Sphere Face Model:A 3D Morphable Model with Hypersphere Manifold Latent Space

Diqiong Jiang
Zhejiang University

Yiwei Jin
Zhejiang University

Fang-Lue Zhang
Victoria University
of Wellington

Zhe Zhu
Duke University

Zhang Yun
Communication University
of Zhejiang

Tong, Ruofeng
Zhejiang University

Tang, Min
Zhejiang University

Abstract

3D Morphable Models (3DMMs) are generative models for face shape and appearance. Recent works impose face recognition constraints on 3DMM shape parameters so that the face shapes of the same person remain consistent. However, the shape parameters of traditional 3DMMs satisfy the multivariate Gaussian distribution. In contrast, the identity embeddings meet the hypersphere distribution, and this conflict makes it challenging for face reconstruction models to preserve the faithfulness and the shape consistency simultaneously. In other words, recognition loss and reconstruction loss can't decrease jointly due to their conflict distribution. To address this issue, we propose the Sphere Face Model(SFM), a novel 3DMM for monocular face reconstruction, preserving both shape fidelity and identity consistency. The core of our SFM is the basis matrix which can be used to reconstruct 3D face shapes, and the basic matrix is learned by adopting a two-stage training approach where 3D and 2D training data are used in the first and second stages, respectively. We design a novel loss to resolve the distribution mismatch, enforcing the shape parameters have the hyperspherical distribution. Our model accepts 2D and 3D data for constructing the sphere face models. Extensive experiments show that SFM has high representation ability and clustering performance in its shape parameter space. Moreover, it produces high-fidelity face shapes consistently in challenging conditions in monocular face reconstruction. The code will be released at `https://github.com/a686432/SIR`.

1. Introduction

The problem of face reconstruction from stills and videos has been attracting considerable attention in the computer vision and computer graphics community. It has a broad range of applications, including AR/VR [14], animation [28, 8], computer games [32], etc. In recent years, there is a growing demand for customizing 3D virtual faces to create game characters [32, 50] or personalized 3D facial editing [66]. In such applications, images from common users usually come from a large diversity of conditions, including occlusion, resolution, pose, expression, illumination, etc. It is thus challenging to reconstruct a face from only a single image requiring both shape faithfulness and identity preservation.

Although previous works [70, 26] claimed to have achieved face reconstruction from a single image, their reconstructed face shapes suffer from inconsistent identity properties when the input images have varying conditions. To address this problem, the follow-up works [44, 57, 33] propose to aggregate shape parameters of the same identity while separate those of different subjects to produce 3D face shapes containing good identity-related features. However, the conflict between the shape loss and the identity loss in their reconstruction pipeline prevents them from achieving both shape fidelity and identity consistency. That conflict comes from the mismatch between the distribution of identity embeddings of face recognition and shape parameters of the previous 3DMMs [39, 30, 21, 11], which maximize their model expression ability while neglecting some distinguishable information of categories.

Therefore, this paper focuses on identity-consistent face reconstruction in a linear model. Before introducing our method, we first introduce the terminologies as well as several key concepts: geometric space, shape parameter space,

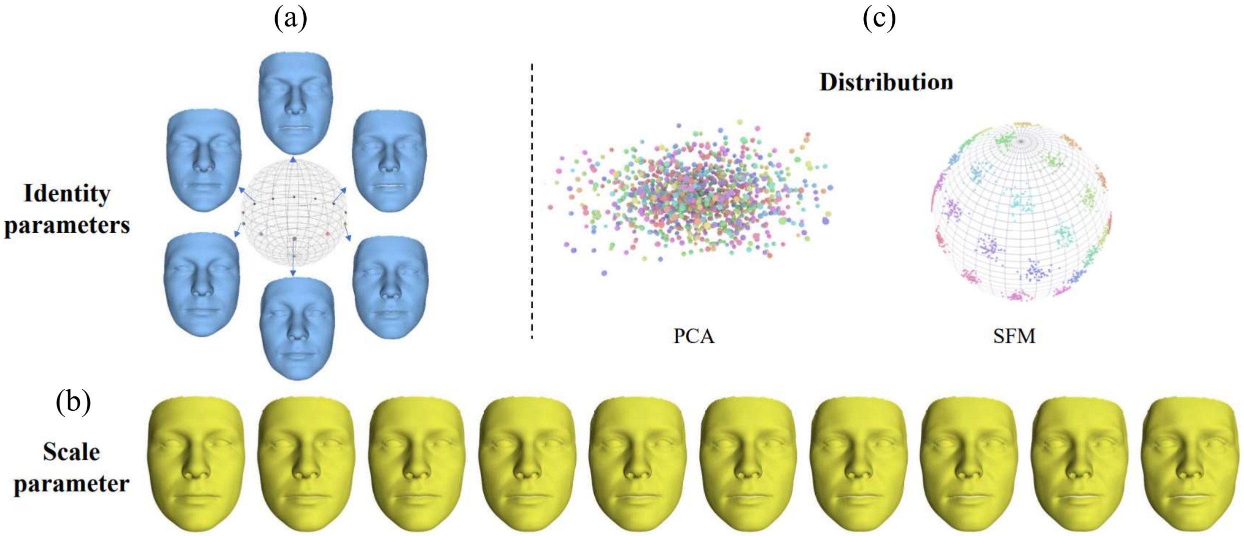

Figure 1. The overview of the sphere face model. (a) The identity parameter of the sphere face model distributed on the hypersphere represents the identity information. The meshes are uniformly sampled from on the hypersphere using the first two dimensions of identity parameters. (b)The scale parameter of the sphere face model is scalar, which controls the distinctiveness to the mean face. (c)The distribution of the parameter. The shape parameter of the PCA-based model has an anisotropic multivariate Gaussian distribution. Our identity parameters are distributed isotropically on the hypersphere and separated between different classes.

and identity latent space. Geometric space Ψ is a set of face meshes, which is formulated as $\Psi \in \mathbb{R}^{N_v}$. N_v is the number of vertices of a face mesh. Shape parameter space Φ is a set of shape parameters of 3DMM, which is formulated as $\Phi \in \mathbb{R}^{N_p}$. N_p is the dimension of shape parameters. Identity latent space Ω is a set of identity embedding which is formulated as $\Omega \in \mathbb{R}^{N_i}$. N_i is the dimension of identity embedding. To resolve the aforementioned distribution mismatch problem, we propose a novel face generation model called the Sphere Face Model (SFM). We add category information while building the basis of SFM and constrain identity parameters over a hypersphere by normalizing the shape parameters to make the shape parameter space of SFM consistent with identity latent space. In this way, we resolve the conflict between the two losses and further improve the identifiability of 3D face models. Moreover, SFM has an essential property that the discrimination of the parameters is transferable to the geometry, which means the Euclidean distance between two sets of 3DMM parameters in shape parameter space and between corresponding mesh vertices in geometric space have a positive correlation. One notable challenge is when the identity parameters are forced to be distributed over a hyperspherical surface, the L2 norm of the parameter vectors become the same. In other words, the reconstructed faces would have the same root mean square errors from the mean face, leading to reduced varieties of generated faces. We use two approaches to address that issue. Algorithmically we add a parameter to control the scale of the shape parameters of each face. While previous approaches mainly use 3D training data, which are limited, we propose a two-stage training approach where we use 3D data only for pre-training and adopt an unsupervised learning approach that can leverage a sufficient amount of 2D face data. Figure 1 highlights the differences between our face model and the previous 3DMMs. The parameter of SFM is composed of a shape parameter and a scale parameter. The identity parameter is the normalized shape parameter, which controls the face's identity attribute. It is distributed on the hypersphere with good separation properties. The scale parameter controls the distance to the average face.

The main contributions of this paper lie in the following three aspects:

- We propose Sphere Face Model (SFM) for 3D face reconstruction from single images with both shape faithfulness and identity consistency.
- We propose a new structure of 3DMMs, where the shape parameter space follows a hyperspherical distribution and the discrimination of shape parameter space is transferable to the geometric space.
- To enable SFMs to reconstruct high-quality 3D face models from single images, we present a learning scheme to train SFMs with both 2D and 3D data.

