

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

Abstract

Most 3D face reconstruction methods rely on 3D morphable models, which disentangle the space of facial deformations into identity geometry, expressions 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 geometry, albedo and expression 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.

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. 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 illumination. This problem has been well-studied, where a lot of success can be owed to the emergence of *3D Morphable Models* [5]. These morphable models define the space of deformations for faces as separate disentangled models such as facial identity, expression and reflectance. These models are widely used in the literature to limit the search space for reconstruction [61, 15]. However, these morphable 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 examined learning face models with better

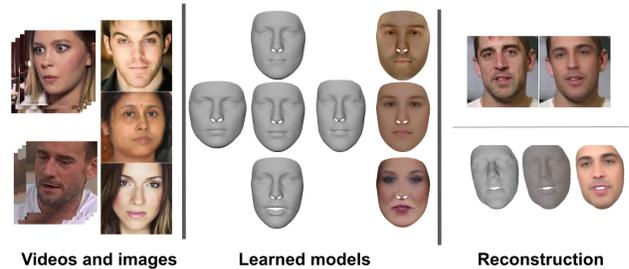


Figure 1. We present the first approach for learning complete 3D morphable face models from in-the-wild images. Our approach learns models for identity geometry, expression and albedo, completely in a self-supervised manner. It achieves good disentanglement of the various facial components and produces high quality photorealism in the final overlay.

generalizability from internet images or videos [51, 52, 54, 55, 56]. 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 [52, 55, 57, 56, 54]. Our objective is to learn face models without using any pretrained models to start with. The closest approach to ours is Tewari *et al.* [51], 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 from 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 pro-

cess multiple frames of video during training, and enables consistent identity reconstructions and expression basis reconstruction. 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 overall 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, and outperforms the state-of-the-art image-based face reconstruction.

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: reflectance, identity geometry and expression. 3D parametric identity [5, 3] and blendshapes [38, 30, 50] are used to represent identity (geometry and reflectance) and facial expressions. This generalizes active appearance models [12] from 2D to 3D space. To model the variations among different people, a parametric PCA space can be learned from a dataset of laser scans [5, 3]. This represents deformations in a low-dimensional subspace. The resulting 3D morphable face model (3DMM) is the most widely used face model in literature [61]. Multi-linear face models extend this base concept; they often use tensor-based representations to better model the correlations between shape identity and expression [13, 6, 16].

Physics-based face models [23, 49] have 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 [30]. Recently, large collections of 3D and 4D

(3D over time) scans have been used to learn face models. In [8] thousands of 4D scans are used to learn a parametric face model. Li *et al.* [32] used 33,000 3D scans to learn the FLAME face model. The model combines a linear shape space with articulated motions and semantic blendshapes.

2.2. Face Reconstruction

Image-based reconstruction methods [61] estimate face reflectance, geometry and/or expressions. 3DMMs [5, 3] is often used as priors for this task. Methods differ in the type of input they use, such as monocular [42], multi-frame [51] or unstructured photo collection input [43]. Current methods can be classified into 1) optimization-based and 2) learning-based. Optimization-based techniques rely on a personalized model [10, 17, 18, 59, 22] or a parametric prior [1, 9, 31, 48] to estimate 3D geometry, often combined with texture and/or illumination, 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 [36, 41, 53, 52, 46, 58, 25]. These methods require ground truth face geometry [58, 27], synthetic data generated from a morphable model [40, 41, 46, 25], or a mixture of both [35, 26, 57]. Tewari *et al.* [53] propose a differentiable rendering-based loss which allows unsupervised training from 2D images. Genova *et al.* [20] learn to regress an image into 3D morphable model coordinates using unlabelled data. They impose identity similarity between the input and the output, in addition to a loop-back loss and a multi-view identity loss. Deng *et al.* [14] combined image-consistency and perceptual loss which leads to improved results. A new multi-image shape confidence learning scheme is also proposed which outperforms naive aggregation and other heuristics. *RingNet* [44] estimates the parameters of the FLAME model [32]. It utilizes multiple images during training, and enforces the shape to be the same for pictures of the same identity, and different for different people. While these techniques are fast and produce good results, reconstructing shape and appearance variations outside the pre-defined 3DMM space is difficult.

