Reconstructing Recognizable 3D Face Shapes based on 3D Morphable Models

Diqiong Jiang[1], Yiwei Jin[1], Risheng Deng[1], Ruofeng Tong[1], Fanglue Zhang[2], Yukun Yai[3] and Ming Tang[1]

[1]Zhejiang University
[2]Victoria University of Wellington
[3]Cardiff University

***Abstract*—Many recent works have reconstructed distinctive 3D face shapes by aggregating shape parameters of the same identity and separating those of different people based on parametric models (e.g., 3D morphable models (3DMMs)). However, despite the high accuracy in the face recognition task using these shape parameters, the visual discrimination of face shapes reconstructed from those parameters is unsatisfactory. The following research question has not been answered in previous works: Do discriminative shape parameters guarantee visual discrimination in represented 3D face shapes? This paper analyzes the relationship between shape parameters and reconstructed shape geometry and proposes a novel shape identity-aware regularization(SIR) loss for shape parameters, aiming at increasing discriminability in both the shape parameter and shape geometry domains. Moreover, to cope with the lack of training data containing both landmark and identity annotations, we propose a network structure and an associated training strategy to leverage mixed data containing either identity or landmark labels. We compare our method with existing methods in terms of the reconstruction error, visual distinguishability, and face recognition accuracy of the shape parameters. Experimental results show that our method outperforms the state-of-the-art methods.**

***Index Terms*—Reconstruction, Shape modeling, 3D morphable model**

I. Introduction

Facial shape estimation from a single RGB image has been an active research topic in both computer vision and computer graphics, and has various applications in fields such as VR/AR, animation, face editing, and biometrics. Early works [1], [2], [3], [4] focused on estimating whether the projection of 3D faces is faithful to the input image by minimizing sparse landmark location losses and dense photometric losses. However, if we use only the supervision signal from the discrepancy between the input image and the projected counterpart, the face shapes reconstructed from different images of the same person might look dissimilar and be difficult to visually recognize. One underlying reason is that the expression and pose could influence the intrinsically identifiable features of the recovered shapes. Empirically speaking, the face shape contributes much less error than the expression and pose in landmark location losses. Therefore, minimizing only the discrepancy between the input image and the projected counterpart makes it difficult to find the optimal face shape consistency among different images of the same person. Based on this observation, learning to regress a recognizable 3D face shape from a single image with varying poses and expressions has attracted much attention in recent years. A straightforward solution to this problem is aggregating the shape parameters of the same identity and separating those of different people. Tran et al.[5] pool shape parameters belonging to the same identity to decrease their intraclass variance. Sonyal et al.[6], and Liu et al.[7] apply shape consistency losses to make shape parameters discriminative and recognizable. Their shape parameters achieve sustainable high performance in face recognition, but the 3D face shape still fails to be visually discriminative, since the authors focus on improving the discrimination of shape parameters while ignoring the relationship between shape parameters and shape geometries. Therefore, to transfer the discrimination of shape parameters to 3D geometries, the relationship between shape parameters and 3D geometries needs to be carefully investigated instead of generally applying shape consistency losses to shape parameters.

The aim of our research is to reconstruct a stable and recognizable 3D face shape. More particularly, the reconstructed results from the proposed method must meet the following criteria: (1) the face shapes of the same identity must have a low root mean squared error (RMSE) with each other. (2) The RMSE of face shapes among different people must be sufficiently high to ensure that the differences can be visually perceived by humans. (3) The reconstructed 3D face shapes under different expressions and poses need to be visually identifiable and the same as those reconstructed from a neutral frontal face image. To achieve the above goals, the following conditions need to be satisfied in the method: (1) the 3D shape parameters can be discriminative by the Euclidean distance; (2) the centers of 3D shape parameters of the same identity are the parameters regressed from natural frontal face images; (3) the 3D shape basis is orthonormal; and (4) the distribution of the 3D shape parameters follows a particular multivariate Gaussian distribution, which is an inherent property when constructing the 3D morphable model(3DMM). In this paper, we design a novel shape identity-aware regularization(SIR) loss, which explicitly imposes shape consistency on the shape parameter space and implicitly guides generated face shape geometry to be visually recognizable. As shown in Table I, the loss functions proposed by existing methods [7], [6] do not satisfy all these conditions at the same time. The lack of a large database containing both identity information and 3D geometric information (3D face geometry or facial

landmarks) also makes the task of learning recognizable face shapes difficult.