2. Related Work

3D morphable models map the high-dimensional face geometry space to the low-dimensional manifold space. Based on 3DMMs, the previous works optimize the low-dimensional 3DMM parameters from the input image to reconstruct high-dimensional face geometries in monocular face reconstruction. Meanwhile, many works introduce identity loss in the face reconstruction pipeline to keep the face shape stable from the various input images. This section introduces the related works from three aspects: 3D morphable model, shape-consistent face reconstruction from monocular images, and deep face recognition.

3D morphable models 3D morphable model is a statistical model of the distribution of the faces, which maps the low dimensional parameter vector to the high dimensional graphic vertices. The groundbreaking work of 3DMMs traces back to Blanz, and Vetter [10], who propose the 3D morphable model using principal component analysis from an example of 200 3D faces. Based on this idea, Paysan et al. [39] provide the first public 3DMM model, BFM 2009 and others [9, 54, 2, 29, 12, 59] extend the model to introduction emotive facial shapes information by adopting an additional expression basis or using bilinear and multilinear. [30] provides the whole head model, FLAME, which introduces an articulated jaw, neck, and eyeballs in linear shape space and global expression to make the model more expressive. Yang et al. [65] present a large-scale detailed 3D face dataset and models the variation of detailed geometry with it. Unlike the previous work, we consider identity information while constructing the 3DMM model, and the shape parameter can be inherently separated among each identity. Blanz and Vetter [10] only use facial meshes of 200 subjects of similar ethnicity and age, which cannot represent the great diversity of the human faces. [11] train the 3DMM with the large scale of 3d data to overcome this limitation, but the 3D data is also limited. [56, 53, 55] use sufficient 2D data to training the 3DMM. However, training with 2D data without 3D prior needs strong regular terms, which leads to a lack of geometric details and diversity. Our method training the model make full of 2D and 3D data. In recent years, with the development of deep learning, [56, 55, 4] propose nonlinear models with encoder-decoder structure. Those nonlinear models do not consider the parameter separation and the property of propagating the discrimination from shape parameter space to geometric space when training the models.

Shape consistence monocular face reconstruction Early works [1, 46, 6, 40, 47], reconstruct 3d face from monocular RGB using the analysis-by-synthesis approach with the prior knowledge of the 3DMM. They often apply the photometric and landmark consistency between the input and the rendered image. In recent years, many researches [22, 44, 17] have proposed the deep network to regress the 3DMM parameters. Applying face recognition loss to the rendered image mainly affects the recognizability of the texture, which has a relatively small impact on shape consistency reconstruction. Adversarial loss, perceptual loss, and identity loss on the rendered image [31, 20, 55, 15] are proposed to generate the high fidelity texture. However, applying face recognition loss to the rendered image mainly affects the recognizability of the texture, which has a relatively small impact on shape consistency reconstruction. Feng et al. [18] replace the shape parameter of the same person and employ the photometric and identity loss on the rendered images. However, it fails to distinguish shape parameters of different people. To reconstruct the stable face shape geometry, Tran et al. [57] label a large number of face images with 3DMM shape parameters using the optimization method, and utilize the deep CNN to learn the mapping from images to shape parameters. But its performance depends on the accuracy of the optimization method. Liu et al. [33], and Sanyal et al. [44] use a face recognition loss to push away the shape parameters of different people while aggregating those of the same person. Jiang et al. [24] propose that simply applying the face recognition loss function to the shape parameter does not guarantee shape consistency. They explore the relationship of shape parameter discrimination and geometric visual discrimination and propose the SIR loss, which increases discriminability in both the shape parameter and shape geometry domain. Since they use the PCA-based face model, it is challenging to preserve faithfulness and shape consistency simultaneously.

Deep face recognition In recent years, many works have achieved incredible face recognition accuracy with the powerful deep convolutions neural network. Most of them focus on cleaning and mining the training data or designing the loss function to maximize the intraclass distance and minimize the interclass distance, which boosts the discrimination of deep feature identity embedding. There are mainly three types of loss functions for face recognition. One utilizes pair or triple training strategy, such as contrastive loss [52] and the Triplet loss [48]. Another type of loss, like the center loss [63], plays as the auxiliary loss to augment the other loss functions. The aim of these loss functions is aggregating features to minimize the inner-class distance. The auxiliary loss can be directly added to the classifier network and learn the discriminative features. The last type of loss is modified softmax [36, 61, 37, 35, 16, 62]. Normface [61] and Cocoloss [37] normalized the weights and features and directly optimize the cosine similarity instead of the inner product. L-softmax [36] and sphereface [35] introduce the multiplicative cosine margin. Cosface [62] and Am-softmax [60] introduce the additive cosine margin, and arcface [16] introduces the additive angular margin. [23, 67, 34] adapt the margin during the training. Current SOTA deep face recognition methods mostly adopt the last type of loss and softmax-based classification loss. Their identity latent space is the hypersphere.

3. Space Distribution

This section elaborates the characteristics the identity latent space needs to have for effective face representation and reconstruction.

Previous 3DMM-based works suffer from the conflict between the losses for face recognition and reconstruction in the shape-consistent face reconstruction pipeline. Taking PCA-based face models as an example, the shape parameters for face reconstruction satisfy the anisotropic multivariate Gaussian distribution [10].

$$p(\alpha) \sim N(0, \Sigma) \tag{1}$$

where α is shape parameter, $\Sigma = \{e_1, e_2, ..., e_n\}$, and the e_i is the ith eigenvalue of shape basis. In contrast, the identity embeddings for face recognition are distributed isotropically on the hypersphere [61].

$$p(\beta) \sim x/||x||_2 \, ; x \sim N(0, 1) \tag{2}$$

where β is identity embedding. The distribution mismatch in the shape parameter space of face reconstruction and identity latent space of face recognition makes the co-convergence of these two loss functions (face recognition loss and face reconstruction loss) very difficult to achieve. More specifically, when conducting the intense face recognition loss, the latent vectors are forced to distribute on a hyper-spherical surface which do not follow the actual distribution of shape parameters and make the reconstruction results inaccurate. On the contrary employing an intense reconstruction loss would probably make the distribution of latent vector to be no longer hyperspherical, resulting in less identity-consistent reconstruction results. Note that nonlinear face models [56, 55, 4], which also belongs to the family of 3DMMs, are not guaranteed to transfer the discrimination of the shape parameter space to the geometric space as explained in Section **??** thus cannot preserve identity information while constructing face models.

To address the above issue, we propose to keep the shape parameter space of SFMs consistent with identity latent space of face recognition. Additionally, it should meet the requirement that discriminability can be transferred between the shape parameter space and the geometric space. Here, we first introduce the identity latent space distribution of identity embeddings and then describe how we design the structure of SFMs and the concrete constraints the SFM should satisfy.

3.1. Hypersphere Manifold of Identity Embedding

Modern face recognition works always adopt the softmax-based classification loss for metric learning, where weights w and identity embeddings l are normalized and the concept of margin [62, 35, 16] is adopted to boost discrimination of deep face features further. In particular, a loss function with margin can be formulated as Equation 3:

$$\begin{aligned} &L_m = \\ &-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{\|x_i\|cos(\alpha\theta_{y_i}+\beta)-\gamma}}{e^{\|x_i\|cos(\alpha\theta_{y_i}+\beta)-\gamma}+\sum_{j=1,j\neq y_i}^{n}e^{\|x_i\|cos(\theta_j)}} \\ &where \qquad cos(\theta_i) = w_i l_i \end{aligned} \tag{3}$$

where $x_i \in \mathbb{R}^d$ denotes the d dimensional deep feature of i-th sample, y_i denotes the label of x_i. w_i is ith column the normalized weight before Softmax[61]. θ_j is the angle between vector x_i and class vector w_j in the identity latent space. m and n denote the batch size and the class number respectively. The parameter α, β and γ encode the margins of different kinds (see SphereFace [35], Cosface [62] and Arcface [16]). The identity embedding trained with softmax-based classification are distributed on a hypersphere. Previous works [33, 24] impose the softmax-based loss on shape parameters. However, the face parameters constrained by the face recognition loss function will make the face parameters tend to have a hyperspherical distribution. On the other hand, these parameters must meet the distribution of PCA-based basis(the anisotropic multivariate Gaussian distribution) to have a better result of face shape reconstruction. Therefore, for the face recognition function to better affect the geometric separation, we must reconstruct and establish a reconstruction base with a similar distribution to identity embedding. However previous works [30, 55] on conducting the shape basis did not emphasize this.

