Towards High-fidelity Nonlinear 3D Face Morphable Model

Luan Tran, Feng Liu, Xiaoming Liu
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{tranluan, liufeng6, liuxm}@msu.edu

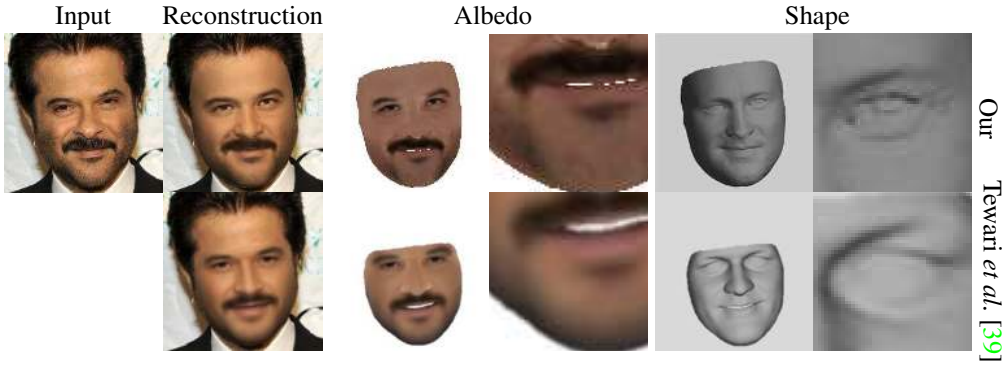

Figure 1: With novel enhancements in both learning objective as well as the network architecture, our proposed nonlinear 3D morphable model enables, for the first time, regressing high-fidelity facial shape (geometry) and albedo (skin reflectence) by directly estimating model latent representations.

Abstract

Embedding 3D morphable basis functions into deep neural networks opens great potential for models with better representation power. However, to faithfully learn those models from an image collection, it requires strong regularization to overcome ambiguities involved in the learning process. This critically prevents us from learning high fidelity face models which are needed to represent face images in high level of details. To address this problem, this paper presents a novel approach to learn additional proxies as means to side-step strong regularizations, as well as, leverages to promote detailed shape/albedo. To ease the learning, we also propose to use a dual-pathway network, a carefully-designed architecture that brings a balance between global and local-based models. By improving the nonlinear 3D morphable model in both learning objective and network architecture, we present a model which is superior in capturing higher level of details than the linear or its precedent nonlinear counterparts. As a result, our model achieves state-of-the-art performance on 3D face reconstruction by solely optimizing latent representations.

Project website: http://cvlab.cse.msu.edu/project-nonlinear-3dmm.html

1. Introduction

Computer vision and computer graphics fields have had much interest in the longstanding problem of 3D face reconstruction — creating a detailed 3D model of a person's face from a collection or a single photograph. The problem is important with many applications, including but not limited to face recognition [1, 26, 50], video editing [12, 41], avatar puppeteering [8, 10, 51] or virtual make-up [13, 24].

Recently, an incredible amount of attention is drawn into the simplest but most challenging form of this problem: monocular face reconstruction. Inferring a 3D face mesh from a single 2D photo is arduous and ill-posed since the image formation process blends multiple facial components (shape, albedo) as well as environment (lighting) into a single color for each pixel. To better handle the ambiguity, one must rely on additional prior assumptions, such as constraining faces to lie in a restricted subspace, e.g., 3D Morphable Models (3DMM) [6] learned from a small 3D scans collection. Many state-of-the-art approaches, either learning-based [33, 34] or optimization-based face reconstruction [5, 14], heavily rely on such priors. While yielding impressive results, these algorithms do not generalize well beyond the underlying model's restricted low-dimensional subspace. As a consequence, the reconstructed 3D face may

fail to recover important facial features, contain incorrect details or not well aligned to the input face.

Recently, with the flourishing in neural network, a few attempts have tried to use deep neural networks to replace the 3DMM basis functions [39, 44]. This increases the model representation power and learns model directly from unconstrained 2D images to better capture in-the-wild variations. However, even with better representation powers, these models still rely on many constraints [39] to regularize the model learning. Hence, their objectives involve the conflicting requirements of a strong regularization for a global shape vs. a weak regularization for capturing higher level details. E.g., in order to faithfully separate shading and albedo, albedo is usually assumed to be piecewise constant [22, 36], which prevents learning albedo with high level of details. In this work, besides learning the shape and albedo, we propose to learn additional shape and albedo proxies, on which we can enforce regularizations. This also allows us to flexibly pair the true shape with strongly regularized albedo proxy to learn the detailed shape or vice versa. As a result, each element can be learned with high fidelity without sacrificing the other element's quality.

On a different note, many 3DMM models fail to represent small details because of their parameterization. Many global 3D face parameterizations have been proposed to overcome the ambiguities associated with single image face fitting such as noise or occlusion. However, because they are designed to model the whole face at once, it is challenging to use them to represent small details. Meanwhile, local-based models can be more expressive than global approaches but with the cost of being less constrained to realistically represent human faces. We propose using dual-pathway networks to provide a better balance between global and local-based models. From the latent space, there is a global pathway focusing on the inference of global face structure and multiple local pathways generating details of different semantic facial parts. Their corresponding features are then fused together for successive process generation of the final shape and albedo. This network also helps to specialize filters in local pathways for each facial part which both improves the quality and saves the computation power.

In this paper, we improve the nonlinear 3D face morphable model in both learning objective and architecture:

- We solve the conflicting objective problem by learning shape and albedo proxies with proper regularization.
- The novel pairing scheme allows learning both detailed shape and albedo without sacrificing one.
- The global-local-based network architecture offers more balance between robustness and flexibility.
- Our model allows high-fidelity 3D face reconstruction by solely optimizing latent representations.

