] Antonio Agudo, Lourdes Agapito, Begoña Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1558–1565, 2014.

[2] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016.

[3] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, pages 641–650, 2003.

[4] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Comput. Graph. Forum*, pages 669–676, 2004.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH's Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

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

[7] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models "in-the-wild". In *CVPR*, 2017.

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

[9] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics*, 32(4):40:1–40:10, 2013.

[10] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 35(4):126:1–126:12, 2016.

[11] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. In *IEEE European Conference on Computer Vision (ECCV)*, 2020.

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

[13] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[14] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 30(6):130:1–10, December 2011.

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, 2019.

[16] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future, 2019.

[17] V. Fernández Abrevaya, S. Wuhrer, and E. Boyer. Multilinear autoencoder for 3d face model learning. In *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, 2018.

[18] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph.*, 34(1):8:1–8:14, 2014.

[19] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian 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, 2013.

[20] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019.

[21] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, June 2018.

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

[23] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph.*, 34(4):45:1–45:14, 2015.

[24] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. Phace: Physics-based face modeling and animation. *ACM Transactions on Graphics*, 36(4):153:1–153:14, 2017.

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*, pages 694–711, 2016.

[26] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. InverseFaceNet: Deep Single-Shot Inverse Face Rendering From a Single Image. In *CVPR*, 2018.

[27] Martin Klaudiny, Steven McDonagh, Derek Bradley, Thabo Beeler, and Kenny Mitchell. Real-Time Multi-View Facial Capture with Synthetic Training. *Comput. Graph. Forum*, 2017.

[28] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *SCA*, pages 10:1–10:10. ACM, 2017.

[29] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019.

[30] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[31] J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. Practice and Theory of Blendshape Facial Models. In Sylvain Lefebvre and Michela Spagnuolo, editors, *Eurographics*, 2014.

[32] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42:1–42:10, 2013.

[33] R. Li, K. Bladin, Y. Zhao, C. Chinara, O. Ingraham, P. Xiang, X. Ren, P. Prasad, B. Kishore, J. Xing, and H. Li. Learning formation of physically-based face attributes. In *Proc. CVPR*, 2020.

[34] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Flame: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017.

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

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

[37] Steven McDonagh, Martin Klaudiny, Derek Bradley, Thabo Beeler, Iain Matthews, and Kenny Mitchell. Synthetic prior design for real-time face tracking. *3DV*, 00:639–648, 2016.

[38] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 35(6), 2016.

[39] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[40] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. Synthesizing realistic facial expressions from photographs. In *ACM Transactions on Graphics*, pages 75–84, 1998.

[41] Ravi Ramamoorthi and Pat Hanrahan. A signal processing framework for inverse rendering. In *ACM Trans. of Graph. (Proceedings of SIGGRAPH)*, pages 117–128. ACM, 2001.

[42] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016.

[43] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, July 2017.

[44] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, pages 986–993, 2005.

[45] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(11):2127–2141, 2017.

[46] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, pages 7763–7772, 2019.

[47] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Real-time avatar animation from a single image. In *Face and Gesture 2011*, pages 213–220, 2011.

[48] Evangelos Sariyanidi, Casey J Zampella, Robert T Schultz, and Birkan Tunc. Inequality-constrained and robust 3d face model fitting. 2020.

[49] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *International Conference on Computer Vision (ICCV)*, 2017.

[50] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *CVPR*, 2018.

[51] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020.

[52] Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.*, 33(6):222:1–222:13, 2014.

[53] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Transactions on Graphics*, 24(3):417–425, July 2005.

[54] J. Rafael Tena, Fernando De la Torre, and Iain Matthews. Interactive region-based linear 3d face models. *ACM Trans. Graph.*, 30(4):76:1–76:10, July 2011.

[55] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhoefer, and Christian Theobalt. FML: Face model learning from videos. In *CVPR*, 2019.

[56] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018.

[57] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, pages 3735–3744, 2017.

[58] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.

[59] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *CVPR*, pages 7346–7355, 2018.

[60] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. June 2019.

[61] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017.

[62] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard 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.

[63] Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhöfer, Christian Theobalt, Markus Gross, and Thabo Beeler. Model-based teeth reconstruction. *ACM Trans. Graph.*, 35(6):220:1–220:13, 2016.

[64] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun 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.

[65] 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.