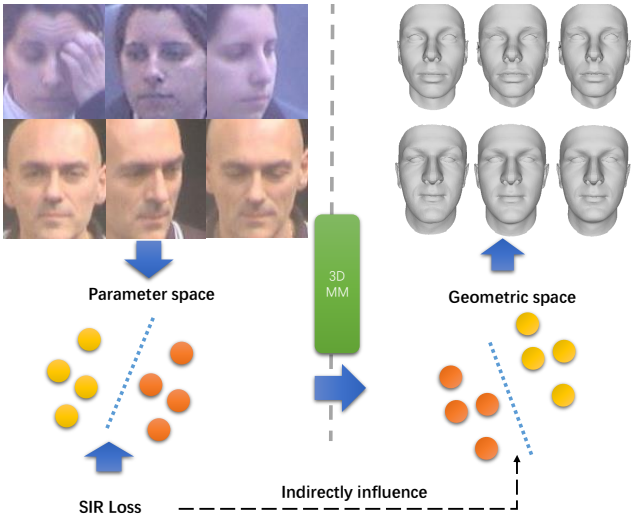

Figure 1. SIR loss is designed according to the relationship between 3DMM parameter space and geometric space. Therefore, although it imposes only shape consistency directly onto the shape parameters, it essentially guides the face geometry. These samples show that the represented face shapes are visually discriminative.

To obtain a sufficient amount of training data, Tran et al. [5] and Liu et al. [7] use 3D face geometries in their methods, which are produced by 3D facial labels estimated from 2D images. Their methods are limited by the capability of geometry reconstruction algorithms. Another work[6] detects facial landmarks annotated by a facial detector, but the detected labels are inaccurate, especially in challenging situations. We apply a more flexible way of dealing with the lack of training data. Our network and training strategy can accept images labeled with identity or geometry information, such as facial landmarks, and combine them during the training process. Consequently, the generation of extra database with annotations of both identity and face geometry is unnecessary. In fact, our network allows us to use any off-the-shelf face recognition and face reconstruction database for training.

This paper investigates the relationship between the 3DMM parameter space and 3D geometric space and presents a method to transfer discrimination from the 3DMM parameter space to the geometric space. We propose a novel SIR loss function for face reconstruction. The SIR loss comprises two terms: an identification term, including inter-class separation loss and intraclass aggregation loss, and a parameter distribution term. As Figure 1 shows, the SIR loss explicitly imposes shape consistency on shape parameters while implicitly guiding face shapes such that they are visually discriminative. The main contributions of this paper include the following:

- We investigate the relationship between the 3DMM parameter space and 3D geometric space and propose that a deep model should follow four principles so that the results are discriminative in both the parameter and geometry domains.
- We propose a deep network that is capable of transferring discriminative features from the shape parameter space to the geometry space with off-the-shelf face recognition and face reconstruction datasets as training data. We also propose an effective training paradigm that leads our network robustly converge with incompletely labeled training data.
- We propose the SIR loss, which explicitly regularizes 3DMM shape parameters to satisfy all four aforementioned conditions while implicitly guiding face shapes to be visually discriminative. The parameter distribution term of the SIR loss ensures that the shape geometry discrimination is also visually discriminative.

