Semi-supervised 3D Face Representation Learning from Unconstrained Photo Collections

Zhongpai Gao[1], Juyong Zhang[2], Yudong Guo[2], Chao Ma[1], Guangtao Zhai[1], and Xiaokang Yang[1]
[1]Artificial Intelligence Institute, Shanghai Jiao Tong University
[2]University of Science and Technology of China

Abstract

Recovering 3D geometry shape, albedo, and lighting from a single image is a typical ill-posed problem. To address this challenging problem, we propose to utilize the joint constraints from unconstrained photo collections of one person to recover his or her identity shape and albedo. Unconstrained photo collections include one's photos captured under different times, backgrounds, and expressions, e.g., photos posted on Instagram. We train our model in a semi-supervised manner with adversarial loss to exploit large amounts of unconstrained facial images. A novel center loss is introduced to make sure that facial images from the same subject have the same identity shape and albedo. Besides, our proposed model disentangles identity, expression, pose, and lighting representations, which improves the overall reconstruction performance and facilitates facial editing applications, e.g., expression transfer. Comprehensive experiments demonstrate that our model produces high-quality reconstruction compared to state-of-the-art methods and is robust to various expression, pose, and lighting conditions.

1. Introduction

Reconstructing 3D face from 2D images enables a wide range of computer vision applications, such as face recognition [3, 27, 23, 43], face puppetry [9], face reenactment [33, 14], virtual make-up [21], etc. However, inferring 3D face shape and texture from 2D images, especially from a single image, is an ill-posed problem due to the missing 3D information during the imaging process. 3D morphable model (3DMM) [2] learned from a collection of 3D face scans often serves as a strong prior assumption for this problem. 3DMM linearly combines a set of bases to provide a statistical parametric representation of 3D faces. Classical optimization-based methods [3, 20, 5] take a local optimal solution by regressing the 3DMM parameters, which is time-consuming due to the high optimization complexity.

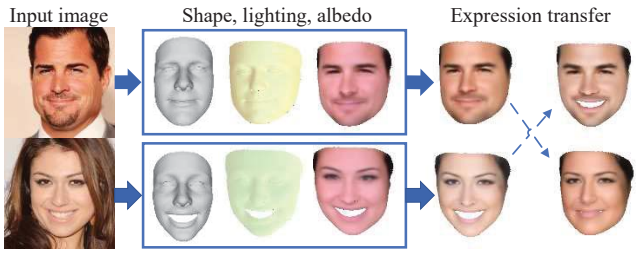


Figure 1: Our model is trained on unconstrained photo collections and extracts four disentangled representations from an input image: identity, expression, pose, and lighting, which allows applications such as expression transfer.

On the other hand, learning-based methods learn deep regression models via convolutional neutral networks (CNN) [29, 31, 15, 46, 36, 13]. Despite the demonstrated success, these methods only search for a solution in the restricted linear low-dimensional subspace of 3DMM and cannot generalize well in the wild.

Furthermore, learning models from a single view causes unresolvable ambiguities due to the lack of reliable 3D constraints. Multi-view geometric constraints using a set of facial images in different views can improve the reliability and achieve favorable results [41]. However, multi-view facial images are difficult and expensive to acquire. On the other hand, people have large amounts of unconstrained photos in smartphone's photo albums and on social media, e.g., Instagram and Wechat Moments. The unconstrained photo collections are captured in various expression, pose, lighting, and occlusion conditions. The joint constraints from unconstrained photo collections of one person can be used to recover his or her identity shape and albedo.

In this paper, we learns 3D face representations from unconstrained photo collections without constrained by a linear 3DMM. We propose a novel encoder-decoder architecture using inverse rendering that bridges computer vision and computer graphics techniques. The vision system (i.e., encoder network) decomposes an input 2D face image into disentangled and semantic representations: identity, ex-

pression, pose, and lighting code. Two decoder networks regress the 3D face shape and albedo from the extracted representations, so that the graphics system can render back a face image to match the input image. As such, we provide a unique opportunity to leverage the vast amounts of readily available unlabeled face images (i.e., without the ground truth of 3D face shapes) from unconstrained photo collections through self-supervised learning.

Since 3D face reconstruction from a 2D image is ambiguous and ill-posed, self-supervised learning with unlabeled data through inverse rendering is not sufficient. In this paper, we train the network in a semi-supervised manner on hybrid batches of large amounts of unlabeled face images and relatively small amounts of labeled face images that are generated from a linear 3DMM using optimization-based methods. Moreover, inspired by generative adversarial networks (GAN) [16], we use a discriminator network to ensure the reconstructed face shape not far away from the distribution of human faces. Since the distribution of human faces is unknown, we sample the 3D face shapes from a linear parametric 3DMM during adversarial training, which prevents our model from generating unrealistic 3D face shapes and constrains our model to be close to but not strictly limited to the linear 3DMM. Importantly, we are able to exploit a large amount of unlabeled face images as training data.

To reconstruct the 3D face shape, we use graph convolutional network (GCN) [12, 19] instead of fully connected layers or CNN as in [35, 34]. Since 3D face shape is usually modeled as a mesh that consists of a collection of vertices, edges, and faces and can be viewed as an unstructured graph. Performing graph convolutions on 3D meshes is memory efficient and allows for processing high resolution 3D structures. GCN-based methods [28, 18, 7] have shown promising results in reconstructing 3D face shapes. To recover 3D face albedo, we first use a GCN that has the same architecture with the shape decoder to learn an illumination-independent face albedo. Then we apply a CNN-based decoder network that has skip connections with the encoder network [30] to capture the facial texture details.