2.3. Joint Modeling and Reconstruction

Recent learning-based methods for monocular face reconstruction [52, 56, 55, 7, 47, 51] allow for capturing variations outside of the 3DMM space by training from in-the-wild data. Tran *et al.* [56] employ two separate convolutional decoders to learn a non-linear model that disentangles shape from appearance. Similarly, Sengupta *et al.* [47] propose residual blocks to produce a complete separation of surface normal and albedo features. Tewari *et al.* [52] learn a corrective space of albedo and geometry which generalizes beyond 3DMMs trained on 3D data. Lin *et al.* [34] refine the initial texture generated by the 3DMM albedo model. Lee *et al.* [29] learn a non-linear identity model,

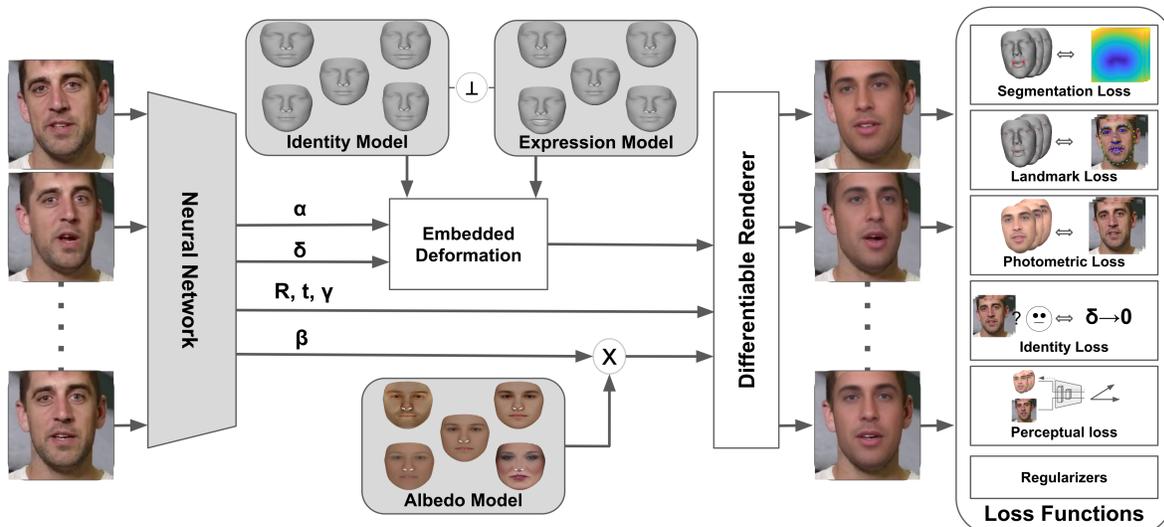


Figure 2. Our approach jointly learns identity, expression and albedo models along with the input-dependent parameters for these models. The network is trained in a siamese manner using differentiable renderer to compute self-supervised loss.

with a fixed existing expression model. Learning morphable models from scratch in a relatively less studied problem. Tewari *et al.* [51] learn the identity (shape and reflectance) 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 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 a linear face models, similar to many existing face models [5, 52, 51], comprising linear models of identity geometry, expression geometry and albedo. (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.* [52] 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 $\phi(v) = Rot(v) + t$, where t includes 3 translation parameters, and rotation $Rot \in SO(3)$ is represented in 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 [39]. The color c of a point with reflectance r and position v can be computed as

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

$\mathbf{H}_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ 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 is appropriate in practice.

3.3. Network Architecture

Our network consists of siamese towers which take as input different frames of a video $F_i \forall i \in [0, N_f]$, 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) $\forall i \in [0, N_f]$. 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). 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 albedo (\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 learned models to the predicted parameters as explained in Eqs. 1 and 3. Thus, for each multi-frame image in a mini-batch, an expression deformation per sub-frame, and consistent identity and reflectance deformations across all sub-frames in the multi-frame image, are computed. The computed reconstructions are then rendered using the differentiable renderer to produce synthetic images $S_i \in \mathbb{R}^{240 \times 240 \times 3}$. We enforce orthogonality between the geometry and expression models to lead $\mathbf{M}_{id} \mathbf{M}_{exp} = 0$. This is done by dynamically constructing \mathbf{M}_{id} in a forward pass by projecting itself onto the orthogonal complement of \mathbf{M}_{exp} [51]. Please see Fig. 2 for more details.