3.2. Shape Parameter Space of Sphere Face Models

As mentioned above, the established SFM should meet the following criteria: (1) the discriminability of the shape parameter space can be transferred to the discriminability of the geometric space; (2) the distribution of the SFM parameters is consistent with the distribution of the face recognition identity embeddings, that is, the isotropic hyperspherical distribution. For the first criteria, SFM shape parameter space have to meet the following conditions:

$$\begin{aligned} &\forall x_1, x_2, x_3; if \; \|x_1 - x_2\| \le \|x_1 - x_3\| \\ &\qquad then \; \|f(x_1) - f(x_2)\| \le \|f(x_1) - f(x_3)\| \end{aligned} \tag{4}$$

If $f(x)$ is a linear function and the basis (mentioned later in Section 4) is orthonormal, the above condition can be met (The property is proved in [24]). Thus we use orthonormal basis in SFM.

To meet the second criteria, we normalize the shape parameters in SFM. As a consequence, the vector of shape

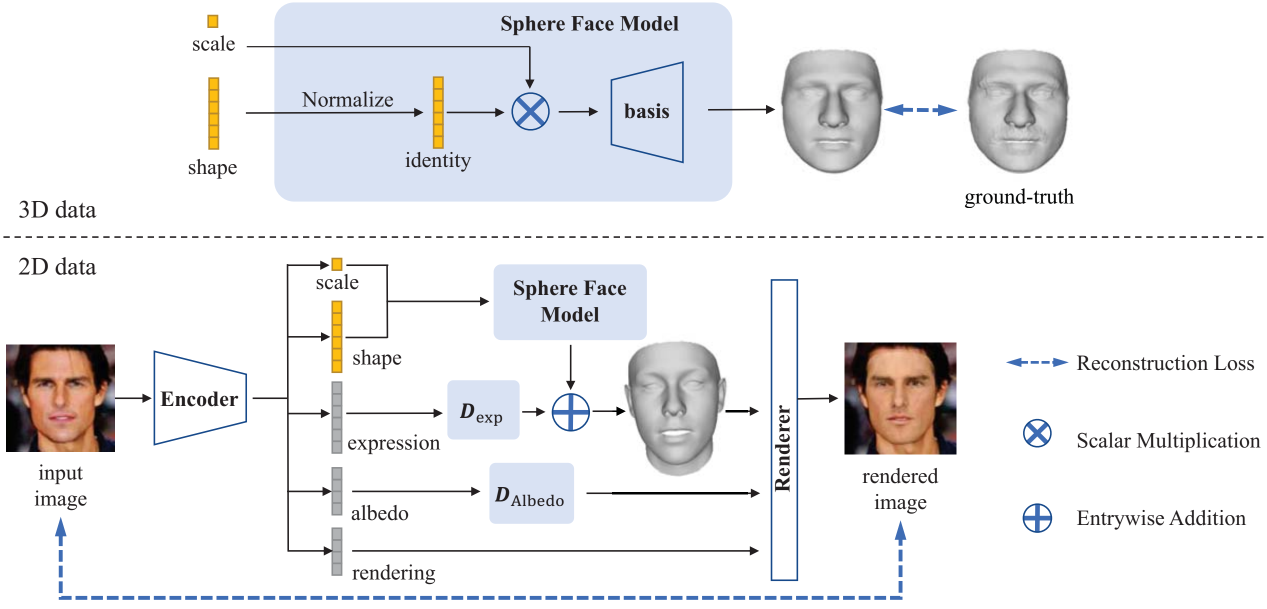

Figure 2. **The framework of our method.** The normalization of x generates the identity parameters distributing on a hypersphere. The normalized identity parameter is multiplied by the scale parameter to get the shape parameter and goes through the basis to get the corresponding mesh. When training on 3D data, we directly optimize s and x. When training on 2D data, we use encoder-decoder because it requires other parameters to render the image.

parameters will be constrained on the hypersphere, leading to the cosine angle between two vectors proportional to their distance in the geometric space. This also brings up a problem that the distance between the result of all human faces and the average face become the same, since all human faces would have the same distance from the origin of the coordinates. Our solution is to add a scalar to control the norm of the face parameters. Similar as [38], we use scale-normalized shape parameters, namely identity parameters, since they are related to identity information. The scale parameter represents the difference with the mean face. Previous work [25] also proposes decomposition networks, but their model did not consider the above situations, making it impossible to use face recognition loss on shape parameters to improve the degree of parameter separation further.

To summarize, our SFM consists of a scale parameter s and a vector of shape parameters $\mathbf{x}$ to describe a face model.

4. Sphere Face Model

Given the shape parameters $\mathbf{x}$ and the scale parameter s, our Sphere Face Model is able to reconstruct the 3D face shape by:

$$\mathbf{M} = \bar{\mathbf{M}} + \mathbf{A}(s * \frac{\mathbf{x}}{\|\mathbf{x}\|}) \tag{5}$$

where $\mathbf{M} \in \mathbb{R}^{3n}$ is a reconstructed 3D face shape with n vertices and $\bar{\mathbf{M}} \in \mathbb{R}^{3n}$ is the mean face shape. The normalized term $\mathbf{x}/\|\mathbf{x}\|$ represents the identity parameters. The orthogonal matrix $\mathbf{A}$ represents the basis of SFM, which is obtained by a 2D-3D joint learning framework based on deep neural networks. This structure guarantees $s * \frac{\mathbf{x}}{\|\mathbf{x}\|}$ located on the hypersphere.

The previous works for constructing parameterized models mainly rely on 2D or 3D datasets. However, only training the model with 3D models would lack face variants because there is no publicly available large 3D face datasets. Training only with a two-dimensional dataset is also difficult to get satisfying results since the large diversity of expressions and poses will affect the identity-related features in the reconstructed face models without 3D shape guidance. The regularization constraint used in these methods [56, 55] also makes the generated mesh similar to with the average face. Tran et al. [55] used the proxy strategy to alleviate that issue but did not fully solve it. Therefore, we propose an effective learning scheme to utilize 2D and 3D data to learn face models with the aforementioned properties.

In the following sub-sections, we introduce the overall framework and then describe how the deep model is trained using 3D and 2D face data.

4.1. Learning Framework

Given the model defined in Equation 5, our goal is to learn the basis matrix $\mathbf{A}$ from face datasets. To achieve so, we adopt a two-stage training strategy as illustrated in Figure 2. In the first stage, we feed the model with scale and shape parameters and force the model to reconstruct the 3D face. We optimize the basis matrix, scale parameters, and shape parameters by minimizing the objective function as shown in Equation 9. After this step, we obtain a basis matrix, which is rough due to the scarcity of the 3D training data. In the second stage, we make use of the large 2D face datasets and train an encoder-decoder style model similar to [55, 53, 68]. The pre-trained SFM can be regarded as a decoder module that can reconstruct the 2D face image along with other decoder modules using the latent vector from the encoder. By optimizing the encoder and decoder, our SFM is finetuned.

More specifically, the encoder regresses the scale, shape, expression, and other rendering parameters, such as albedo, illumination, pose, and camera parameters. In the decoder part, we have four components, each of which is to be trained in this stage: (1) The trained shape basis of SFM, (2) The expression basis D_{exp} from bfm2017 [21], (3)the albedo basis D_{albedo} from [51], (4) the rendering layer takes the geometric, albedo, illumination, pose parameter, and camera parameter and renders 224×224 RGB images, which is based on Pytorch3d [42]. The illumination model is a spherical harmonic illumination model.