We apply a face recognition loss and a center loss [40] to learn the identity representation (i.e., one's identity) from one's unconstrained multiple face images. The center loss ensures the identity representation's compactness for each person and separability for different people, so that the identity representation is disentangled from the pose, lighting, and expression representations. In order to further disentangle the identity and expression representations, we adopt pairwise training approaches. Given a pair of labeled face data, we keep the identity codes and interchange the expression codes of 3DMM to generate new 3D shapes as supervision. Comprehensive evaluation experiments show that the proposed method achieves state-of-the-art performance in 3D face reconstruction and can easily be used for the applications of face recognition and facial expression transfer. The main contributions of this paper are summarized below:

- We propose an efficient semi-supervised and adversarial training process to fully exploit unconstrained photo collections and go beyond the limitation of a linear 3DMM.
- We design a novel framework to extract nonlinear disentangled representations from a face image with the help of face recognition losses and shape pairwise loss.
- Extensive experiments show that our model achieves state-of-the-art performance in face reconstruction.

2. Related work

Linear 3D face models Blanz & Vetter [2] proposed the first linear parametric 3DMM using principal component analysis (PCA) to model the shape and texture of 3D faces. Booth *et al.* [6] built a linear face model from around 10,000 facial scans of more diverse subjects but only in neutral expressions. Vlasic *et al.* [39], Cao *et al.* [10], and Li *et al.* [22] developed bilinear/multilinear face models with separate attributes of identity and expression to support a wide variety of face manipulation applications. Bolkart & Wuhrer [4] proposed a multilinear face model by jointly optimizing the model parameters and the facial scan registrations. The most popular 3DMM [45] was built by merging Basel Face Model (BFM) [27] with only 200 subjects in neutral expressions and FaceWarehouse [10] with 150 subjects in 20 different expressions.

Linear 3D face models based reconstruction 3DMM is a strong prior for monocular 3D face reconstruction. The methods of fitting 3DMM can be grouped into two types: optimization-based approaches [3, 20, 5] that obtain the 3DMM parameters by solving complex optimization problems and learning-based approaches [29, 31, 15, 46, 36] that directly regress the 3DMM parameters using CNN. However, these 3DMM fitting methods are based on a linear 3DMM learned from limited facial scans via PCA. Linear statistical models have limitations to construct 3D faces with various ethnic groups, ages, occlusions, lightings, and facial expressions. Tewari *et al.* [32], Tran *et al.* [38], and Guo. *et al.* [17] further proposed 3D face models composed of two networks: a coarse-scale linear 3DMM network and a fine-scale corrective network. Even though the fine-scale corrective model can generate more details, 3D face reconstruction will fail if the foundation face shape generated by the linear 3DMM network is not good enough. Liu *et al.* [23] proposed an encoder-decoder network to extract disentangled shape features from single images and directly regress 3D face shapes from the features. Although this method is not constrained by a pre-existing linear 3DMM, it is still a linear face model since the decoders were imple-

mented as a fully connected (FC) layer.

Nonlinear 3D face models based reconstruction Tran *et al.* [35, 34] proposed encoder-decoder networks to regress the face shape and texture directly. The nonlinear networks have higher representation power compared to a linear model and are able to reconstruct high-fidelity facial texture. However, the nonlinear models were only trained on the 300W-LP dataset [44] that was generated based on a linear 3DMM with a face profiling technique. Besides, the model training process does not consider each image's facial identity. The face albedo and shape are decoded from an albedo code and a shape code separately. In fact, facial images from the same person have the same face albedo and identity shape. Learning the albedo and shape parameter separately is difficult to disentangle the face albedo from lightings and occlusions. As a result, the albedo decoder may reconstruct high-fidelity face albedo without aligning with the face shape and thus fails to contribute to the face shape reconstruction. At last, the identity and expression representations are entangled in these methods, disabling a large number of applications, such as face recognition, face animation, and face reenactment.

3. Background

This section describes some background information related to our work, including face representations in conventional linear 3DMM and the rendering process.

Linear 3DMM We first recap the conventional linear 3DMM. As described in [11], the linear 3DMM constructed from facial scans via PCA can be expressed as:

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{A}_{id}\boldsymbol{\alpha}_{id} + \boldsymbol{A}_{exp}\boldsymbol{\alpha}_{exp}, \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^{3N\times 1}$ is a 3D face shape with N vertices, $\bar{\boldsymbol{s}} \in \mathbb{R}^{3N\times 1}$ is the mean shape, $\boldsymbol{A}_{id} \in \mathbb{R}^{3N\times K}$ is the first K principle components trained on facial scans with neutral expression and $\boldsymbol{\alpha}_{id} \in \mathbb{R}^{K\times 1}$ is the identity parameter, $\boldsymbol{A}_{exp} \in \mathbb{R}^{3N\times L}$ is the first L principle components trained on the offset between neutral scans and expression scans and $\boldsymbol{\alpha}_{exp} \in \mathbb{R}^{M\times 1}$ is the expression parameter.

The texture of 3D face can also be modeled via PCA as:

$$\boldsymbol{t} = \bar{\boldsymbol{t}} + \boldsymbol{A}_{tex}\boldsymbol{\alpha}_{tex}, \tag{2}$$

where $\boldsymbol{t} \in \mathbb{R}^{3N\times 1}$ is a 3D face texture, $\bar{\boldsymbol{t}} \in \mathbb{R}^{3N\times 1}$ is the mean texture, $\boldsymbol{A}_{tex} \in \mathbb{R}^{3N\times M}$ is the first M principle components trained on facial textures and $\boldsymbol{\alpha}_{tex} \in \mathbb{R}^{M\times 1}$ is the texture parameter.