3.4. Dataset

We use two datasets to train our approach: *VoxCeleb* [11] and *EmotionNet* [2]. *VoxCeleb* is a multi-frame dataset consisting of over 140k videos covering 6000 different celebrities crawled from YouTube. The multi-frame nature of *VoxCeleb* allow us to train our Siamese network by feeding multiple frames (multi-frame images) for the same identity (see Fig. 2). We sample 4 images per identity from the same video clip to avoid unwanted variations due to aging, accessories and so on. 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.

We also use *EmotionNet* [2]. It is a large-scale still 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 faced images to enforce model disentanglement between the identity and expression geometry (Sec. 3.5.1).

3.5. Loss Functions

We formulate several loss functions to train our network. 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 M_{id} , M_{exp} and M_R . All the estimated parameters \mathbf{p}_i can be parametrized using

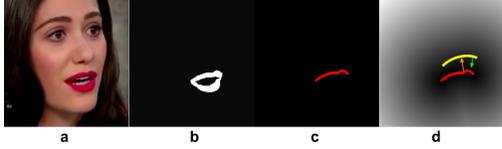


Figure 3. For a given image [a], we obtain the segmentation masks [b], its boundary [c] and distance transform (DT) image [d] of [c]. We employ a segmentation loss which tries to move the vertices on the projected mesh contour (yellow) to a lower energy position in DT. In addition, each pixel in the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

these learnable variables. Our loss function consists of:

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

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

Landmark Consistency For each frame F_i , we automatically annotate 66 sparse 2D keypoints [45] $l \in \mathbb{R}^{66}$. We compare these 2D landmarks with sparse vertices of the reconstruction which corresponds to these landmarks.

$$\mathcal{L}_{land}(\mathbf{x}) = \sum_i^{N_f} \sum_{k=0}^{66} \|l_k - p(v_k(\mathbf{x}))\|^2 . \quad (6)$$

Here, $v_k(\mathbf{x}) \in \mathbb{R}^3$ indicates the position of the k th landmark vertex, and $p(v_k(\mathbf{x}))$ is its 2D projection (Sec. 3.2). While most face landmarks can be manually annotated once 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 for the lips [28]. 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 $\mathbf{C}_a^b(x)$, where each element of $\mathbf{C}_a^b(x)$ stores a 2D pixel location. 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.

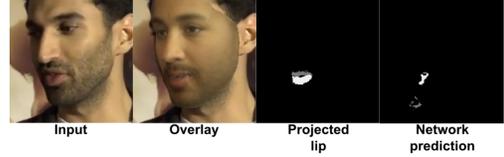


Figure 4. At test time, our approach produces plausible upper (gray) and lower (white) lip segmentation even when the images are of bad quality, contain extreme poses or occlusions. For such cases [28] struggles to produce acceptable segmentation (column 4).

$$\begin{aligned} \mathcal{L}_{seg}(\mathbf{x}) = & \sum_{i=0}^{N_f} \sum_{\forall(a,b)} [\sum_{\forall(x,y) \in \mathbf{C}_a^b(\mathbf{x})} \mathbf{D}_a^b(x,y) + \\ & \sum_{\{(x,y) | \mathbf{D}_a^b(x,y)=0\}} \|(x,y) - closest(\mathbf{C}_a^b(\mathbf{x}), (x,y))\|^2] . \end{aligned} \quad (7)$$

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 to (x,y) . We use our differentiable renderer to calculate $\mathbf{C}_l^{inner}(x)$ for the rolling inner contours. The outer contour $\mathbf{C}_l^{outer}(x)$ is 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 image F and the rendered face S .

$$\mathcal{L}_{pho}(x) = \sum_{i=0}^{N_f} \| \langle M_i, (F_i - S_i(x)) \rangle \|^2 . \quad (8)$$

M_i is a mask image with pixel value 1 where the reconstructed mesh projects to.

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 [24] to get the intermediate features for both input frames and the output synthetic frames. We then minimize the cosine distance between these features.

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

where $f_l(\cdot)$ denotes the output of the l th intermediate layer for input x and $\langle \cdot, \cdot \rangle$ denotes the inner product.