2. Prior Work

Linear 3DMM. The first generic 3D face model is built by Blanz and Vetter [6] using principal component analysis (PCA) on 3D scans. Since this seminal work, there has been a large amount of effort on improving 3DMM modeling mechanism. Paysan *et al*. [30] replace the previous UV space alignment [6] by Nonrigid Iterative Closest Point [2] to directly align 3D scans. Vlasic *et al*. [49] use a multilinear model to describe the combined effect of expression and identity variation on the facial geometry. On the texture side, Booth *et al*. [7] explore feature-based texture model to represent in-the-wild texture variations.

Nonlinear face model. Recently, there is a great interest to use deep neural networks to present the 3DMM. Early work by Duong *et al*. [29] use Deep Boltzmann Machines to present 2D Active Appearance Models. Bagautdinov *et al*. [3] use Variational Autoencoder (VAE) to learn to model facial geometry directly from 3D scans. On another direction, Tewari *et al*. [39] and Tran and Liu [44] attempt to learn 3DMM models from a 2D image collection. Tewari *et al*. [39] embed shape and albedo bases in multi-layer perceptions. Meanwhile, Tran and Liu [44] use convolution neural networks by representing both geometry and skin reflectance in UV space. Despite having greater representation power, these models still have difficulty in recovering small details in the input images due to strong regularizations in their learning objectives.

Global/local-based facial parameterization. Although, global 3D face parameterizations [23, 49] can remedy the vagueness associated with monocular face tracking [4, 11]; they can't represent small geometry details without making them exceedingly large and unwieldy. Hence, region or local-based models are proposed to overcome this problem. Blanz and Vetter [6] and Tena *et al*. [37] learn a region-based PCA, where Blanz and Vetter [6] segment the face into semantic subregions (eyes, nose, mouth), while Tena *et al*. [37] further split into smaller regions to increase the model's expressiveness. Other approaches include a region-based blendshape [18] or localized multilinear model [9]. All these models bring more flexibility than the global one but at the cost of being less constrained on realistically representing human faces. Our approach offers a balance between global and local models by using a dual-pathway network architecture. Bagautdinov *et al*. [3] try to achieve a similar objective with compositional VAE by introducing multiple layers of hidden variables, but at a cost of extremely large numbers of hidden variables.

Residual learning. Residual learning has been used in many vision tasks. In super resolution, Kim *et al*. [21] propose to learn the difference between the high-resolution target and the low-resolution input rather than estimating the target itself. In face alignment [19], or missing data imputation task [46], residual learning is used in many cascade

of networks to iteratively refine their estimation by learning the difference with the true target. In this work, we leverage residual learning idea but with a different purpose to overcome conflicting objectives in learning 3D models.

3. Proposed Method

For completeness, we start by briefly summarizing the traditional linear 3DMM, the recently proposed nonlinear 3DMM learning method including their limitations. Then we introduce our proposed improvements in both learning objective and network architecture.

3.1. Linear 3DMM

The 3D Morphable Model (3DMM) [6] provides parametric models representing faces using two components: shape (geometry) and albedo (skin reflectance). Blanz *et al*. [6] describe the 3D face space with PCA. The 3D face mesh $\mathbf{S} \in \mathbb{R}^{3Q}$ with Q vertices is computed as:

$$\mathbf{S} = \mathcal{F}_S(\mathbf{f}_S|\mathbf{\Theta}_S) = \mathbf{\Theta}_S\mathbf{f}_S, \tag{1}$$

where $\mathcal{F}_S(\mathbf{f}_S|\mathbf{\Theta}_S)$ is a function of $\mathbf{f}_S \in \mathbb{R}^{l_S}$, parameterized by $\mathbf{\Theta}_S$. In linear model, $\mathcal{F}_S$ is simply a matrix multiplication (the mean shape is omitted for clarity).

The albedo of the face $\mathbf{A} \in \mathbb{R}^{3Q}$ is defined within a template shape, describing the R, G, B colors of Q corresponding vertices. $\mathbf{A}$ is also formulated in a similar fashion:

$$\mathbf{A} = \mathcal{F}_A(\mathbf{f}_A|\mathbf{\Theta}_A) = \mathbf{\Theta}_A\mathbf{f}_A. \tag{2}$$

To synthesize 2D face images, the 3D mesh is projected onto the image plan with the weak perspective projection model. Then, the texture and 2D image is rendered using an illumination model, i.e., Spherical Harmonics [32].

3.2. Nonlinear 3DMM

Recently, Tewari *et al*. [39], Tran and Liu [44, 45] concurrently propose to use deep neural network to present 3DMM bases. Essentially, mappings $\mathcal{F}_S$ and $\mathcal{F}_A$ are now represented as neural networks with parameters Θ_S, Θ_A respectively. Tewari *et al*. [39] straightforwardly use multilayer perceptron as their networks. Meanwhile, Tran and Liu [44] leverage spatial relation of vertices by presenting both $\mathbf{S}$ and $\mathbf{A}$ in a UV space, denoted $\mathbf{S}^{\text{UV}}, \mathbf{A}^{\text{UV}}$. Mappings $\mathcal{F}_*$ are convolution neural networks (CNNs) with an extra sampling step converting from $\mathbb{R}^{\text{UV}}$ to $\mathbb{R}^{3Q}$. To make the framework end-to-end trainable, they also learn a model fitting module, $\mathcal{E}$, which is another CNN. Beside estimating shape, albedo latent vectors $\mathbf{f}_S, \mathbf{f}_A$, the encoder $\mathcal{E}$ also estimates projection matrix $\mathbf{M}$ as well as lighting coefficients $\mathbf{L}$. The objective of the whole network is to reconstruct the original input image via a differentiable rendering layer $\mathcal{R}$:

$$\begin{aligned} &\underset{\mathcal{E},\mathcal{D}_S,\mathcal{D}_A}{\arg\min} \sum_{\mathbf{I}} \mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}, \mathbf{I}), \\ &\hat{\mathbf{I}} = \mathcal{R}\left(\mathcal{E}_M(\mathbf{I}), \mathcal{E}_L(\mathbf{I}), \mathcal{F}_S(\mathcal{E}_S(\mathbf{I})), \mathcal{F}_A(\mathcal{E}_A(\mathbf{I}))\right). \end{aligned} \tag{3}$$