Rendering process The 3D face modeled by 3DMM is projected onto an image plane with weak perspective projection:

$$\boldsymbol{s}_{2D} = f * \boldsymbol{Pr} * \boldsymbol{R} * \boldsymbol{s} + \boldsymbol{t}_{2D}, \tag{3}$$

where $\boldsymbol{s}_{2D} \in \mathbb{R}^{2\times N}$ is the face shape located on the image plane after projection, $\boldsymbol{Pr} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$ is the orthographic projection matrix, $\boldsymbol{R}$ is the rotation matrix constructed from Euler angles (i.e., *pitch*, *yaw*, and *roll*), $\boldsymbol{t}_{2D} = [t_x, t_y]^\intercal$ is the translation vector on the image plane, and f is the scale factor.

Following [17], we assume the face is Lambertian surface and the global illumination is approximated using the spherical harmonics (SH) basis function. The first three bands of SHs are used for the illumination model. $\boldsymbol{\gamma} \in \mathbb{R}^{27\times 1}$ is the illumination parameter for the RGB channels' SH illumination coefficient. Thus, the rendering process depends on the parameter set $\boldsymbol{\chi} = \{\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}, \boldsymbol{\alpha}_{tex}, \mathit{pitch}, \mathit{yaw}, \mathit{roll}, f, \boldsymbol{t}_{2D}, \boldsymbol{\gamma}\}$.

4. Method

We design an encoder-decoder architecture that allows end-to-end semi-supervised adversarial training to extract disentangled semantic representations of a single image, as shown in Figure 2. We adopt inverse rendering technique that utilizes parameterized illumination model and differentiable renderer to render back the input face image under varying identity, expression, pose, and lighting conditions. Our model is trained on hybrid batches of unlabeled face images from CelebA [24] and labeled face images from 300W-LP [44].

4.1. Encoder-decoder network

Encoder As shown in Figure 2, the encoder network is a multi-task learning network, which takes a face image as input and extracts its identity, expression, pose, and lighting representations. A pre-trained ResNet-50 network is used as the backbone of the encoder network. The ResNet-50 network is followed by four branches of fully connected layers with outputs of 128-D identity code ($\boldsymbol{c}_{id}$), 64-D expression code ($\boldsymbol{c}_{exp}$), 6-D pose code ($\boldsymbol{c}_{pose}$), and 27-D lighting code ($\boldsymbol{c}_{lgt}$).

Shape decoder The shape decoder network is a graph convolutional network modified from the COMA architecture [28] with an extra graph convolutional layer and up-sampling layer at the beginning. We concatenate the identity code and expression code extracted from the encoder network to get a 192-D vector as the input of the shape decoder network. The output of the shape decoder is the vertices of the corresponding 3D face shape in the standard position (i.e., without any translations or rotations). We denote as $FC(d)$ a fully connected layer, l the number of vertices after the last down-sampling layer, $GC(k, w)$ a graph convolutional layer with k kernel size and w filters, and $US(p)$ a up-sampling layer by a factor of p, respectively. The shape decoder network is listed as follows: $FC(l * 256) \rightarrow US(2) \rightarrow GC(6, 256) \rightarrow US(4) \rightarrow$

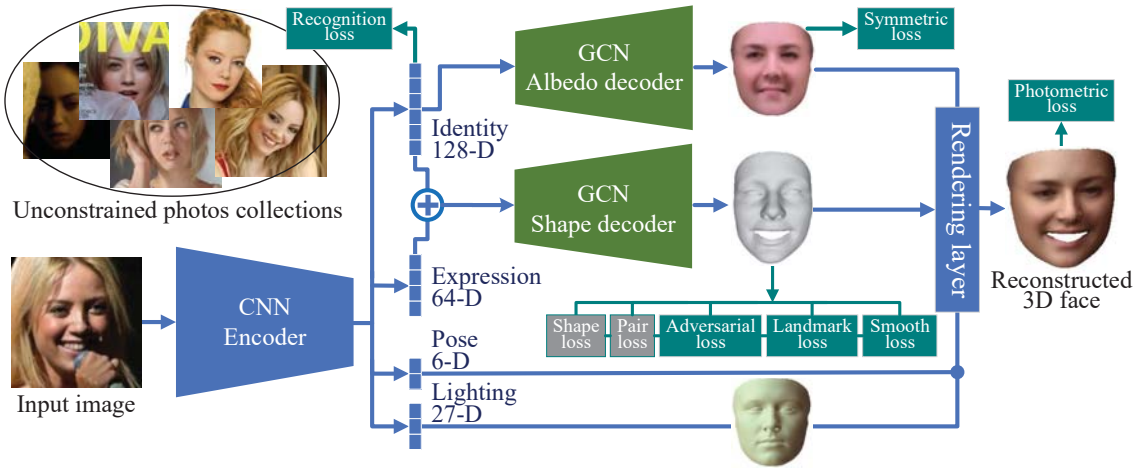


Figure 2: Framework overview. The encoder network takes an input face image and extracts four disentangled representations: identity code ($\boldsymbol{c}_{id}$), expression code ($\boldsymbol{c}_{exp}$), pose code ($\boldsymbol{c}_{pose}$), and lighting code ($\boldsymbol{c}_{lgt}$). The albedo decoder network reconstructs the face albedo from the identity code. The shape decoder network reconstructs the face shape from the combination of the identity code and expression code. The rendering layer takes the face albedo, face shape, pose, and lighting to render back the face image. Multiple losses are applied on our network. Losses in gray rectangles are only used on labeled face images and in green rectangles are used on all face images.

$GC(6,128) \rightarrow US(4) \rightarrow GC(6,64) \rightarrow US(4) \rightarrow GC(6,32) \rightarrow US(4) \rightarrow GC(6,16) \rightarrow GC(6,3)$.