In previous works, [55] did not use 3d prior when constructing face models from 2D data; [53] creates a new basis besides the 3DMM to correct face shape; [68] directly regresses the residual displacement in geometric space to correct the face shape. In contrast, our work directly corrects the 3d prior basis by decoupling the expression and appearance information in 2D data, which is able to learn better identity-related features for face reconstruction.

4.2. Data Preparation

3D data. FRGC v2.0 database [41] contains 4007 3D face scans of 466 subjects and is acquired by a Minolta Vivid 900/910 series sensor under controlled illumination conditions. In the preprocessing, we use a non-rigid iterative closest point algorithm [3] to register the 3D face raw scans to the topology of BFM2017 [21] and remove the sample with radical expressions. The registered 3D models face the positive direction of the z-axis, and their centers are coincident with the origin. Note that the unit of the registered 3D model is the millimeter.

2D data. The second stage is trained with 300W-LP [69] and VGGFace2 [13]. VGGface2 contains 3.31 million images of 9131 subjects covering a large range of poses, ages, and ethnicities. 300W-LP is a synthetically generated dataset based on the 300-W database [43] containing 61,255 samples across various poses. In our preprocessing stage, the faces are aligned using similarity transformation and cropped to 224×224 in the RGB format with its landmark of 300W-LP.

4.3. Training Sphere Face Model with 3D Data

SFMs are first trained with 3D data to learn the shape basis using the following loss function:

Loss function. To assemble the identity parameters of the same identity and separate those of different identities in cosine distance, we apply the modified-Softmax loss with normalized shape parameters and normalized weight, which is introduced by the Normface [61]:

$$L_m = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{\frac{\mathbf{x}_{y_i}}{\|\mathbf{x}_{y_i}\|}*\frac{w_{y_i}}{\|w_{y_i}\|}}}{\sum_{j=1}^{n}e^{\frac{\mathbf{x}_j}{\|\mathbf{x}_j\|}*\frac{w_j}{\|w_j\|}}} \tag{6}$$

where n is the number of classes and m is the number of samples of the batch. y_i the groundtruth label. w_j represents the jth row of the basis $\mathbf{A}$. At the same time, we aggregate the scaled identity parameters $s*\frac{\mathbf{x}}{\|\mathbf{x}\|}$ of the same identity to its center c and separate the centers of different identities in Euclidean distance:

$$L_c = \frac{\left\|s*\frac{\mathbf{x}_{y_i}}{\|\mathbf{x}_{y_i}\|}-c_{y_i}\right\|^2}{\frac{1}{n}\sum_{i\neq j}\|c_i-c_j\|^2} \tag{7}$$

where c_i represents the center of the ith class. Finally, we minimize the reconstruction error with basis regularization:

$$L_s = \left\|M-\hat{M}\right\|^2 + w_a\left\|\mathbf{A}^T\mathbf{A}-\mathbf{I}\right\|^2 \tag{8}$$

where M is the ground-truth mesh and $\hat{M}$ is the reconstructed mesh. I is the identity matrix and w_a is the weight of the loss function. Finally, we optimize the following objective function and solve the target basis $\mathbf{A}$:

$$\min_{x,s,c,\mathbf{A}} w_m L_m + w_c L_c + w_s L_s \tag{9}$$

Hyperparameter setting. We use the Adam optimizer, where the initial learning rate of x and s is 0.02 and that of the learning rate of $\mathbf{A}$ is 0.005. The batch size is 512, and the learning rate is reduced to one-tenth for every 20 epochs.

4.4. Training Sphere Face Model with 2D Data

In the second stage, we train a model to reconstruct the 2D face image. Here, the decoder is initialized by the first stage and will be finetuned during this stage. Here, ε denotes the weight of a loss term.

loss function The loss function consists of three components: landmark loss, photometric loss, and recognition loss. The landmark loss and recognition loss would take effect according to the label of training data as follows:

$$L = \begin{cases} L_{pix}(I_r, I) + \varepsilon_l L_{land} + \varepsilon_r L_{reg} & I \in S_{recon} \\ L_{pix}(I_r, I) + \varepsilon_s L_{id} + \varepsilon_r L_{reg} & I \in S_{id} \end{cases} \tag{10}$$

where L_{pix} is the photometric loss, L_{land} is the landmark loss, and L_{id} is the recognition loss. I_r is the rendered image and I is the input image. The set S_{recog} represents the training data with landmark annotations and S_{id} is the the training data with identity annotations. We explain these losses in detail below.

The landmark term L_{land} uses the L_1 loss between projected landmarks $\hat{V}_{2d}$ and ground-truth landmarks V_{2d}:

$$L_{land} = \frac{1}{N} \left\| V_{2d} - \hat{V}_{2d} \right\|_2 \tag{11}$$

where N is the number of landmarks.

Face recognition loss includes three components as shown in Equation(12): a softmax-based loss, a centerness loss, and a Kullback-Leibler loss.

$$L_{id} = L_{soft} + \varepsilon_{center} L_{center} + \varepsilon_{kl} L_{kl} \tag{12}$$

We use cosloss [62] L_{soft} as the softmax-based loss, which applies to the identity parameters. The Kullback-Leibler loss [27] L_{kl} and L_{center} center loss [63] are applied to the scale parameter.

Photometric loss measures the difference between the rendered image and the input image using pixel-wise differences to measure the absolute errors between each corresponding pixel pair with the weights of a confidence map [64], which aims to deal with occlusions or other challenging appearance variations such as beard and hair. The weighted pixel-wise loss is defined as follows:

$$L_{pix}(I_r, I) = -\frac{1}{|\Omega|} \sum_{uv \in \Omega} \ln \frac{1}{\sqrt{2}\sigma_{uv}} \exp - \frac{\sqrt{2}\ell_{1,uv}}{\sigma_{uv}} \tag{13}$$

where $\ell_{1,uv} = |I_r^{uv} - I^{uv}|$ is the L_1 distance between the intensity of input image I and the reconstructed image I_r at location (u, v) and $\sigma \in \mathbb{R}_+^{W \times H}$ is the confidence map. Ω is the 2D image space.

As shown in Equation 14, the regularization term L_{reg} consists of two parts: parameter-level regularity loss L_{preg} and mesh-level regularity loss L_{mreg}.

$$L_{reg} = L_{preg} + \varepsilon_{mreg} L_{mreg} \tag{14}$$

The regularization term of L_{preg} for 3DMM coefficients is defined as:

$$L_{preg} = \varepsilon_{id} \sum_{j=1}^{m_{id}} \alpha_{id_j}^2 + \varepsilon_{exp} \sum_{j=1}^{m_{exp}} \frac{\alpha_{exp_j}^2}{\sigma_{exp_j}^2} + \varepsilon_{alb} \sum_{j=1}^{m_{alb}} \frac{\alpha_{alb_j}^2}{\sigma_{alb_j}^2} \tag{15}$$

where σ_{exp} is an eigenvalue of the expression basis and σ_{alb} is an eigenvalue of the albedo basis. α_{id}, α_{exp} and α_{alb} are the 3DMM parameters which are regressed by the encoder network as shown in Figure 2; m_{id}, m_{exp} and m_{alb} are the dimensions of the shape, expression and albedo parameters respectively.

The mesh-level regular loss consists of the smooth loss, the symmetrical loss and the residual loss.

$$L_{mreg} = L_{smooth} + L_{sym} + L_{res} \tag{16}$$

$$L_{smooth}(\mathbf{G}) = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{G}_i - \frac{1}{|\mathcal{N}_i|} \sum_{\mathbf{G}_j \in \mathcal{N}_i} \mathbf{G}_j \right\|_2 \tag{17}$$

where G is the reconstructed face shape, $\mathcal{N}_i$ denotes a set of a neighboring vertices $\mathbf{G}_i$ and N is the number of vertices.