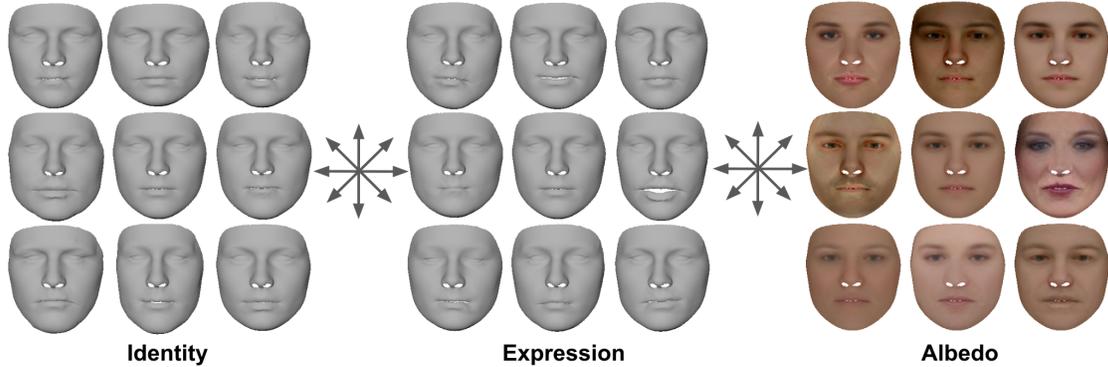


Figure 5. Visualization of learned models. Faces in each direction of indicated arrows is obtained by linearly scaling individual component of respective models. Identity geometry captures variations in face shape (second column), lips (top left to bottom right) and jaw (top right to bottom left), while expressions capture variations due to mouth opening (second row), smile (second column) and eye movement (top right to bottom left). Albedo spans a variety of skin color (second column), eye color (top right to bottom left) and gender specific features e.g. facial hair, make-up (second row).



Figure 6. Our approach reconstructs all facial components with high fidelity and good disentanglement and results a photorealistic overlay.

Geometry Smoothness To ensure smoothness of the final geometry, we use a smoothness loss at the graph level. Let $G_i \in \mathbb{R}^{N_g \times 3}$ 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.

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

where $\mathcal{N}(g)$ is the neighbourhood of node g , $g(x)$ and $n(x) \in \mathbb{R}^3$.

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 disentanglement of expressions from identity geometry.

Disentanglement Loss One example of a failure mode would be when M_{id} collapses to a zero matrix. Here, all

geometric deformations including identity would 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 3.4, a subset of our dataset includes images which correspond to neutral expression. For these images, we employ a loss function which minimizes the geometry deformations due to expressions.

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

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.

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

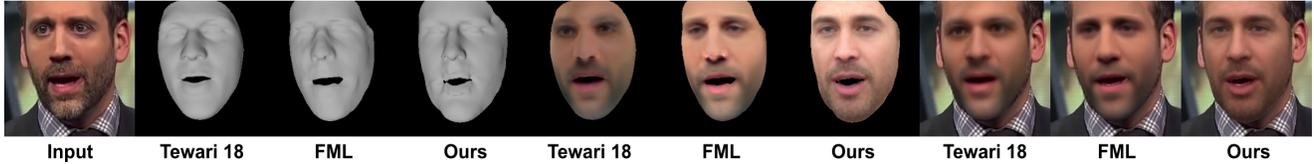


Figure 7. Our approach produces better geometry, including detailed mouth shapes compared to Tewari *et al.* [53] and FML [51]. Our albedo is also more detailed and better disentangled from the illumination.



Figure 8. Both approaches of Tran *et al.* [56, 54] do not disentangle identity geometry from expressions. Our technique, however, estimates and disentangles all facial components. It also produces produces more accurate mouth shapes.

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. 5 visualizes the different modes of 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 movements produced by the mouth and eyes, while the albedo captures different skin color, and gender specific features such as facial hair and make-up. Fig. 6 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.