Albedo decoder The albedo decoder network is also a graph convolutional network and has the same architecture as the shape decoder. The albedo decoder takes only the identity code as input since the face albedo should be independent of different expression, pose, and lighting conditions. Furthermore, the face albedo should be consistent across one's different photos with various facial occlusions, such as different hair styles, glasses, microphones, etc. Thus, we apply face segmentation by [26] to eliminate the effect of facial occlusions. Note that, we do not consider aging, injury, or other factors that may affect one's face albedo.

Lighting-independent albedo is critical when using inverse rendering to improve 3D face shape reconstruction. If the learned albedo is entangled with lighting and shadow, the albedo may not well align with the 3D face shape. After the lighting representation is learned, we change the GCN-based albedo decoder network to a CNN network that has skip connections with the encoder network to improve the details of facial texture. The architecture of the encoder and CNN-based albedo decoder with skip connections is similar to U-Net [30]

4.2. Loss functions

Our network is trained with multi-task losses that enable us to regress the 3D face shape and albedo end-to-end. The loss function combines face recognition loss, photometric loss, sparse landmark loss, pairwise shape loss, adversarial loss, and other regularization.

Face recognition loss In order to extract the identity code that only represents the photo's facial identity, we apply face recognition loss as follows:

$$L_{recog} = L_{soft} + \lambda_{center} L_{center}, \tag{4}$$

where L_{soft} is the softmax loss that classifies each photo to a specific identity class, L_{center} is the center loss to improve the discriminative power of the deeply learned identity code [40], and λ_{center} is used for balancing the two loss functions. Face recognition loss is essential to learn the facial identity without being influenced by other factors such as facial expressions, poses, lightings, occlusions, etc.

Photometric loss The rendering layer renders back an image to compare with the input image. The photometric loss is formulated as

$$L_{photo} = \boldsymbol{M} \odot (\|\hat{\boldsymbol{I}} - \boldsymbol{I}\|_2^2 + L_{gdl}), \tag{5}$$

where $\odot$ is the element-wise Hadamard product, $\boldsymbol{I}$ is the input image, $\hat{\boldsymbol{I}}$ is the rendered image, and $\boldsymbol{M}$ is the mask obtained by [26] to eliminate the effect of facial occlusions such as hair, glasses, and microphone. Moreover, a gradient difference loss (GDL) [25], denoted as L_{gdl}, is applied to recover more details in the albedo reconstruction.

Sparse landmark loss We add sparse landmark loss to help learn the face pose and achieve better face reconstruction. The sparse landmark loss is defined as

$$L_{lmk} = \|\hat{\boldsymbol{s}}_{2D}[:, \mathcal{L}] - \boldsymbol{U}\|_2^2 + L_{gdl,lmk}, \tag{6}$$

where $\hat{\boldsymbol{s}}_{2D}$ is the projected face shape from our network, $\mathcal{L}$ is the vertex indexes of the 68 landmarks in the 3D face shape, $\boldsymbol{U}$ is considered as the ground truth of the corresponding sparse 2D landmarks on the input image and is obtained by [8]. The idea of GDL is also applied on the sparse landmarks, denoted as $L_{gdl,lmk}$, which describes the distance of two different landmarks should be close to the corresponding distance in the ground truth. Especially, it is important for the distances from the upper eyelids to the lower eyelids and the upper lip to the lower lip that represent the conditions of eye's opening and mouth's opening, respectively.

Shape loss In order to prevent the network from either generating unrealistic 3D face shapes or being under the constrain of a linear 3DMM, we train our network in a semi-supervised manner on hybrid batches of unlabeled and labeled face images. For the labeled face images, we choose 300W-LP dataset that contains 122,450 images with fitted 3DMM shapes across large poses and was created by [44] with face profiling technique, while we exclude half of the dataset that are horizontally flipped images. The BFM template that has 53,215 vertices is used for the fitted 3DMM shapes. The 3DMM parameters $\boldsymbol{\alpha}_{exp}$ and $\boldsymbol{\alpha}_{exp}$ are provided to calculate each of the fitted 3DMM shapes, as presented in Eq. (1). In this paper, we remove the neck and ears of the BFM model to create our own face shape template with 37,202 vertices. The shape loss for the 300W-LP dataset is formulated as

$$L_{shp} = \|\hat{\boldsymbol{s}} - \boldsymbol{s}[:, \mathcal{T}]\|_1, \tag{7}$$

where $\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{A}_{id}\boldsymbol{\alpha}_{id} + \boldsymbol{A}_{exp}\boldsymbol{\alpha}_{exp}$ is considered as the ground truth of the face shape, $\hat{\boldsymbol{s}}$ is the 3D face shape reconstructed by our network, and $\mathcal{T}$ is the vertex indexes of our face template in the BFM model.

Pairwise shape loss To further disentangle the identity code and expression code, we train the 300W-LP dataset in pairwise manner. Given an input image, the corresponding 3DMM parameters $\boldsymbol{\alpha}_{exp}$ and $\boldsymbol{\alpha}_{exp}$ are provided. For a pair of input images, $\boldsymbol{I}_A$ and $\boldsymbol{I}_B$, we interchange the expression parameters $\boldsymbol{\alpha}_{exp,A}$ and $\boldsymbol{\alpha}_{exp,B}$ to get the 3D face shape of A's identity with B's expression. The pairwise shape loss for the 300W-LP dataset is expressed as

$$L_{pair} = \|f_{shape}([\boldsymbol{c}_{id,A}, \boldsymbol{c}_{exp,B}]) - \boldsymbol{s}_{A,B}[:, \mathcal{T}]\|_1, \tag{8}$$

where $f_{shape}(\cdot)$ is the shape decoder, $[\boldsymbol{c}_{id,A}, \boldsymbol{c}_{exp,B}]$ means concatenation of A's identity code and B's expression code from the encoder network, and $\boldsymbol{s}_{A,B} = \bar{\boldsymbol{s}} + \boldsymbol{A}_{id}\boldsymbol{\alpha}_{id,A} + \boldsymbol{A}_{exp}\boldsymbol{\alpha}_{exp,B}$ is the 3DMM shape of A's identity parameter with B's expression parameter.