Reconstruction loss. There are many design options for the reconstruction loss. The straightforward choice is comparing images in the pixel space, with typical l_1 or l_2 loss. To better handle outliers, the robust $l_{2,1}$ is adopted, where the distance in the RGB color space is based on l_2 and the summation is based on l_1-norm to enforce sparsity [41, 42]:

$$\mathcal{L}^i_{\text{rec}} = \frac{1}{|\mathcal{V}|} \sum_{q \in \mathcal{V}} \left\| \hat{\mathbf{I}}(q) - \mathbf{I}(q) \right\|_2, \tag{4}$$

where $\mathcal{V}$ is the set of pixels covered by the estimated mesh.

The closeness between images $\hat{\mathbf{I}}$ and $\mathbf{I}$ can also be enforced in the feature space (perceptual loss):

$$\mathcal{L}^f_{\text{rec}} = \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} \frac{1}{W_j H_j C_j} ||\varphi_j(\hat{\mathbf{I}}) - \varphi_j(\mathbf{I})||_2^2. \tag{5}$$

The loss is summed over $\mathcal{C}$, a subset of layers of the network φ. Here $\varphi_j(\mathbf{I})$ is the activations of the j-th layer of φ with dimension $W_j \times H_j \times C_j$ obtained when processing $\mathbf{I}$.

The final reconstruction loss is a weighted average between the image and feature reconstruction losses:

$$\mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}, \mathbf{I}) = \mathcal{L}^i_{\text{rec}}(\hat{\mathbf{I}}, \mathbf{I}) + \lambda_f \mathcal{L}^f_{\text{rec}}(\hat{\mathbf{I}}, \mathbf{I}). \tag{6}$$

Sparse Landmark Alignment. To help achieve better model fitting, which in turn helps to improve the model learning itself, the landmark alignment loss is used as an auxiliary task. The loss is defined by Euclidean distance between estimated and groundtruth landmarks:

$$\mathcal{L}_{\text{lan}} = \left\| \mathbf{M} * \begin{bmatrix} \mathbf{S}(:, \mathbf{d}) \\ \mathbf{1} \end{bmatrix} - \mathbf{U} \right\|_2^2, \tag{7}$$

where $\mathbf{U} \in \mathbb{R}^{2\times 68}$ is the manual labels of 2D landmark locations, $\mathbf{d}$ stores the indexes of 68 vertices corresponding to the sparse 2D landmarks in the 3D face mesh. In [44,45], the landmark loss is only applied on $\mathcal{E}$ to prevent learning implausible shapes as the loss only affects a tiny subsets of vertices related to the keypoints.

Different regularization. To overcome ambiguity and faithfully recover different elements (shape, albedo, lighting), many regularizations are needed.

Albedo Symmetry:

$$\mathcal{L}_{\text{sym}}(\mathbf{A}) = \left\| \mathbf{A}^{\text{uv}} - \text{flip}(\mathbf{A}^{\text{uv}}) \right\|_1, \tag{8}$$

where flip() is a horizontal image flip operation.

Albedo Constancy:

$$\mathcal{L}_{\text{con}}(\mathbf{A}) = \sum_{\mathbf{v}^{\text{uv}}_j \in \mathcal{N}_i} \omega(\mathbf{v}^{\text{uv}}_i, \mathbf{v}^{\text{uv}}_j) \left\| \mathbf{A}^{\text{uv}}(\mathbf{v}^{\text{uv}}_i) - \mathbf{A}^{\text{uv}}(\mathbf{v}^{\text{uv}}_j) \right\|_2^p. \tag{9}$$

The weight $\omega(\mathbf{v}^{\text{uv}}_i, \mathbf{v}^{\text{uv}}_j) = \exp\left(-\alpha \left\| \mathbf{c}(\mathbf{v}^{\text{uv}}_i) - \mathbf{c}(\mathbf{v}^{\text{uv}}_j) \right\|\right)$, helps to penalize more on pixels with the same chromaticity

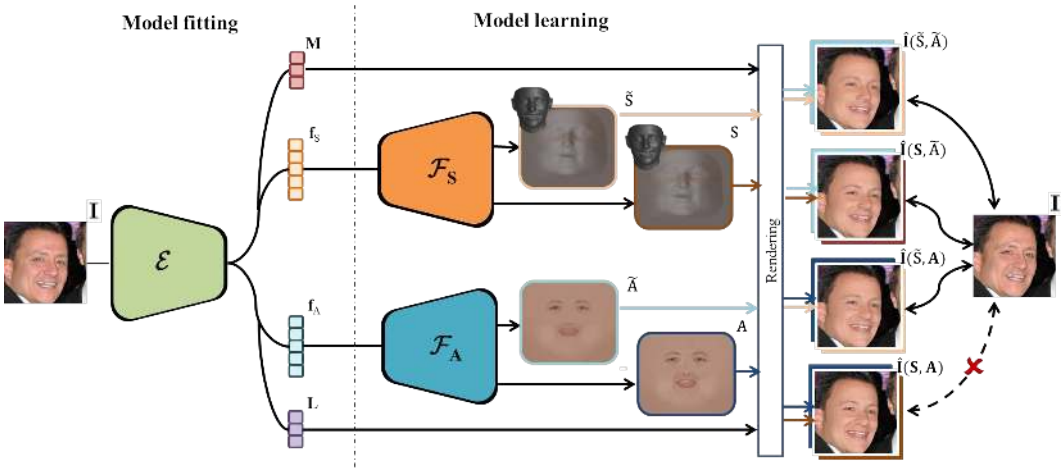

Figure 2: The proposed framework. Each shape or albedo decoder consist of two branches to reconstruct the true element and its proxy. Proxies free shape and albedo from strong regularizations, allow them to learn models with high level of details.