Condition	(1)	(2)	(3)	(4)
Liu et al.[7]			✓	
Ringnet[6]	✓		✓	
Ours	✓	✓	✓	✓

Table I
THE CONDITION IS SATISFIED BY DIFFERENT METHODS. (1) THE 3D SHAPE PARAMETERS ARE DISCRIMINATIVE WITH REGARD TO THE EUCLIDEAN DISTANCE; (2) THE CENTERS OF 3D SHAPE PARAMETERS WITH THE SAME IDENTITY ARE PARAMETERS REGRESSED FROM A NEUTRAL FRONTAL FACE IMAGE; (3) THE 3D SHAPE BASIS IS ORTHONORMAL; (4) THE 3D SHAPE PARAMETER SATISFIES A PARTICULAR MULTIVARIATE GAUSSIAN DISTRIBUTION.

II. Related Work

In the 3D face reconstruction field, the 3D faces are reconstructed from various inputs, including a depth map [8], video[9], [10], [11], [12], [13], multi-view images[14], [15], [16] and a single image[17], [18], [19], [20]. Among them, reconstructing a 3D face from a single image has attracted more attention because of its simplicity and usability. After several years of research, monocular reconstruction was generalized from coarse-level reconstruction by parametric face models[2] to medium-level[21], [22] and fine-level shape[17], [23] corrections. Recently, some works[5], [7], [6] considered using shape consistency to make reconstructed face shapes recognizable. In the rest of this section, we focus on 3DMM face reconstruction and shape-consistent face reconstruction which are more closely related to our work.

Monocular 3D face reconstruction based on 3DMM. The groundbreaking work of monocular 3D face reconstruction with statistical models can be traced back to Blanz and Vetter[1], [2] which recovered facial geometries by solving an optimization problem constrained by linear statistical model, i.e., 3DMMs. Paysan et al. [3] and Zhu et al. [4] extend the 3DMM with pose and expression parameters. In recent years, Deep convolutional neural networks (DCNNs) have shown strong capabilities in many computer vision tasks. The existing literature [24], [25], [5], [20], [26], [17] reveals that CNNs can effectively regress the 3DMM parameters with sufficient training data. They provide comparable reconstruction precision with much less computation and adapt to input images under challenging conditions. Richardson et al.[27] build a synthetic dataset using the 3DMM with random shape, expression, and pose parameters, and render them as 2D images with different levels of illumination. However, the synthesized data cannot capture the complexity of the real

world. Zhu et al.[28] fit 3D shapes with traditional methods and augment data by applying the image warping technique to simulate in-plane and out-of-plane head rotation. They build the 300W-LP dataset, which covers various head poses and facial expressions with labeled 3DMM coefficients. In this way, the shape labels are ambiguous and inaccurate because they are constrained only by sparse facial landmarks in the fitting process. Sanyal et al.[6] regress the 3D shape parameters without any supervised 2D-to-3D training data. The landmark labels are detected by a face detection algorithm, which is not very precise in challenging conditions (e.g., large poses and poor lighting conditions). Otherwise, sparse landmarks cannot capture sufficient recognizable features in face geometries. With the development of generic differentiable rendering[29], [30], [31], [18], [32], [33] train networks without shape labels in an unsupervised or weakly-supervised way by constraining the consistency between rendered images and raw image signals. The photometric consistency can capture more geometric details, especially from the frontal face. Our method uses the 300W-LP facial landmarks and the pixelwise photometric difference as our reconstruction training losses.

Shape-consistent face reconstruction. Many works [25], [19], [34] pursue alignment accuracy or pixelwise appearance accuracy to get precise face geometries. However, the final face geometry is composed of face shape, expression and pose parameters. Any of those parameters could dominate the reconstruction if the model is poorly trained. Therefore, a well-aligned face geometry does not guarantee the accuracy of a face shape. To reconstruct a stable and visually discriminative face shape, Tran et al. [5] label a very number of face images with 3DMM shape parameters and develop a deep CNN to learn the mapping from images to shape parameters. During training, they pool coefficients that belong to the same identity to give their output feature lower intraclass variance. Liu et al.[7] propose a multitask deep CNN to disentangle identity from "residual attribute" to learn the 3D face shape and discriminative authentication feature together. They use a softmax loss function to directly push away the shape parameters of different people while aggregating those of the same person. Compared with pooling, their loss function can achieve even better face recognition accuracy with a simpler network structure. Sanyal et al.[6] achieve a similar goal by introducing a shape consistency loss embodied in a ring-structured network. Their methods aim only to achieve "shape parameter consistency" rather than "shape geometric consistency visually", and does not attempt to explain the relationship between shape parameter consistency and visual shape geometric consistency. Our work takes both shape parameter and geometry discrimination into consideration and proposes SIR loss to separate shape parameters explicitly and distinguish face geometries implicitly.