Shape smooth loss Laplacian regularization is used on the shape vertex to help remove undesired noise of 3D face shapes. Conventional Laplacian smoothing assumes all the vertices satisfy the equation $\boldsymbol{X}_i = \frac{1}{|\mathcal{M}_i|}\sum_{j\in\mathcal{M}_i}\boldsymbol{X}_j$, where $\boldsymbol{X}_i$ is the ith vertex and $\mathcal{M}_i$ is the vertex indexes of the first order neighbors of $\boldsymbol{X}_i$. However, some vertices, like on the edges, in the nostrils, at the eye corners, etc, do not satisfy the Laplacian equation. In this paper, we propose a novel shape smooth loss that calculates the difference of each vertex with the mean of its first order neighbors bo be close to the corresponding difference of the shape template,

$$L_{smth} = \sum_{i\in\mathcal{N}} |(\hat{\boldsymbol{s}}_i - \frac{1}{|\mathcal{M}_i|}\sum_{j\in\mathcal{M}_i}\hat{\boldsymbol{s}}_j) - (\tilde{\boldsymbol{s}}_i - \frac{1}{|\mathcal{M}_i|}\sum_{j\in\mathcal{M}_i}\tilde{\boldsymbol{s}}_j)|, \tag{9}$$

where $\tilde{\boldsymbol{s}}$ is our face shape template cropped from the BFM model.

Albedo symmetry loss Facial symmetry is a strong prior for face albedo learning, which helps to disentangle facial expression, lighting, and occlusions from the face albedo. The albedo symmetry loss is defined as

$$L_{symm} = \|\boldsymbol{A} - flip(\boldsymbol{A})\|_1, \tag{10}$$

where $\boldsymbol{A}$ is the output face albedo of the GCN-based albedo decoder and $flip(\cdot)$ is an operation of flipping face albedos left and right.

Adversarial loss Semi-supervised learning is not sufficient to generate realistic 3D face shape for the unlabeled face images. Following the idea of generative adversarial network (GAN), an adversarial loss is used to train the encoder-decoder network and a discriminator network alternatively based on WGAN-div [42]. The discriminator network D is a GCN-based encoder network and is used to discriminate the fake shapes (i.e., shapes reconstructed from our network) and real shapes (i.e., shapes sampled from the linear 3DMM), so that the reconstructed face shapes will not be too far away from the distribution of the linear 3DMM. The min-max optimization problem can be written as

$$\min_G \max_D \mathop{\mathbb{E}}_{\hat{\boldsymbol{s}}\sim\mathbb{P}_g}[D(\hat{\boldsymbol{s}})] - \mathop{\mathbb{E}}_{\boldsymbol{s}[:,\mathcal{T}]\sim\mathbb{P}_r}[D(\boldsymbol{s}[:,\mathcal{T}])] - k\mathop{\mathbb{E}}_{\dot{\boldsymbol{s}}\sim\mathbb{P}_u}[\nabla_{\dot{\boldsymbol{s}}}\|D(\dot{\boldsymbol{s}})\|^p] \tag{11}$$

where $L_{adv} = -D(\hat{\boldsymbol{s}})$ is the adversarial loss, $\hat{\boldsymbol{s}}$, $\boldsymbol{s}[:,\mathcal{T}]$ are the fake and real face shapes satisfying the probability measures $\mathbb{P}_g$, $\mathbb{P}_r$, and $\mathbb{P}_u$ is the distribution obtained by sampling uniformly along straight lines between points from the real and fake face shape distributions.

5. Experiments

We train our model on hybrid batches of unlabeled face images from CelebA dataset [24], totally 10,176 identities and 200,405 face images after removing some low quality images and labeled face images from 300W-LP dataset [44], totally 3,837 identities and 61,225 face images (horizontal flipped images in the dataset are not included). The training images are augmented on the fly with random horizontal flip and random scaling of [0.8,1.0].

The whole model is trained in two steps. First, we train the model with GCN-based albedo decoder for 100 epochs.

	Method	Cooperative Mean	Cooperative Std.	Indoor Mean	Indoor Std.	Outdoor Mean	Outdoor Std.
Average over identity	Tran17 [37]	1.93	0.27	2.02	0.25	1.86	0.23
	Genova18 [15]	1.50	0.13	1.50	0.11	1.48	0.11
	Ours	**1.17**	**0.27**	**1.20**	**0.28**	**1.21**	**0.28**
Average over three-view	Tran17 [37]	1.40	0.29	1.38	0.32	–	–
	Tewari17 [31]	1.37	0.32	1.29	0.27	–	–
	Genova18 [15]	1.37	0.35	1.26	0.31	–	–
	MVF-Net [41]	1.22	0.25	1.23	0.24	–	–
	Ours	**1.16**	**0.29**	**1.24**	**0.30**	**1.22**	**0.28**

Table 1: Mean error comparison on the MICC dataset. Note that MVF-Net [41] is a multi-view 3D face reconstruction method.