(i.e., $\mathbf{c}(x) = \mathbf{I}(x)/|\mathbf{I}(x)|$), where the color is referenced from the input image using the current estimated projection. $\mathcal{N}_i$ denotes a set of 4-pixel neighborhood of pixel $\mathbf{v}_i^{\text{uv}}$.

Shape Smoothness: This is a Laplacian regularization on the vertex locations.

$$\mathcal{L}_{\text{smo}}(\mathbf{S}) = \sum_{\mathbf{v}_i^{\text{uv}} \in \mathbf{S}^{\text{uv}}} \left\| \mathbf{S}^{\text{uv}}(\mathbf{v}_i^{\text{uv}}) - \frac{1}{|\mathcal{N}_i|} \sum_{\mathbf{v}_j^{\text{uv}} \in \mathcal{N}_i} \mathbf{S}^{\text{uv}}(\mathbf{v}_j^{\text{uv}}) \right\|_2 . \quad (10)$$

The overall objective can be summarized as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_{\text{lan}} + \mathcal{L}_{\text{reg}}, \quad (11)$$

$$\text{with } \mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{sym}}(\mathbf{A}) + \lambda_{\text{con}}\mathcal{L}_{\text{con}}(\mathbf{A}) + \lambda_{\text{smo}}\mathcal{L}_{\text{smo}}(\mathbf{S}). \quad (12)$$

3.3. Nonlinear 3DMM with Proxy and Residual

Proxy and Residual Learning. Strong regularization has been shown to be critical in ensuring the plausibility of the learned models [39,45]. However, the strong regularization also prevents the model from recovering high-level details in either shape or albedo. Hence, this prevents us from achieving the ultimate goal of learning a high-fidelity 3DMM model.

In this work, we propose to learn additional **proxy shape** ($\tilde{\mathbf{S}}$) and **proxy albedo** ($\tilde{\mathbf{A}}$), on which we can apply the regularization. All presented regularizations will now be moved to proxies:

$$\mathcal{L}^*_{\text{reg}} = \mathcal{L}_{\text{sym}}(\tilde{\mathbf{A}}) + \lambda_{\text{con}}\mathcal{L}_{\text{con}}(\tilde{\mathbf{A}}) + \lambda_{\text{smo}}\mathcal{L}_{\text{smo}}(\tilde{\mathbf{S}}). \quad (13)$$

There will be no regularization applied directly to the actual shape $\mathbf{S}$ and albedo $\mathbf{A}$, other than a weak regularization encouraging each to be close to its proxy:

$$\mathcal{L}_{\text{res}} = \|\Delta\mathbf{S}\|_1 + \|\Delta\mathbf{A}\|_1 = \left\|\mathbf{S} - \tilde{\mathbf{S}}\right\|_1 + \left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_1 . \quad (14)$$

By pairing two shapes $\mathbf{S}, \tilde{\mathbf{S}}$ and two albedos $\mathbf{A}, \tilde{\mathbf{A}}$, we can render four different output images (Fig. 2). Any of them can be used to compare with the original input image. We rewrite our reconstruction loss as:

$$\begin{aligned}\mathcal{L}^*_{\text{rec}} = &\mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\tilde{\mathbf{S}}, \tilde{\mathbf{A}}), \mathbf{I}) \\ &+ \mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\tilde{\mathbf{S}}, \mathbf{A}), \mathbf{I}) \\ &+ \mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\mathbf{S}, \tilde{\mathbf{A}}), \mathbf{I}). \end{aligned} \quad (15)$$

Pairing strongly regularized proxies and weakly regularized components is a critical point in our approach. Using proxies allows us to learn high-fidelity shape and albedo without sacrificing quality of either component. This pairing is inspired by the observation that Shape from Shading techniques are able to recover detailed face mesh by assuming over regularized albedo or even using the mean albedo [34]. Here, $\mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\mathbf{S}, \tilde{\mathbf{A}}), \mathbf{I})$ loss promotes $\mathbf{S}$ to recover more details as $\tilde{\mathbf{A}}$ is constrained by piece-wise constant $\mathcal{L}_{\text{con}}(\tilde{\mathbf{A}})$ objective. Vice versa, $\mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\tilde{\mathbf{S}}, \mathbf{A}), \mathbf{I})$ aims to learn better albedo. In order for these two losses to work as desired, proxies $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{A}}$ should perform well enough to approximate the input images by themselves. Without $\mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\tilde{\mathbf{S}}, \tilde{\mathbf{A}}), \mathbf{I})$, a valid solution that minimizes $\mathcal{L}_{\text{rec}}(\hat{\mathbf{I}}(\mathbf{S}, \tilde{\mathbf{A}}), \mathbf{I})$ is combination of a constant albedo proxy and noisy shape creating surface normal with dark shading in necessary regions, i.e., eyebrows.