We assume that the human faces in natural expressions are symmetric about the center axis and add the face shape geometry symmetrical loss:

$$L_{sym}(\mathbf{G}) = \| \mathbf{G} - filp(\mathbf{G}) \|_1 \tag{18}$$

where $flip()$ is the operation to flip the face shape geometry.

The residual loss is:

$$L_{res}(\mathbf{G}) = \left\| \mathbf{G} - \bar{\mathbf{G}} \right\|_1 \tag{19}$$

where $\bar{\mathbf{G}}$ is the mean face geometry.

More Training Details Currently, there are no large public databases that contain both face identity labels and landmark labels. Moreover, since the results of existing face detectors are unsatisfactory in challenging conditions, we do not automatically generate landmarks in the face recognition dataset. Therefore, we use the mixed data from 300W-LP [69] and VGGFace2 [13]. To successfully train our model with the mixed dataset, we use the following strategy to achieve convergence:

(1)*Switch the loss function:* Because the labels in the mixed database are deficient, we determine which loss terms take effect according to the labels of the training samples. For example, if the training sample is from VGGface2, we enable face recognition loss and photometric loss. Otherwise, the landmark loss and photometric loss take effect as shown in Equation 10.

(2) *Warm up the network:* To warm up the network, we train our network on the 300W-LP [69] database only using S_{recon}, then train the mixed database with the full loss function shown in 10.

(3) *Balance the data from different datasets:* Because the VGGface2 contains 3.31 million images while 300W-LP [43] contains 61,255 samples, which are extremely un-

Model	PCA	Linear	Sphere-Linear	SFM
RMSE	**0.2777**	0.2916	0.2808	0.2827
SCE	-0.0490	-0.0492	-0.0490	**0.1193**
SCC	-0.1068	-0.1073	-0.1068	**0.2038**
CH	15.86	15.89	15.86	**25.81**

Table 1. The results of model representation ability and its shape parameter separability in FRGCv2 database.

Model	PCA	Linear	Sphere-Linear	SFM
RMSE	**0.3747**	0.3924	0.3790	0.3863
SCE	0.1061	0.1059	0.1061	**0.2236**
SCC	0.1474	0.1470	0.1473	**0.3513**
CH	9.01	9.03	9.01	**10.28**

Table 2. The results of model representation ability and its shape parameter separability in Bosphorus database.

Model	BFM17	FLAME	SFM	SFM*
RMSE	0.6137	0.4820	**0.3501**	0.4355
SCE	0.1674	0.0928	0.2247	**0.2953**
SCC	0.2451	0.1700	0.3616	**0.4514**
CH	6.87	3.85	9.50	**11.04**

Table 3. Compare the model representation ability and its shape parameter separability with BFM17 [21] and FLAME [30] in Bosphorus database.The SFM* is the SFM finetuned with 2d data.

balanced, we design a sampling scheme where the probability of selecting samples from the VGGFace2 is given by:

$$P = \frac{N_{recon}}{N_{recog} + N_{recon}} \tag{20}$$

Here, N_{recon} is the number of samples in 300W-LP dataset and N_{recog} is that in VGGFace2 dataset. The probability of selecting samples from 300W-LP database is $1 - P$.

5. Experiment

Comparing with the previous methods, SFMs have the following properties: (1) The shape parameter space of SFMs has inherent separation property between the various classes; (2) The shape parameter space distribution of SFM is similar to that of identity embeddings, so that the losses for face recognition and face reconstruction can be easily optimized together in the pipeline of shape-consistent face reconstruction; (3) SFM has better capabilities for face representation. Therefore, in this section, we evaluate SFMs from the following three aspects: model representation ability, shape parameter space separability, and shape-consistent monocular reconstruction performance.

5.1. Model Representation Ability

To validate the expressive ability of face representation models, we reconstruct 3D meshes on the training and testing database, respectively, with the same dimension of the latent vector(all of 199 dimensions in this paper). Evaluation of the training database shows the ability of the models to recover the meshes of the training data. We also verify the generalizability of our model by fitting meshes for the testing database. We also present the result of the parameter interpolation.

Our training dataset is FRGCv2 [41], and the testing dataset is the Bosphorus Database [45], which contains 4666 3D face models of 105 people. The models for each person have various expressions, poses, and occlusions. In our experiment, we select the face with a frontal natural expression for each person, and register all the data on the BFM 2017 [4] template. We first use rigid registration [7] to align the template with the point cloud roughly and then use non-rigid ICP [3]. When performing non-rigid registration, we first register with strong, rigid regular parameters and then use smaller regular parameters to perform more delicate meshes registration.

In our experiment, we select the face with a natural frontal expression for each person and register all the data on the BFM 2017 [4] template using non-rigid ICPs [3]. We use the Adam optimizer to optimize the face parameters of the model. The initial learning rate is 0.02 and reduced by a factor of 0.5 for every 128th iteration. The total optimization iteration number is 1000.

The Root Mean Square Error(RMSE) between the reconstructed meshes and the ground truth for the training dataset is shown in Table 1, and that for the testing dataset is in Table 2. We use the face model trained with FRGCv2 but use different methods when generating the above results. “PCA” means the face model is directly established by the PCA method. “Linear” means the face model is established by optimizing an orthogonal linear basis. “Sphere-Linear” refers to using the structure of SFM without the loss of face recognition when constructing the face shape. The expression ability of our SFM basis is slightly better than that of the linear basis but worse than PCA. Because when the face model has a linear orthogonal basis, the basis solved by PCA has the smallest reconstruction error, which is the optimal solution. Our reconstruction accuracy is slightly lower than PCA’s, but has a better separation in the shape parameter space.

Table 3 shows the comparison between SFMs with the shape models of BFM2017 and FLAME on the Bosphorus database. We crop the face area for fitting because other areas (ears, neck) are irrelevant to our task and can largely influence the RMSE. We use the point-to-plane error to calculate RMSE. The results show that our face model has fewer reconstruction errors than others. Figure 3 shows some fitting results on the Bosphorus database. SFMs are competitive among all the validated 3DMMs in terms of expressive ability, with the best visual quality of the generated recon-

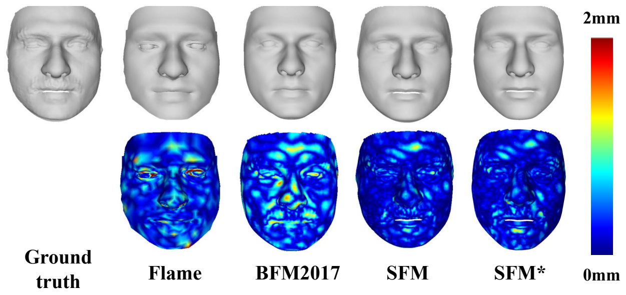

Figure 3. The fitting results of BFM17 [21], FLAME [30] and ours. The first row is the fitted mesh and the second row is the error map with ground truth. SFM* is the SFM fine-tuned with 2d data.

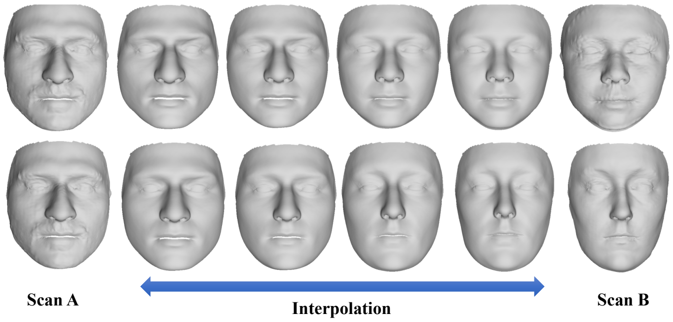

Figure 4. On the left and right are two different scanning models. We first find their identity parameters and scale parameters. Then we perform the interpolation of the identity parameters on the hypersphere and perform linear interpolation on the scale parameter. Columns 2-5 are the result of interpolation.