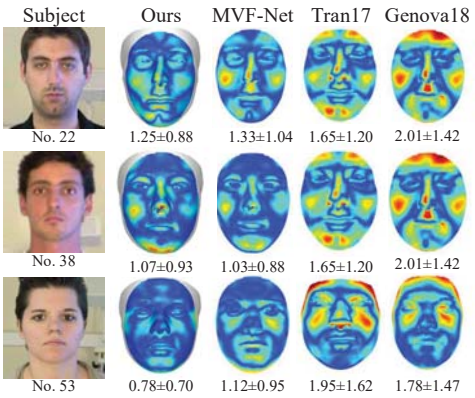


Figure 3: Error map comparison of the examples used in MVF-Net [41].

Then, we replace the GCN-based albedo decoder with the CNN-based albedo decoder and train for another 100 epochs to improve the details of facial texture. The encoder-decoder networks and discriminator network are optimized using Adam optimizer with a learning rate of 0.0001 and RMSprop optimizer with a learning rate of 0.00005, respectively. MICC Florence dataset [1] and AFLW2000-3D dataset [44] are selected for the quantitative and qualitative evaluations. The face region of the BFM model is cropped as the 3D face mesh template (i.e., 37202 out of the 53215 vertices).

5.1. Comparisons to the state-of-the-art

We evaluate our model quantitatively on the MICC Florence dataset [1], which contains the ground truth scans of 53 subjects in neutral expressions. Each subject is recorded in three videos: *Cooperative*, *Indoor*, and *Outdoor* with increasingly challenging conditions. We use the same evaluation metric in [15] — the face region of 95mm around the nose tip of the ground truth scan is cropped to calculate the point-to-plane L2 errors with the predicted face shape. Two evaluation methods are used in this paper. First, as described in [15], we run our method on each frame of the videos to extract each frame's facial identity, average the identity codes over each video, and combine with the neutral expression code to obtain a single reconstruction for each video, called 'Average over identity'. Second, we use the same setting in [41]. The left, frontal, and right view of each subject are selected from the *Cooperative* and *Indoor* videos. The predicted 3D face shape is obtained by averaging over the three reconstructed 3D face shapes, called 'average over three-view'.

Table 1 shows that the proposed method outperforms other single-view reconstruction methods. For the evaluation of 'average over identity', our method is stable among the conditions of *Cooperative*, *Indoor*, and *Outdoor*, which means the identities extract from the frames of these three

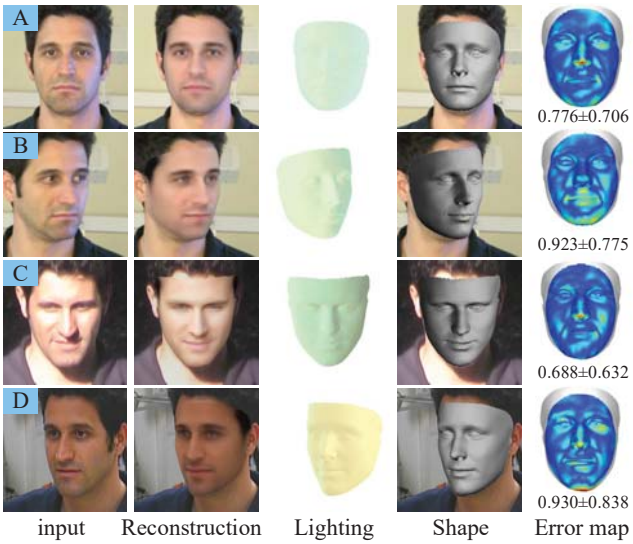


Figure 4: Examples with different lightings and poses of subject No. 05 from the MICC dataset. A and B are from the video of *Cooperative*. C and D are from the videos of *Outdoor* and *Indoor*, respectively.

videos are very close and our model is robust to facial identity recognition. The variance of our method is relative higher because a few of the ground truth face shapes were scanned not in a neutral expression, e.g., subject No. 10 was scanned with a smile expression and the error is 1.89 in the *Cooperative* condition. Compared to the multi-view reconstruction method (MVF-Net) [41], we achieve better results in the *Cooperative* condition and have slightly worse results in the *Indoor* condition. Figure 3 presents three examples (i.e., subject No. 22, No. 38, and No. 53) of detailed error maps. Figure 4 shows the reconstruction results of face images from the same subject (No. 05) in the *Cooperative*, *Indoor*, and *Outdoor* videos with different lightings and poses. The reconstruction errors are small across dif-

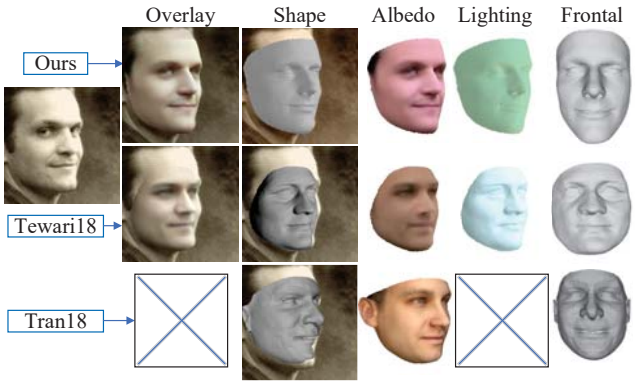


Figure 5: 3D reconstruction comparisons with [32] and [38]

ferent conditions.

We further evaluate our model qualitatively on the AFLW2000-3D datasets [44]. Both Tewari *et al.* [32] and Tran *et al.* [38] proposed two-stage models: a coarse-scale linear model and a fine-scale corrective model. Even though the fine-scale corrective model is able to add more details on top of the linear model, the reconstructed face shape will fail when the foundation face shape generated in the first stage is not good. The foundation face shape is restricted by the linear 3DMM and cannot generalize well in the wild conditions with true diversity of poses, expressions, lightings, and occlusions. As shown in Fig. 5, the face shape reconstructed by our model aligns better with the input face image and looks more realistic from the frontal view. Moreover, the proposed method can reconstruct the facial texture in more detail.