Another notable design choice is that we intentionally left out the loss function on $\hat{\mathbf{I}}(\mathbf{S}, \mathbf{A})$, even though this theoretically is the most important objective. This is to avoid the case that the shape $\mathbf{S}$ learns an in-between solution that works well with both $\tilde{\mathbf{A}}, \mathbf{A}$ and vice versa.

Occlusion Imputation. With proposed objective function, our model is able to faithfully reconstruct input images. However, we empirically found that besides high-

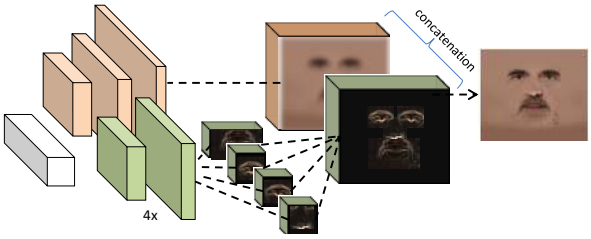

Figure 3: The proposed global-local-based network architecture.

fidelity visible regions, the model tends to keep invisible region smooth. The reason might be that, there is no supervision on those areas other than the residual magnitude loss pulling the shape and albedo closer to their proxies. To learn a more meaningful model, which is beneficial to other applications, i.e., face editing or face synthesis, we propose to use a soft symmetry loss [43] on occluded regions:

$$\mathcal{L}_{\text{res-sym}}(\mathbf{S}) = \|\mathbf{T} \odot (\Delta\mathbf{S}_z^{\text{uv}} - \text{flip}(\Delta\mathbf{S}_z^{\text{uv}}))\|_1 , \qquad (16)$$

where $\mathbf{T}$ is a visibility mask of each pixel in UV space, approximated based on estimated surface normal direction. Even though the shape itself is not symmetric, i.e., face with asymmetric expression; we enforce symmetrical property on its depth residual $\Delta\mathbf{S}_z$ (only use shape's z-dimension).

3.4. Global-Local-Based Network Architecture

While global-based models are usually robust to noise and mismatches, they are usually over-constrained and do not provide sufficient flexibility to represent high-frequency deformations as local-based models. In order to take the best of both worlds, we propose to use dual-pathway networks for our shape and albedo decoders.

Here, we transfer the success of combining local and global models in image synthesis [15, 28] to 3D face modeling. The general architecture of a decoder is shown in Fig. 3. From the latent vector, there is a global pathway focusing on inferring the global structure and a local pathway with four small sub-networks generating details of different facial parts, including eyes, nose and mouth. The global pathway is built from fractional strided convolution layers with five up-sampling steps. Meanwhile, each sub-network in the local pathway has the similar architecture but shallower with only three up-sampling steps. Using different small sub-networks for each facial part offers two benefits: i) with less up-sampling steps, the network is better able to represent high-frequency details in early layers; ii) each sub-network can learn part-specific filters, which is more computationally efficient than applying across global face.

As shown in Fig. 3, to fuse two pathways' features, we firstly integrate four local pathways' outputs into one single feature tensor. Different from other works that synthesize face images with different yaw angles [20, 47, 48] with no fixed keypoints' locations, our 3DMM generates facial albedo as well as 3D shape in UV space with predefined topology. Merging these local feature tensors is efficiently done with the zero padding operation. The max-pooling fusion strategy is also used to reduce the stitching artifacts on the overlapping areas. Then the resultant feature is simply concatenated with the global pathway's feature, which has the same spatial resolution. Successive convolution layers integrate information from both pathways and generate the final albedo/shape (or their proxies).

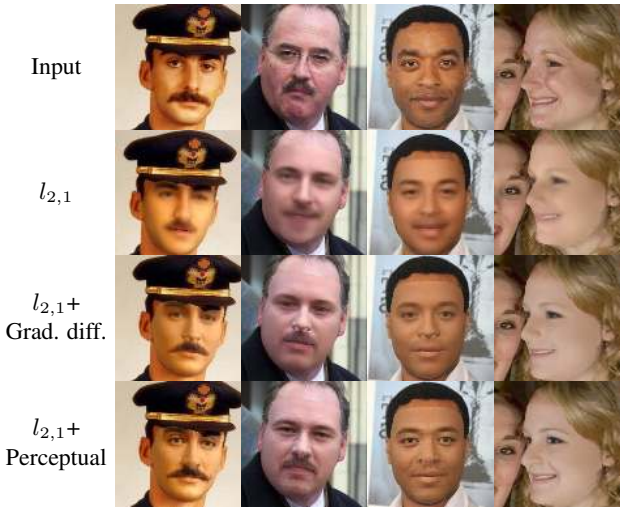

Figure 4: Reconstruction results with different loss functions.

4. Experimental Results

We study different aspects of the proposed framework, in terms of framework design, model representation power, and applications to facial analysis.

The training is similar to [45], which also include a pretrain stage with supervised losses. Adopting Basel Face Model (BFM) [30]'s facial mesh triangle topology, we use a subset of $Q = 39,111$ vertices on the face region only. The model is trained on 300W-LP dataset [52], which contains $122,450$ in-the-wild face images, in a wide pose range.

The model is optimized using Adam optimizer with a learning rate of 0.001. We set the following parameters: $U = 192, V = 224$, $l_S = l_A = 320$. λ values are selected to bring losses to similar magnitudes.