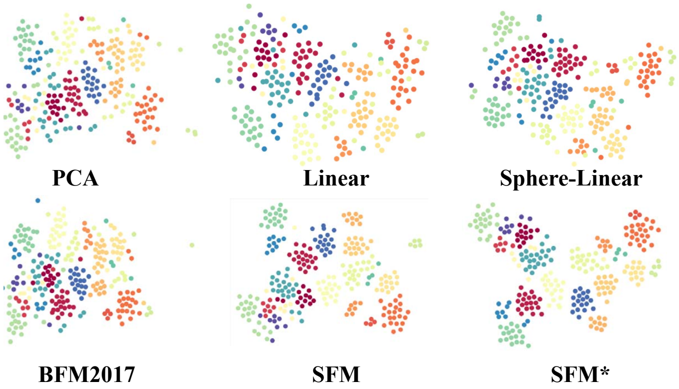

Figure 5. The latent vector distributions of different methods. We select 20 people on FRGCv2, fitting the shape parameter, and then use t-SNE to reduce the shape parameter to two dimensions and display it on this figure. Different colors represent different people.The SFM* is the SFM fine-tuned with 2d data.

Method	LFW	CFP-FP	YTF
Cosine similarity			
3DMM-CNN	90.53	-	88.28
Lui et al.	94.40	-	88.74
D3FR	88.98	66.58	81.00
TDDFA	64.90	57.57	58.50
MGCNet	82.10	70.87	75.58
RingNet	79.40	71.41	71.02
DECA	81.70	65.98	78.64
Jiang et al	95.36	83.34	89.07
SFM	97.23	89.12	91.35
SFM*	**98.23**	**91.12**	**93.86**
Euclidean similarity			
D3FR	87.63	66.50	81.10
TDDFA	63.45	55.49	58.16
MGCNet	80.87	66.01	72.36
RingNet	80.05	69.46	72.40
DECA	80.32	63.49	76.46
Jiang et al	94.47	80.78	86.40
SFM	97.07	87.12	90.43
SFM*	**98.03**	**90.79**	**92.60**

Table 4. Face verification accuracy(%) on the LFW, CFP-FP and YTF datasets. Our results are obtained using the weighted center loss. We compare our results with 3DMM-CNN [57], Liu et al. [33], D3FR [17], TDDFA [22], MGCNet [49], Jiang et al [24], RingNet [44] and DECA [18]. SFM* is the SFM fine-tuned with 2d data.

struction results.

Figure 4 shows that the parameters of our basis have an excellent interpolation performance. We use the geodesic distance to interpolate the identity parameters and directly interpolate the scale parameters linearly.

5.2. Separability of Shape Parameter Space

After fitting all the 3D scans of a database, we get the parameters of the corresponding 3DMM model in this database. We can evaluate the clustering properties of these parameters to estimate the degree of separation of shape parameter space. The performance of clustering can be evaluated with the following metrics: the Silhouette Coefficient score with Euclidean distance(SCE), Silhouette Coefficient with Cosine distance(SCC), and Calinski-Harabasz score indicators(CH). The Silhouette Coefficients are given as:

$$s = \sum_{i=1}^{n} \frac{a_i - b_i}{max(a_i, b_i)} \tag{21}$$

where a_i is the mean distance between ith sample and all other points of the same class and b_i is the mean distance between ith sample and all other points of the nearest cluster. n is the number of the sample. The score is the ratio of the sum of between-cluster dispersion and of within-cluster dispersion for all clusters (where dispersion is defined as the sum of squared distances). The Calinski-Harabasz

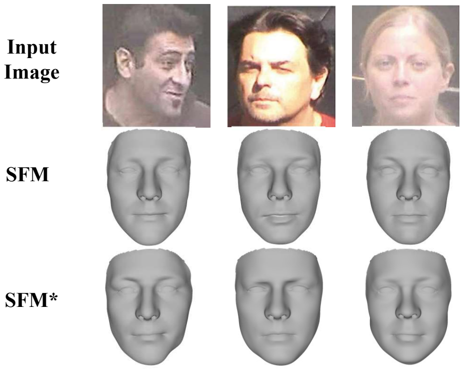

Figure 6. Ablation experimentson samples from MICC [5] dataset. SFM* is the SFM fine-tuned with 2d data.

score(CH) is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$s = \frac{tr(\mathbf{B}_k)}{tr(\mathbf{W}_k)} \times \frac{n_E - k}{k - 1} \tag{22}$$

where $\mathbf{B}_k$ is the trace of the between-cluster dispersion matrix and $\mathbf{W}_k$ is the trace of the within-cluster dispersion matrix defined by:

$$\begin{aligned} \mathbf{W}_k &= \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \\ \mathbf{B}_k &= \sum_{q=1}^{k} (n_q)(c_q - c_e)(c_q - c_E)^T \end{aligned} \tag{23}$$

with C_q the set of points in cluster q, c_q the center of cluster q, c_E the center of E, and n_q the number of points in cluster q.

Table1 and Table 2 show the results of shape parameter space separation of SFMs and the shape basis constructed by other methods. We add the face recognition loss while establishing the SFM basis, significantly improving shape parameter space separation. Table 3 shows the comparison between our basis and other basis. The separability of our shape parameter space is also higher than other models. In order to present the distribution of shape parameter space more intuitively, we use t-SNE [58] to project the shape parameters of different bases to two dimensions. As shown in Figure 5, the intra-class distance of the shape parameter space of SFM is small, and the inter-class distance is large. Compared with other methods, the shape parameter space of our basis shows a much better separation.

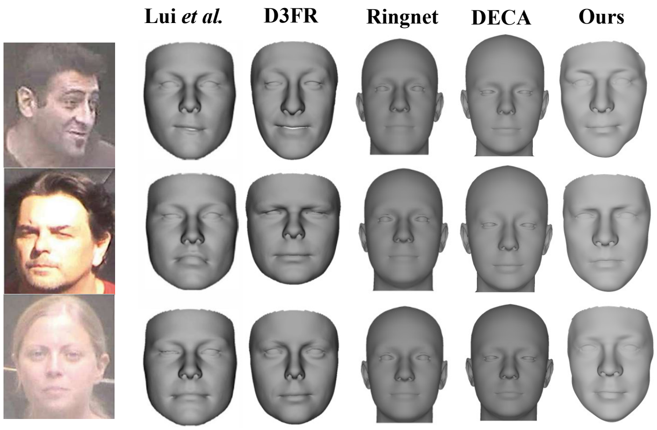

Figure 7. Comparison with liu et al. [33], Ringnet [44], D3FR[17] and DECA [18] on three MICC [5] subjects. Our reconstruction results capture more face details.

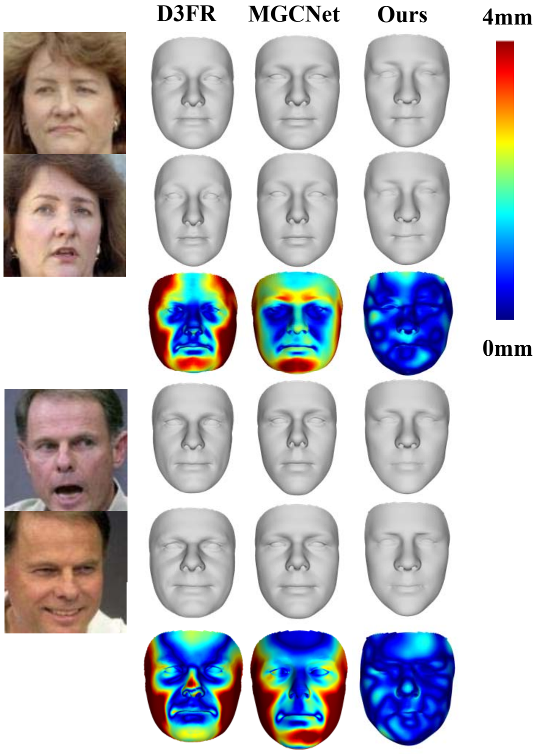

Figure 8. Comparison with D3FR [17], MGCNet [49] on LFW samples The reconstruction results are the same person under different conditions. The third and sixth rows are the error between two meshes in the same column.