Fig. 4 shows that our method can produce better lip segmentation than the approach used for generating the training data [28] 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. Fig. 10 shows that including perceptual loss in our training (Sec. 3.5) clearly improves the photorealism of the albedo and the final overlay. *Comparisons:* Fig. 8 - 10 compare our approach to several state-of-the-art face

reconstruction techniques. Tran *et al.* [56, 54] learns a combined geometry for identity and expressions, while we learn a separate model for each (Fig. 8). *MoFA* [53] and *GANFIT* [19] geometry are limited by a pretrained 3DMM model and hence lead to less detailed shapes than ours (Fig. 9). While *GANFIT* produces detailed textures, it can contain artifacts. Like other 3DMM based approaches, *RingNet* [44], which estimates the parameters of a pre-trained face model [33], struggles with out-of-space variations especially in the mouth region (Fig. 10). Tewari *et al.* [52] refines a pretrained 3DMM model on an image dataset. We can disentangle the reflectance and illumination better (Fig. 7). *FML* [51] is constrained by a pretrained expression model and thus produces less convincing shape reconstructions than us (Fig. 7). In addition, our reflectance estimates are more detailed compared to [51]. Thus, 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.

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. We evaluate this metric on the BU3DFE dataset [60] where the ground-truth geometry is obtained using 3D scans. Tab. 1 reports the results over 324 images. Our approach outperforms the approaches of MoFA [53], Tewari *et al.* [52] and *FML* [51]. 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

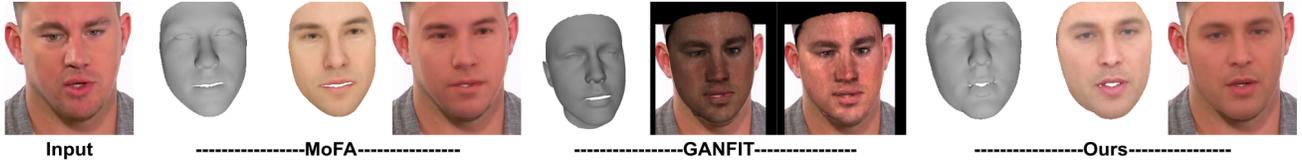


Figure 9. MoFA [53] and GANFIT [19] produce less accurate mouth shape than our technique. GANFIT [19] can produce artifacts in the albedo and final overlay, especially around the eyes.



Figure 10. Left: Our technique better captures mouth shape and eye geometry than *RingNet* [44]. It also produces a photorealistic overlay. Right: Albedo and overlay is noticeably improved with perceptual loss \mathcal{L}_{per} (see 3rd and 5th column).

shapes and significantly outperform *FML* [51] (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 average of expression deformation for images with neutral faces. We tested our approach on 1864 neutral faces mined using the same strategy described in Sec. 3.4. Tab. 3 reports the average ℓ_2 value of the expression deformations, as estimated using different approaches. The lower the expression strength, the better the disengagement. Our approach achieves significantly better expression and identity disentanglement over *FML* [51] and *MoFA* [53].

Verification Metric: To further evaluate disentanglement, we use the LFW dataset [21], which includes face image pairs of same as well as different identities. We render the identity component of the reconstructions with the predicted pose and lighting parameters. Face embeddings are computed as the average pooled version of `conv5_3` output of VGG-Face [37]. 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. Then, we compute the verification metric as the Earth Movers Distance (EMD) between these two distributions. Our method achieves an EMD of 0.15, compared to 0.09 of *FML*. A larger distance implies that the network can better represent the difference between different identities due to better disentanglement.

5. Conclusion

We presented the first approach for learning a full face model, including learned identity, reflectance and expres-

	Ours	MoFA	FML	Fine [52]	Coarse [52]
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 [60]. Our technique outperforms *MoFA* [53], coarse and fine models of Tewari *et al.* [52] and *FML et al.* [51].

	W/o \mathcal{L}_{seg}	With \mathcal{L}_{seg}	FML
UL IoU	0.49	0.52	0.51
LL IoU	0.53	0.61	0.56

Table 2. Intersection over Union (IoU) between ground-truth lip mask and the segmentation produced by different techniques. Our segmentation consistency term produces better IoU and leads to noticeably better performance than *FML* [51]

	W/o \mathcal{L}_{dis}	With \mathcal{L}_{dis}	FML	MoFA
AE	2.5075	0.0116	2.0329	0.4056

Table 3. Our identity disentanglement term results in lesser leakage of identity geometry into expression component in neutral faces. It performs better than *FML* [51] and *MoFA* [53]

sion 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.

Acknowledgments. This work was supported by the ERC Consolidator Grant 4DReply (770784).