4.1. Ablation Study

Reconstruction Loss Functions. We study effects of different reconstruction losses on quality of the reconstructed images (Fig. 4). As expected, the model trained with $l_{2,1}$ loss only results in blurry reconstruction, similar to other l_p loss. To make the reconstruction more realistic, we explore other options such as gradient difference [27] or perceptual loss [17]. While adding the gradient difference loss creates more details in the reconstruction, combining perceptual loss with $l_{2,1}$ gives the best results with high level of details and realism. For the rest of the paper we will refer

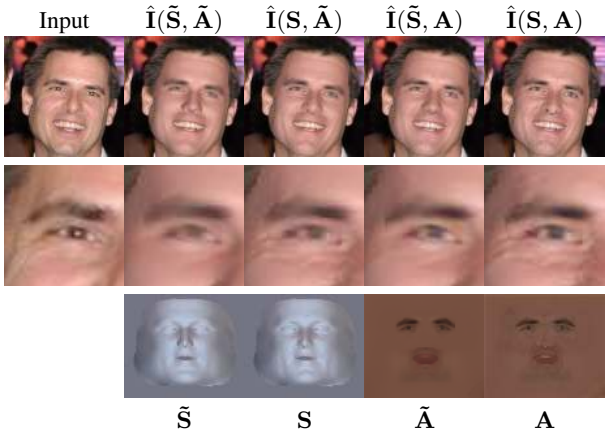

Figure 5: Image reconstruction with our 3DMM model using the proxy and the true shape and albedo. Our shape and albedo can faithfully recover details of the face. Note: for the shape, we show the shading in UV space – a better visualization than the raw $\mathbf{S}^{\text{UV}}$.

Figure 6: Affect of soft symmetry loss on our shape model.

to the model trained using this combination.

Understanding image pairing. Fig. 5 shows fitting results of our model on a 2D face image. By using the proxy or the final components (shape or albedo) we can render four different reconstructed images with different quality and characteristics. The image generated by two proxies $\tilde{\mathbf{S}}, \tilde{\mathbf{A}}$ is quite blurry but is still be able to capture major variations in the input face. By pairing $\mathbf{S}$ and the proxy $\tilde{\mathbf{A}}$, S is enforced to capture high level of details to bring the image closer to the input. Similarly, $\mathbf{A}$ is also encouraged to capture more details by pairing with the proxy $\tilde{\mathbf{S}}$. The final image $\hat{\mathbf{I}}(\mathbf{S}, \mathbf{A})$ inherently achieves high level of details and realism even without direct optimization.

Residual Soft Symmetry Loss. We study effects of the residual soft symmetry loss on recovering details on occluded face region. As shown in Fig. 6, without $\mathcal{L}_{\text{res-sym}}$, the learned model can result in an unnatural shape, in which one side of the face is over-smooth, on occluded regions, while the other side still has high level of details. Our model learned with $\mathcal{L}_{\text{res-sym}}$ can consistently create details across the face, even in occluded areas.

4.2. Representation Power

We compare the representation power of the proposed nonlinear 3DMM with Basel Face Model [30], the most commonly used linear 3DMM. We also make comparisons with the recently proposed nonlinear 3DMM [44].

Texture. We evaluate our model's power to represent

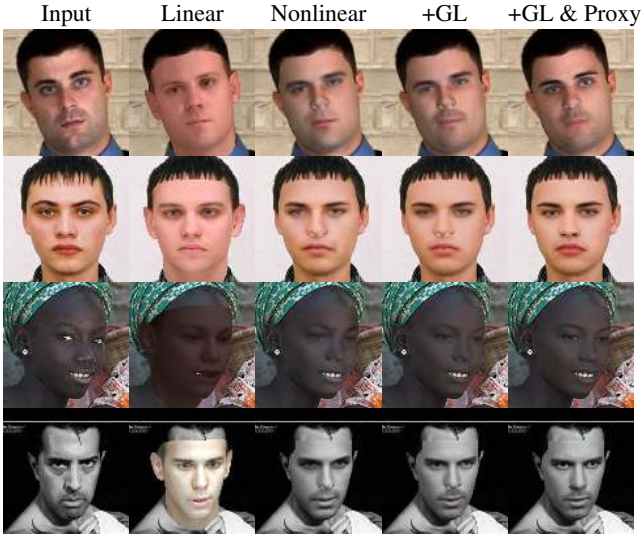

Figure 7: Qualitative comparisons on texture representation power. Our model can better reconstruct in-the-wild facial texture.

Table 1: Texture representation power quantitative comparison (Average reconstruction error on non-occluded face portion.)

Method	Reconstruction error ($l_{2,1}$)
Linear [52]	0.1287
Nonlinear [45]	0.0427
Nonlinear + GL (Ours)	0.0386
Nonlinear + GL + Proxy (Ours)	**0.0363**

in-the-wild facial texture on AFLW2000-3D dataset [52]. Given a face image, also with the groundtruth geometry and camera projection, we can jointly estimate an albedo parameter $\mathbf{f}_A$ and a lighting parameter $\mathbf{L}$ whose decoded texture can reconstruct the original image. To accomplish this, we use SGD on $\mathbf{f}_A$ and $\mathbf{L}$ with the initial parameters estimated by our encoder $\mathcal{E}$. For the linear model, Zhu *et al*. [52] fitting results of Basel albedo using Phong illumination model [31] is used. As in Fig. 7, nonlinear model significantly outperforms the Basel Face model. Despite, being close to the original image, Tran and Liu [45] model reconstruction results are still blurry. Using global-local-based network architecture ("+GL") with the same loss functions helps to bring the image closer to the input. However, these models are still constrained by regularizations on the albedo. By learning using proxy technique ("+Proxy"), our model can learn more realistic albedo with more high frequency details on the face. This conclusion is further supported with quantitative comparison in Tab. 1. We report the averaged $l_{2,1}$ reconstruction error over the face portion of each image. Our model achieves the lowest averaged reconstruction error among four models, 0.0363, which is a 15% error reduction of the recent nonlinear 3DMM work [45].