Tran *et al.* [34] proposed a nonlinear 3DMM and is the most related work. The face shape and albedo are reconstructed from CNN-based decoders and have higher representation power compared to a linear 3DMM. However, the model was trained on the 300W-LP dataset generated based on a linear 3DMM. Even with higher representation power, the nonlinear model is limited to fit the 300W-LP dataset. Moreover, the identity and expression of face shape are entangled, resulting in poor performance on face images with diverse expressions. As shown in Figure 6, the face shapes reconstructed by [34] tend to have smaller mouth opening and some artifacts are introduced to the face shapes and textures in challenging conditions. The proposed model achieves better performance across various conditions: exaggerated expressions, large poses, diverse lighting, different ethnic groups, and different occlusions as presented in the figures.

5.2. Ablation study

Shape reconstruction We study the effects of face recognition loss, adversarial loss, and the proposed shape smooth loss on the quality of 3D face shape reconstruction.

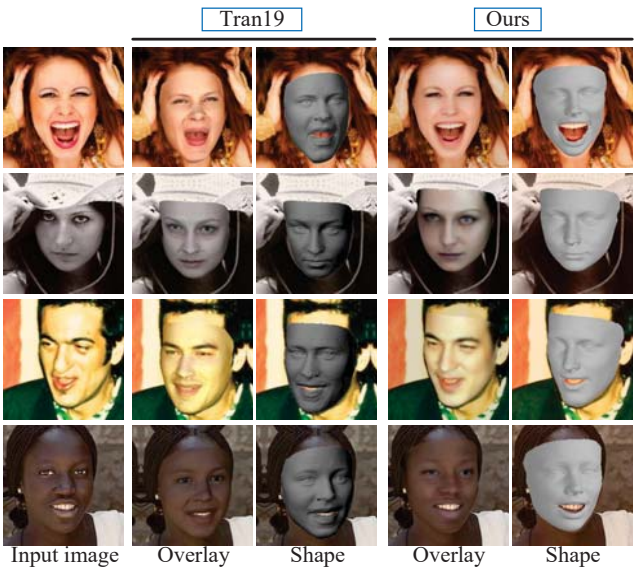


Figure 6: 3D reconstruction comparisons with [34].

Method	Cooperative		Indoor		Outdoor	
	Mean	Std.	Mean	Std.	Mean	Std.
w/o L_{recog}	1.39	0.73	1.40	0.82	1.54	0.76
w/o L_{Adv}	1.27	0.27	1.23	0.26	1.24	0.26
Full model	**1.17**	**0.27**	**1.20**	**0.28**	**1.21**	**0.28**

Table 2: Shape ablation test of mean error comparison on the MICC dataset using the evaluation method in [15].

Table 2 shows the quantitative results on the MICC dataset without (i.e., w/o) face recognition loss (L_{recog}) and adversarial loss (L_{Adv}) using the evaluation method of 'average over identity'. Figure 7 shows two qualitative results. Without L_{recog} results in much higher reconstruction errors and variances compared with the full model because the facial identity extracted from each frame is not consistent over each video. Without L_{Adv} results in higher reconstruction errors as well. Table 2 shows the degradation is not very obvious because only frontal facial images with neutral expressions are tested on the MICC dataset. The degradation is more severe for images with large poses and expressions, where the reconstructed face's eyebrows extrude out and two sides are shrunk, as shown in Figure 7.

Since our face model is not constrained by a pre-existing linear 3DMM, the face meshes can potentially be deformed to any shapes. The conventional smoothing loss causes abnormal effects on the edges and nostrils of face shapes. The vertices on the mouth's inner edge distance away from their neighbors. The nostrils are prone to be flat or even sticking out of the nose. This is because the vertices on the edges and nostrils are not satisfied with the Laplacian regularization which forces each vertex locates at the mean of its first order neighbors.

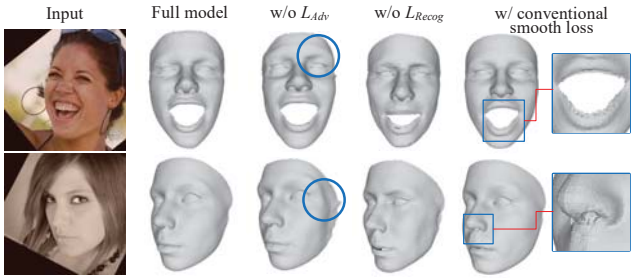


Figure 7: Shape ablation test showing failures when without adversarial loss, face recognition loss, and with conventional shape smooth loss.

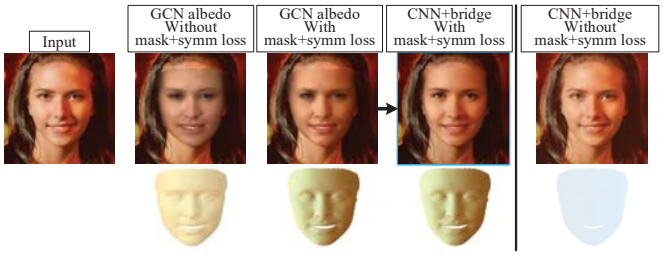
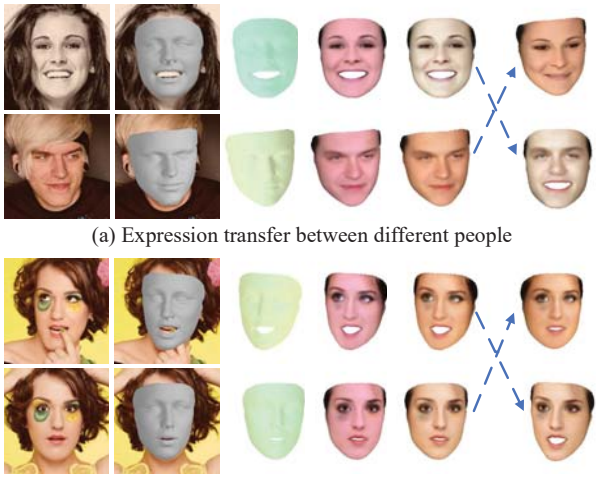