References

- [1] 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 *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] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Computer graphics forum*, pages 641–650, 2003.
- [4] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. In *Comput. Graph. Forum*, pages 669–676, 2004.
- [5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH's Computer Graphics and Interactive Techniques*, pages 187–194, 1999.
- [6] T. Bolkart and S. Wuhler. A robust multilinear model learning framework for 3d faces. In *CVPR*, pages 4911–4919. IEEE Computer Society, 2016.
- [7] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild". In *CVPR*, 2017.
- [8] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3D morphable models. *International Journal of Computer Vision*, 126(2):233–254, April 2018.
- [9] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics*, 32(4):40:1–40:10, 2013.
- [10] C. Cao, H. Wu, Y. Weng, T. Shao, and K. 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] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.
- [12] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [13] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister. Video face replacement. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 30(6):130:1–10, December 2011.
- [14] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, 2019.
- [15] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhler, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter. 3d morphable face models – past, present and future, 2019.
- [16] V. Fernández Abrevaya, S. Wuhler, and E. Boyer. Multilinear autoencoder for 3d face model learning. In *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, 2018.
- [17] 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.
- [18] 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, 2013.
- [19] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019.
- [20] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, June 2018.
- [21] G. B. Huang, M. Ramesh, T. Berg, and E. 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.
- [22] 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.
- [23] A.-E. Ichim, P. Kadlecěk, L. Kavan, and M. Pauly. Phace: Physics-based face modeling and animation. *ACM Transactions on Graphics*, 36(4):153:1–153:14, 2017.
- [24] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*, pages 694–711, 2016.
- [25] 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 *CVPR*, 2018.
- [26] M. Kludiny, S. McDonagh, D. Bradley, T. Beeler, and K. Mitchell. Real-Time Multi-View Facial Capture with Synthetic Training. *Comput. Graph. Forum*, 2017.
- [27] 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.
- [28] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019.
- [29] G.-H. Lee and S.-W. 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.
- [30] J. P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. Pighin, and Z. Deng. Practice and Theory of Blendshape Facial Models. In S. Lefebvre and M. Spagnuolo, editors, *Eurographics*, 2014.
- [31] 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.
- [32] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Flame: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017.

- [33] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.
- [34] J. Lin, Y. Yuan, T. Shao, and K. 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.
- [35] S. McDonagh, M. Kludiny, D. Bradley, T. Beeler, I. Matthews, and K. Mitchell. Synthetic prior design for real-time face tracking. *3DV*, 00:639–648, 2016.
- [36] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 35(6), 2016.
- [37] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.
- [38] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. In *ACM Transactions on Graphics*, pages 75–84, 1998.
- [39] R. Ramamoorthi and P. Hanrahan. A signal processing framework for inverse rendering. In *ACM Trans. of Graph. (Proceedings of SIGGRAPH)*, pages 117–128. ACM, 2001.
- [40] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016.
- [41] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, July 2017.
- [42] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, pages 986–993, 2005.
- [43] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(11):2127–2141, 2017.
- [44] S. Sanyal, T. Bolkart, H. Feng, and M. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, pages 7763–7772, 2019.
- [45] J. M. Saragih, S. Lucey, and J. F. Cohn. Real-time avatar animation from a single image. In *Face and Gesture 2011*, pages 213–220, 2011.
- [46] M. Sela, E. Richardson, and R. Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *International Conference on Computer Vision (ICCV)*, 2017.
- [47] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, 2018.
- [48] 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.
- [49] E. Sifakis, I. Neverov, and R. Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Transactions on Graphics*, 24(3):417–425, July 2005.
- [50] J. R. Tena, F. De la Torre, and I. Matthews. Interactive region-based linear 3d face models. *ACM Trans. Graph.*, 30(4):76:1–76:10, July 2011.
- [51] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhoefer, and C. Theobalt. FML: Face model learning from videos. In *CVPR*, 2019.
- [52] 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 *CVPR*, 2018.
- [53] 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*, pages 3735–3744, 2017.
- [54] L. Tran, F. Liu, and X. Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.
- [55] L. Tran and X. Liu. Nonlinear 3D face morphable model. In *CVPR*, pages 7346–7355, 2018.
- [56] L. Tran and X. Liu. On learning 3d face morphable model from in-the-wild images. June 2019.
- [57] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017.
- [58] 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.
- [59] 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.
- [60] 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 (FG06)*, pages 211–216, 2006.
- [61] 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.