Shape. Similarly, we also compare models' power to represent real-world 3D scans. Using ten 3D face meshes

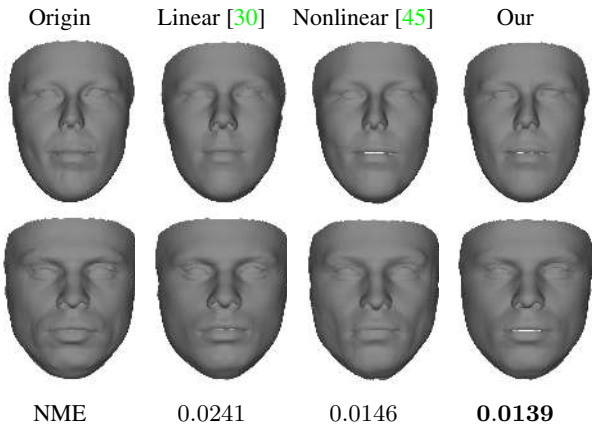

Figure 8: Shape representation power comparison. Given a 3D shape, we optimize the feature $\mathbf{f}_S$ to approximate the original one.

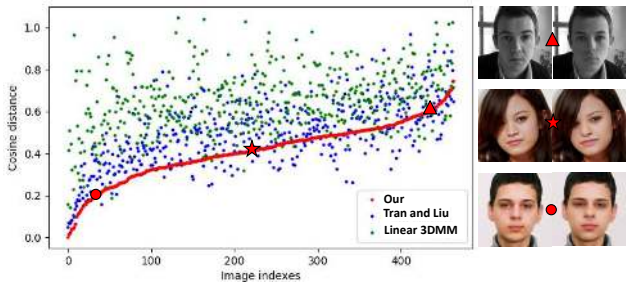

Figure 9: The distance between the input images and their reconstruction from three models. For better visualization, images are sorted based on their distance to our model's reconstructions.

provided by [30], which share the same triangle topology with us, we can optimize the shape parameter to generate, through the decoder, shapes matching the groundtruth scans. The optimization objective is defined based on vertex distances (Euclidean) as well as surface normal direction (cosine distance), which empirically improves reconstructed meshes' fidelity compared to optimizing the former only. Fig. 8 shows the visual comparisons between different reconstructed meshes. Our reconstructions closely match the face shapes details. To make quantitative comparisons, we use NME — averaged per-vertex Euclidean distances between the recovered and groundtruth meshes, normalized by inter-ocular distances. The proposed model has a significantly smaller reconstruction error than the linear model, and is also smaller than the nonlinear model by Tran and Liu [45] (0.0139 vs. 0.0146 [45], and 0.0241 [30]).

4.3. Identity-Preserving

We explore the effect of our proposed 3DMM on preserving identity when reconstructing face images. Using DR-GAN [48], a pretrained face recognition network, we can compute the cosine distance between the input and its reconstruction from different models. Fig. 9 shows the plot of these score distributions. At each horizontal mark, there

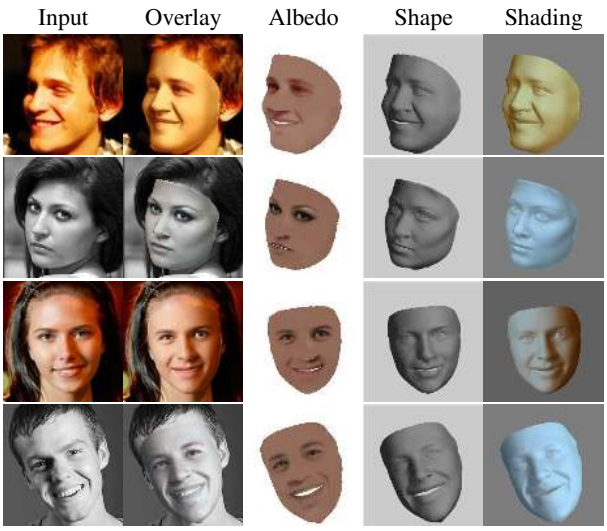

Figure 10: Model fitting on faces with diverse skin color, pose, expression, lighting. Our model faithfully recovers these cues.

Figure 11: 3D reconstruction comparison to Tewari *et al.* [40].

are exactly three points presenting distances between an image with its reconstructions from three models. Images are sorted based on the distance to our reconstruction. For the majority of the cases (77.2%), our reconstruction has the smallest difference to the input in the identity space.

4.4. 3D Reconstruction

Using our model $\mathcal{F}_S, \mathcal{F}_A$, together with the model fitting CNN $\mathcal{E}$, we can decompose a 2D photograph into different components: 3D shape, albedo and lighting (Fig. 10). Here we compare our 3D reconstruction results with different lines of works: linear 3DMM fitting [40], nonlinear 3DMM fitting [39, 45] and approaches beyond 3DMM [16, 35]. Comparisons are made on CelebA dataset [25].

For linear 3DMM model, the representative work, MoFA by Tewari *et al.* [38, 40], learns to regress 3DMM parameters in an unsupervised fashion. Even being trained on in-the-wild images, it is still limited to the linear subspace, with limited power to recovering in-the-wild texture. This results in the surface shrinkage when dealing with challenging texture, i.e., facial hair as discussed in [39, 44, 45]. Besides, even with regular skin texture their reconstruction is

Figure 12: 3D reconstruction comparisons to nonlinear 3DMM approaches by Tewari *et al*. [39] or Tran and Liu [45]. Our model can reconstruct face images with higher level of details. Please zoom-in for more details. Best view electronically.

Figure 13: 3D reconstruction comparisons to Sela *et al*. [35] or Tran *et al*. [43], which go beyond latent space representations.

still blurry and has less details compared to ours (Fig. 11).