5.3. Monocular Reconstruction

To test the face monocular reconstruction, we use the same encoder-decoder network in the second training stage as the shape-consistent face reconstruction pipeline. Unlike the training phase, when performing inference for monocu-

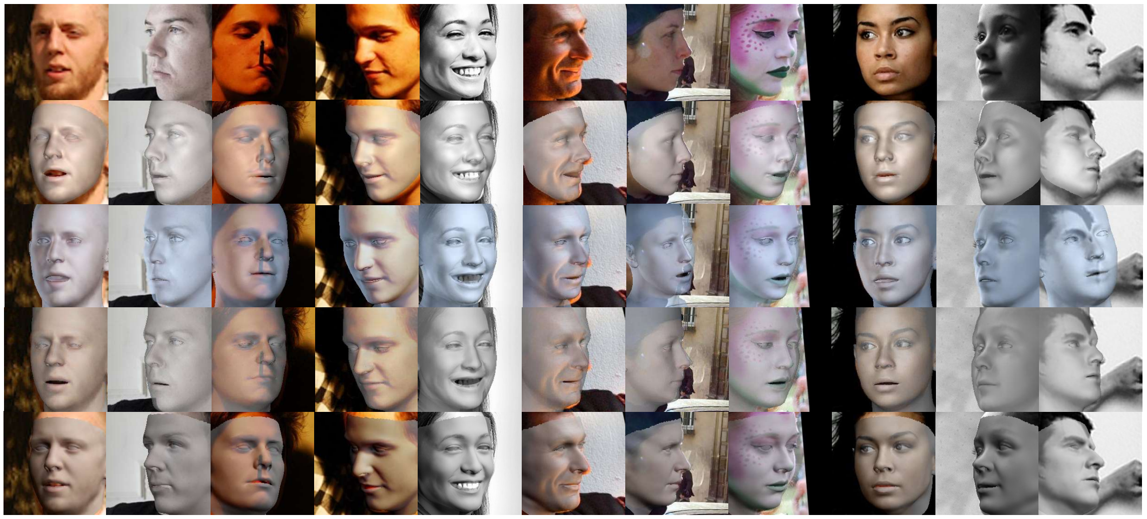

Figure 9. The visualization results on ALFW2000 dataset. The first row: images from ALFW2000 dataset. The second row: the result of 3DDFA v2 [22]. The third row: the results of ringnet [44]. The forth row: the results of DECA [18]. The last row: the results of ours.

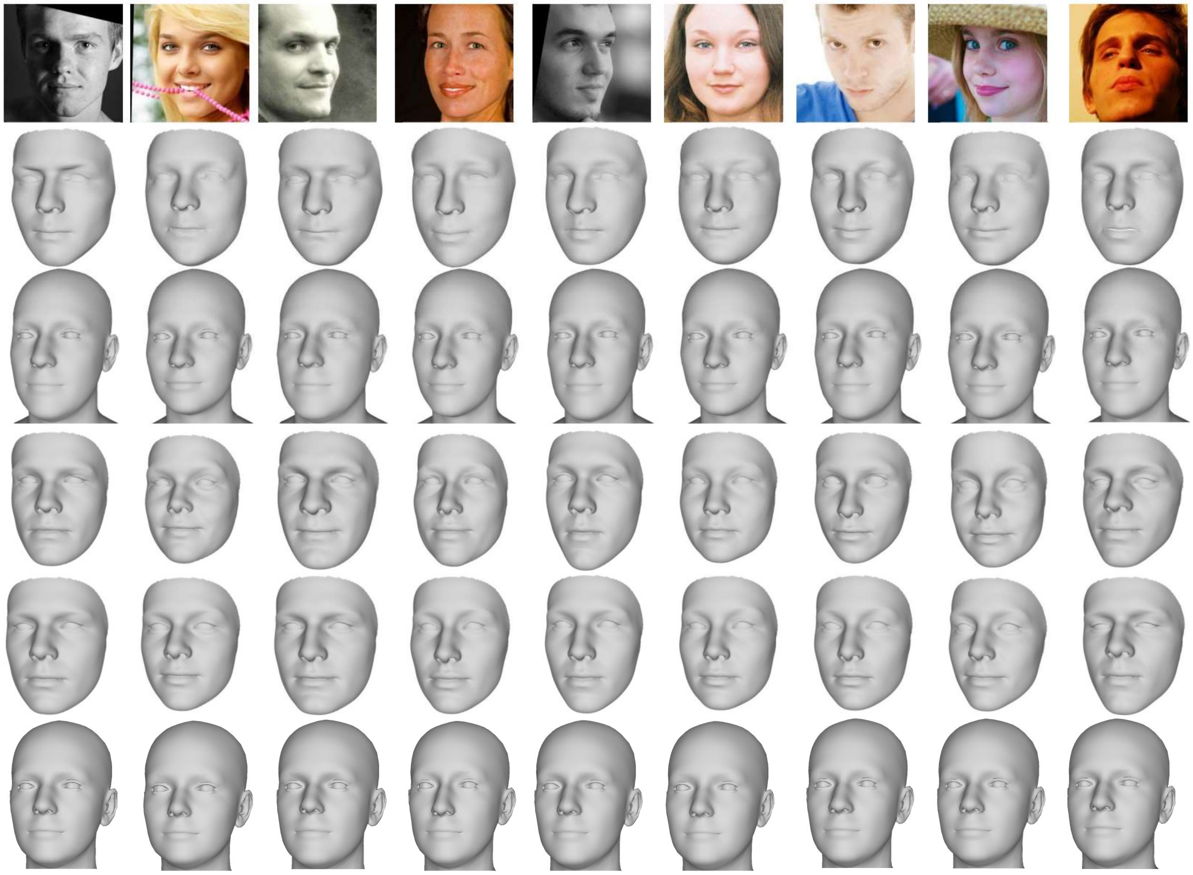

Figure 10. Some samples of user study, The first row: samples from ALFW2000 dataset. The second row: our results. The third row: the results of ringnet [44]. The forth row: the results of D3FR [17]. The fifth row: the results of MGCNet [49]. The last row: the results of DECA [18].

lar reconstruction, we fix the weight of the shape basis and retrain the encoder to regress the parameters. In this section, we evaluate the faithfulness and shape consistency of monocular face reconstruction results using SFM. In terms of faithfulness, we compared the visual results with other face reconstruction methods. Moreover, we compare the accuracy of 3D face alignment. In terms of shape consistency, we compare the accuracy of the face recognition using the shape parameters and the visual results of the same person reconstructed in different environments. In this subsection, when comparing with methods, "ours" means that we use the SFM finetuned with 2D data.

Face shape consistency evaluation. We use the cosine distance and Euclidean distance as the similarity measurements between two groups of shape parameters, when evaluating the face recognition accuracy. The result of face recognition performance is shown in Table 4. The accuracy of our face recognition parameters is higher than other methods. The results get better after SFM is finetuned with 2D data because finetuning with 2D face data results in a more robust generalization model. Figure 8 shows the visualization results of the 3D face reconstructed by the same person in different environments. We have smaller errors among the meshes reconstructed for the same person.

Face faithfulness evaluation. As shown in Figure 6, finetuning with 2d data can improve the expression ability of the model and generate faces with more details. Figure 7 shows that our reconstruction results capture more face details compared to other face reconstruction methods. Figure 11 shows the cumulative errors distribution curve of 3D face alignment compared with other methods, Figure 9 shows the visual results of face alignment and Figure 10 shows the visual results of face shape. Both quantitative and visual evaluations show that in terms of face faithfulness our method has better performance than previous methods.

User study. We conducted a user study to compare the visual diversity and the degree of retention of the reconstructed face shape on the identity information. We randomly selected 20 face images from ALFW2000 and reconstructed 3D face models using the following methods: RingNet [44], D3FR [17], MGCNet [49] and our SFM, and in turn asked 5 participants to evaluate the reconstructed faces' diversity and the retention of identity information of the reconstructed faces from the input image with a score from 0 to 10. Participants were told that the reconstruction results with more identity information maintained or more diversity of different people should be scored higher. The average scores of the results from different methods are shown in the Figure 12. The "identity" means the degree of identity preservation, and the "diversity" means the diversity among the 3D faces reconstructed from different people. Results and comparisons vividly show the advantages of our method.