Figure 8: Texture ablation test showing failures of lighting caused when without facial mask and albedo symmetric loss (i.e., w/o *mask+symm loss*). We denote GCN-based albedo decoder as *GCN albedo*, and CNN-based albedo decoder with skip connections as *CNN+bridge*.

Texture reconstruction Figure 8 shows the effects of different albedo decoders, the facial mask, and the albedo symmetric loss. As shown in the last column, the reconstructed face texture is almost identical with the original face image, which means CNN-based albedo decoder with skip connections to the encoder has very high representation power for face texture reconstruction.

The facial mask and albedo symmetric loss are crucial for lighting representation learning so that shade, lighting, and facial occlusions are not confounded with the facial albedo. When without applying facial mask and albedo symmetric loss, especially if the representation power of the albedo decoder is high (e.g., using the CNN-based albedo decoder), the model may fail to learn the lighting even though the generated texture looks very close to the input image, as shown in the last column of Figure 8. As a result, reconstructing high fidelity texture makes limited contributions to the face shape reconstruction because the high fidelity texture may not align with the face shape and looks odd when viewing from a different pose. The facial mask, albedo symmetric loss, and the GCN-based albedo decoder that is solely determined by the identity code help disentangle the albedo from lighting and facial occlusions. Once the lighting is learned, the CNN-based albedo decoder with skip connections to the encoder is used to improve the detail of facial albedo.



(a) Expression transfer between different people

(b) Expression transfer between the same person

Figure 9: Expression transfer between different face images. The left side is the expression transfer between different people and right side is the expression transfer between the same persian.

5.3. Applications

Disentangled representations of our model not only can improve the performance of face reconstruction, but also can facilitate many facial editing applications, such as face recognition, face puppetry, face replacement, face reenactment, expression transfer, and so forth. Figure 9 demonstrates the function of expression transfer between different face images. We keep the face image's identity representation and replace the pose, lighting, and expression representations from another face image to generate a realistic new face image with the same identity but another face's pose, lighting, and expression. When we apply the expression transfer on different images of the same person, the results are consistent after the expression transfer, demonstrating high robustness of our model.

6. Conclusion

This paper proposes an encoder-decoder architecture to reconstruct 3D face from a single image with disentangled representations: identity, expression, pose, and lighting. We develop an effective semi-supervised training scheme to fully exploit the value of large amount of unlabeled face images from unconstrained photo collections. An adversarial loss is applied to prevent our model from generating unrealistic 3D faces. We evaluate our model quantitatively and qualitatively. Our model outperforms state-of-the-art single-view reconstruction methods and can effectively disentangle identity, expression, pose, and lighting features.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (61901259) and China Postdoctoral Science Foundation (BX2019208).

References

[1] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2D/3D hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80. ACM, 2011. 6

[2] 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*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1, 2

[3] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, Sep. 2003. 1, 2

[4] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[5] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou. 3D reconstruction of in-the-wild faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2638–2652, Nov 2018. 1, 2

[6] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3D morphable model learnt from 10,000 faces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, June 2016. 2

[7] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 5

[9] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):43:1–43:10, July 2014. 1

[10] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, March 2014. 2

[11] Baptiste Chu, Sami Romdhani, and Liming Chen. 3D-aided face recognition robust to expression and pose variations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3

[12] Michal Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems 29*, pages 3844–3852. Curran Associates, Inc., 2016. 2

[13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1

[14] 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. *Computer Graphics Forum*, 34(2):193–204, 2015. 1

[15] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 6, 7

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[17] Yudong Guo, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1294–1307, June 2019. 2, 3

[18] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3D face shape. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[19] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[20] Martin D. Levine and Yingfeng Chris Yu. State-of-the-art of 3D facial reconstruction methods for face recognition based on a single 2D training image per person. *Pattern Recognition Letters*, 30(10):908 – 913, 2009. 1, 2

[21] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[22] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, Nov. 2017. 2

[23] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 3, 5

[25] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 4

[26] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and

face perception. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 98–105, May 2018. 4

[27] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, Sep. 2009. 1, 2

[28] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

[29] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469, Oct 2016. 1, 2

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2, 4

[31] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 1, 2, 6

[32] Ayush Tewari, Michael Zollhfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Prez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 7

[33] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2Face: Real-time face capture and reenactment of RGB videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[34] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 7

[35] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[36] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*, 2019. 1, 2

[37] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Grard 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. 6

[38] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Grard Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[39] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 426–433, New York, NY, USA, 2005. ACM. 2

[40] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision - ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing. 2, 4

[41] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. MVF-Net: Multi-view 3D face morphable model regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 6

[42] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. Wasserstein divergence for GANs. In *Computer Vision – ECCV 2018*, pages 673–688, Cham, 2018. Springer International Publishing. 5

[43] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3D-aided dual-agent GANs for unconstrained face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2380–2394, Oct 2019. 1

[44] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3, 5, 6, 7

[45] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, June 2015. 2

[46] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face alignment in full pose range: A 3D total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):78–92, Jan 2019. 1, 2