The most related work to our proposed model is Tewari *et al*. [39], Tran and Liu [45], in which 3DMM bases are embedded in neural networks. With more representation power, these models can recover details that the traditional 3DMM usually can't, i.e. make-up, facial hair. However, the model learning process is attached with strong regularization, which limits their ability to recover high-frequency details of the face. Our propose model enhances the learning process in both learning objective and network architecture to allow higher-fidelity reconstructions (Fig. 12).

To improve 3D reconstruction quality, many approaches also try to move beyond the 3DMM such as Richardson *et al*. [34], Sela *et al*. [35] or Tran *et al*. [43]. The current state-of-the-art 3D monocular face reconstruction method by Sela *et al*. [35] using a fine detail reconstruction step to help reconstructing high fidelity meshes. However, their first depth map regression step is trained on synthetic data generated by the linear 3DMM. Besides domain gap between synthetic and real, it faces a more serious problem of lacking facial hair in the low-dimension texture. Hence, this network's output tends to ignore these unexplainable regions, which leads to failure in later steps. Our network is more robust in handling these in-the-wild variations (Fig. 13). The approach of Tran *et al*. [43] shares a similar objective with us to be both robust and maintain high level of details in 3D reconstruction. However, they use an over-constrained foundation, which loses personal characteristics of the each face mesh. As a result, the 3D shapes look similar across different subjects (Fig. 13).

Figure 14: Adding stickers to faces. The sticker is naturally added into faces following the surface normal or lighting.

4.5. Facial Editting

With more precise 3D face mesh reconstruction, the quality of successive tasks is also improved. Here, we show an application of our model on face editing: adding stickers or tattoos onto faces. Using the estimated shape as well as the projection matrix, we can unwrap the facial texture into the UV space. Thanks to the lighting decomposition, we can also remove the shading from the texture to get the detailed albedo. From here we can directly edit the albedo by adding sticker, tattoo or make-up. Finally, the edited images can be rendered using the modified albedo together with other original elements. Fig. 14 shows our editing results by adding stickers into different people's face.

5. Conclusions

In realization that the strong regularization and global-based modeling are the roadblocks to achieve high-fidelity 3DMM model, this work presents a novel approach to improve the nonlinear 3DMM modeling in both learning objective and network architecture. Hopefully, with insights and findings discussed in the paper, this work can be a step toward unlocking the possibility to build a model which can capture mid and high-level details in the face. Through which, high-fidelity 3D face reconstruction can be achieved solely by doing model fitting.

References

[1] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *FG*, 2008. 1

[2] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *CVPR*, 2007. 2

[3] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional VAEs. In *CVPR*, 2018. 2

[4] Michael J Black and Yaser Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV*, 1995. 2

[5] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*. Wiley Online Library, 2003. 1

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 1, 2, 3

[7] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3D face morphable models "In-the-wild". In *CVPR*, 2017. 2

[8] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM TOG*, 2013. 1

[9] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *ECCV*, 2014. 2

[10] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM TOG*, 2014. 1

[11] Douglas DeCarlo and Dimitris Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *CVPR*, 1996. 2

[12] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer Graphics Forum*. Wiley Online Library, 2015. 1

[13] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM TOG*, 2013. 1

[14] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM TOG*, 2016. 1

[15] Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017. 5

[16] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *ICCV*, 2017. 7

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[18] Pushkar Joshi, Wen C Tien, Mathieu Desbrun, and Fr'ed'eric Pighin. Learning controls for blend shape based realistic facial animation. In *ACM Siggraph 2006 Courses*, 2006. 2

[19] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3D face alignment. In *ICCV*, 2015. 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 5

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[22] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 1971. 2

[23] Manfred Lau, Jinxiang Chai, Ying-Qing Xu, and Heung-Yeung Shum. Face poser: Interactive modeling of 3D facial expressions using facial priors. *ACM TOG*, 2009. 2

[24] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *CVPR*, 2015. 1

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7

[26] Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 1

[27] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv:1511.05440*, 2015. 5

[28] Umar Mohammed, Simon JD Prince, and Jan Kautz. Visiolization: generating novel facial images. *ACM TOG*, 2009. 5

[29] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D Bui. Beyond principal components: Deep Boltzmann Machines for face modeling. In *CVPR*, 2015. 2

[30] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, 2009. 2, 5, 6, 7

[31] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 1975. 6

[32] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 3

[33] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016. 1

[34] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017. 1, 4, 8

[35] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017. 7, 8

[36] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 2

[37] J Rafael Tena, Fernando De la Torre, and Iain Matthews. Interactive region-based linear 3D face models. In *ACM TOG*, 2011. 2

[38] Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *TPAMI*, 2018. 7

[39] 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. 1, 2, 3, 4, 7, 8

[40] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 7

[41] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *CVPR*, 2016. 1, 3

[42] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. FaceVR: Real-time facial reenactment and eye gaze control in virtual reality. *ACM TOG*, 2018. 3

[43] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gerard Medioni. Extreme 3D face reconstruction: Looking past occlusions. In *CVPR*, 2018. 5, 8

[44] Luan Tran and Xiaoming Liu. Nonlinear 3D morphable model. In *CVPR*, 2018. 2, 3, 6, 7

[45] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *arXiv:1808.09560*, 2018. 3, 4, 5, 6, 7, 8

[46] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*, 2017. 2

[47] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 5

[48] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *TPAMI*, 2018. 5, 7

[49] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM TOG*, 2005. 2

[50] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 1

[51] Eduard Zell, JP Lewis, Junyong Noh, Mario Botsch, et al. Facial retargeting with automatic range of motion alignment. *ACM TOG*, 2017. 1

[52] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. 5, 6