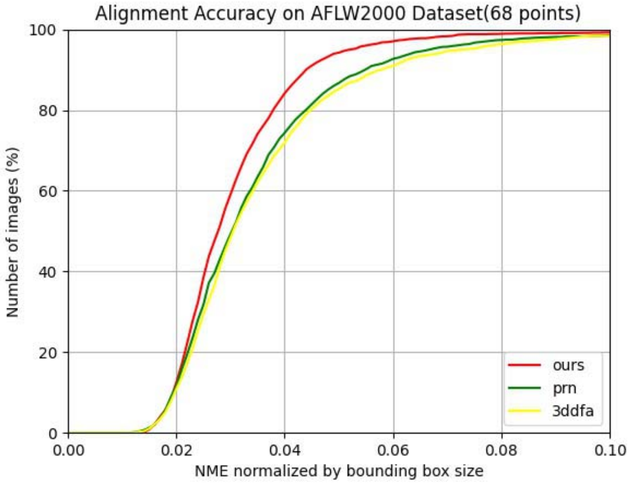

Figure 11. The cumulative errors distribution curve of 3d face alignment accuracy on the ALFW2000 Dataset. Compared with PRNet [19], 3ddfa [70], our method produces better results.

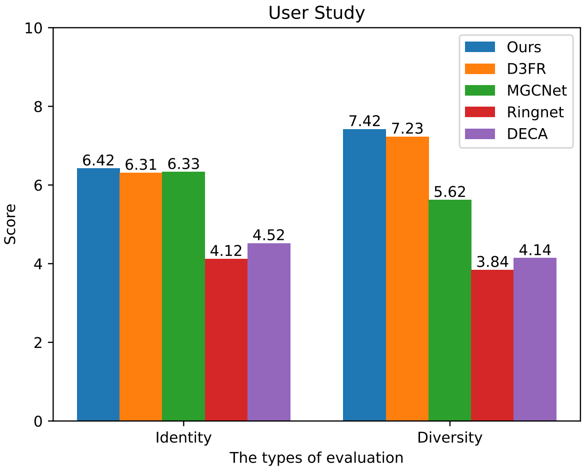

Figure 12. The results of user study. We compare our SFM with RingNet [44], D3FR [17], MGCNet [49] and DECA [18] in terms of identity and diversity, and our results are more satisfactory.

6. Conclusion

We have proposed a novel 3D morphable model with a hypersphere manifold shape parameter space for face generation. We have also proposed a two-stage training framework where both 3D and 2D data were utilized. Our model outperformed previous models on the consistency and the fidelity of the reconstructed faces. Experimental results validated that our method is superior to previous methods objectively, and user study showed that our model can provide visually better face reconstruction results.

References

[1] O. Aldrian and W. A. Smith. Inverse rendering in suv space with a linear texture model. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 822–829. IEEE, 2011. 3

[2] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3d face recognition with a morphable model. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008. 3

[3] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 6, 8

[4] T. Bagautdinov, C. Wu, J. Saragih, P. Fua, and Y. Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 3, 4, 8

[5] A. D. Bagdanov, A. Del Bimbo, and I. 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, 2011. 10

[6] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision*, pages 377–391. Springer, 2016. 3

[7] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 8

[8] S. Bian, A. Zheng, L. Gao, G. Maguire, W. Kokke, J. Macey, L. You, and J. J. Zhang. Fully automatic facial deformation transfer. *Symmetry*, 12(1):27, 2020. 1

[9] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library, 2003. 3

[10] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3, 4

[11] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2016. 1, 3

[12] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG)*, 32(4):1–10, 2013. 3

[13] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6, 7

[14] S.-Y. Chen, L. Gao, Y.-K. Lai, P. L. Rosin, and S. Xia. Real-time 3d face reconstruction and gaze tracking for virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 525–526. IEEE, 2018. 1

[15] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, and L. Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020. 3

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3, 4

[17] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 9, 10, 11, 12

[18] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *arXiv preprint arXiv:2012.04012*, 2020. 3, 9, 10, 11, 12

[19] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 12

[20] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019. 3

[21] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 1, 6, 8, 9

[22] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. *arXiv preprint arXiv:2009.09960*, 2020. 3, 9, 11

[23] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 3

[24] D. Jiang, Y. Jin, R. Deng, R. Tong, F.-L. Zhang, Y.-K. Lai, and M. Tang. Reconstructing recognizable 3d face shapes based on 3d morphable models. *arXiv preprint arXiv:2104.03515*, 2021. 3, 4, 9

[25] Z.-H. Jiang, Q. Wu, K. Chen, and J. Zhang. Disentangled representation learning for 3d face shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[26] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016. 1

[27] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 7

[28] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2020. 1

[29] H. Li, T. Weise, and M. Pauly. Example-based facial rigging. *Acm transactions on graphics (tog)*, 29(4):1–6, 2010. 3

[30] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1, 3, 4, 8, 9

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

[32] J. Lin, Y. Yuan, and Z. Zou. Meingame: Create a game character face from a single portrait. *arXiv preprint arXiv:2102.02371*, 2021. 1

[33] F. Liu, R. Zhu, D. Zeng, Q. Zhao, and X. Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5216–5225, 2018. 1, 3, 4, 9, 10

[34] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019. 3

[35] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017. 3, 4

[36] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516, 2016. 3

[37] Y. Liu, H. Li, and X. Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*, 2017. 3

[38] A. Patel and W. A. Smith. Manifold-based constraints for operations in face space. *Pattern recognition*, 52:206–217, 2016. 5

[39] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. 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. Ieee, 2009. 1, 3

[40] P. Paysan, M. Lüthi, T. Albrecht, A. Lerch, B. Amberg, F. Santini, and T. Vetter. Face reconstruction from skull shapes and physical attributes. In *Joint Pattern Recognition Symposium*, pages 232–241. Springer, 2009. 3

[41] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005. 6, 8

[42] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6

[43] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 6, 7

[44] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 1, 3, 9, 10, 11, 12

[45] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *European workshop on biometrics and identity management*, pages 47–56. Springer, 2008. 8

[46] A. Schneider, S. Schonborn, L. Frobeen, B. Egger, and T. Vetter. Efficient global illumination for morphable models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3865–3873, 2017. 3

[47] S. Schönborn, B. Egger, A. Morel-Forster, and T. Vetter. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, 123(2):160–183, 2017. 3

[48] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[49] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 9, 10, 11, 12

[50] T. Shi, Z. Zuo, Y. Yuan, and C. Fan. Fast and robust face-to-parameter translation for game character auto-creation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1733–1740, 2020. 1

[51] W. A. Smith, A. Seck, H. Dee, B. Tiddeman, J. B. Tenenbaum, and B. Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020. 6

[52] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014. 3

[53] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 3, 6

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

[55] L. Tran, F. Liu, and X. Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 3, 4, 5, 6

[56] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 3, 4, 5

[57] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 1, 3, 9

[58] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 10

[59] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*, pages 24–es. Association for Computing Machinery, 2006. 3

[60] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 3

[61] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 3, 4, 6

[62] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 3, 4, 7

[63] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 3, 7

[64] S. Wu, C. Rupprecht, and A. Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 7

[65] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 601–610, 2020. 3

[66] L. Yang, J. Wu, J. Huo, Y.-K. Lai, and Y. Gao. Learning 3d face reconstruction from a single sketch. *Graphical Models*, 115:101–102, 2021. 1

[67] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019. 3

[68] W. Zhu, H. Wu, Z. Chen, N. Vesdapunt, and B. Wang. Reda: reinforced differentiable attribute for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4958–4967, 2020. 6

[69] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 6, 7

[70] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017. 1, 12