III. Our Method

This section first introduces the parametric face model. Then, we investigate the relationship between the 3DMM parameter space and 3D geometry space and propose principles that the deep neural network should follow to make the results discriminative in both the parameter and geometry domains. Finally, the network, loss function, and training strategy are designed to make our deep neural network satisfy these principles.

A. Parametric Face Model

We follow the previous work[25] which combines the Basel Face Model-09 [3] and FaceWarehouse [35] by Equation(1) for our 3DMM representation to describe the geometry of a 3D face model

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp} \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{3n}$ is a reconstructed 3D face with n vertices, which is controlled by the shape parameter vector α_{id} and the expression parameter vector α_{exp} for representing various shape identities and expressions. $\bar{\mathbf{S}} \in \mathbb{R}^{3n}$ is the mean face shape. The orthogonal matrices $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ are the bases of shape and expression, respectively.

Six degrees-of-freedom (rotation and translation) are required to describe the camera pose. More specifically, 3DMM meshes are transformed by the camera pose $[\mathbf{R}|\mathbf{t}_{3d}] \in SE3$ by the following Equation(2)

$$\mathbf{V}_{3d} = \mathbf{R} \cdot (\bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}) + \mathbf{t}_{3d} \tag{2}$$

where $\mathbf{V}_{3d}$ denotes the 3D vertices of the transformed 3DMMs in the camera coordinate system. $\mathbf{t}_{3d} \in \mathbb{R}^{3n}$ is the translation matrix and we use a quaternion $\mathbf{R}$ to represent rotations.

We apply weak projections to project the 3DMM meshes to the image plane so that a scalar f can be introduced as the focal length to perform the projection as in Equation(3).

$$\mathbf{V}_{2d} = f \cdot \mathbf{Pr} \cdot \mathbf{R} \cdot (\bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}) + \mathbf{t}_{2d} \tag{3}$$

where $\mathbf{V}_{2d}$ denotes the projected 2D coordinates of the 3D model vertex matrix $\left(\begin{smallmatrix}1&0&0\\0&1&0\end{smallmatrix}\right)$.

B. Properties of 3DMM model

In this subsection, we explore the underlying relationship between 3DMM shape parameters and 3D shape geometries and the conditions in which the separable shape parameters lead to visually distinguishable face geometries. For simplicity, we focus only on face shapes regardless of expression and pose in this sub-section, .

From parameter discrimination to shape geometry discrimination. We denote $A \in \mathbb{R}^{3n\times m}$ as the shape basis, $\alpha \in \mathbb{R}^m$ as the shape parameters and $\bar{\mathbf{S}} \in \mathbb{R}^{3n}$ as the mean face vertices. m is the dimension of the shape parameters, and n is the number of vertices of the face shape. The face shape $\mathbf{S}$ is represented by Equation(4).

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{A}\alpha \tag{4}$$

Suppose x, $y \in \mathbb{R}^m$ are the shape parameters of two faces. Denote $e = \frac{1}{m}\|x-y\|_2^2$ as the square of their Euclidean distance. Accordingly, X, $Y \in \mathbb{R}^{3n}$ are their corresponding face shapes and the square of their Euclidean distance is $E = \frac{1}{n}\|X-Y\|_2^2$. According to Equation(4), E can be calculated by:

$$E = \frac{1}{n}\|X-Y\|_2^2 = \frac{1}{n}\|\mathbf{A}(x-y)\|_2^2 \tag{5}$$

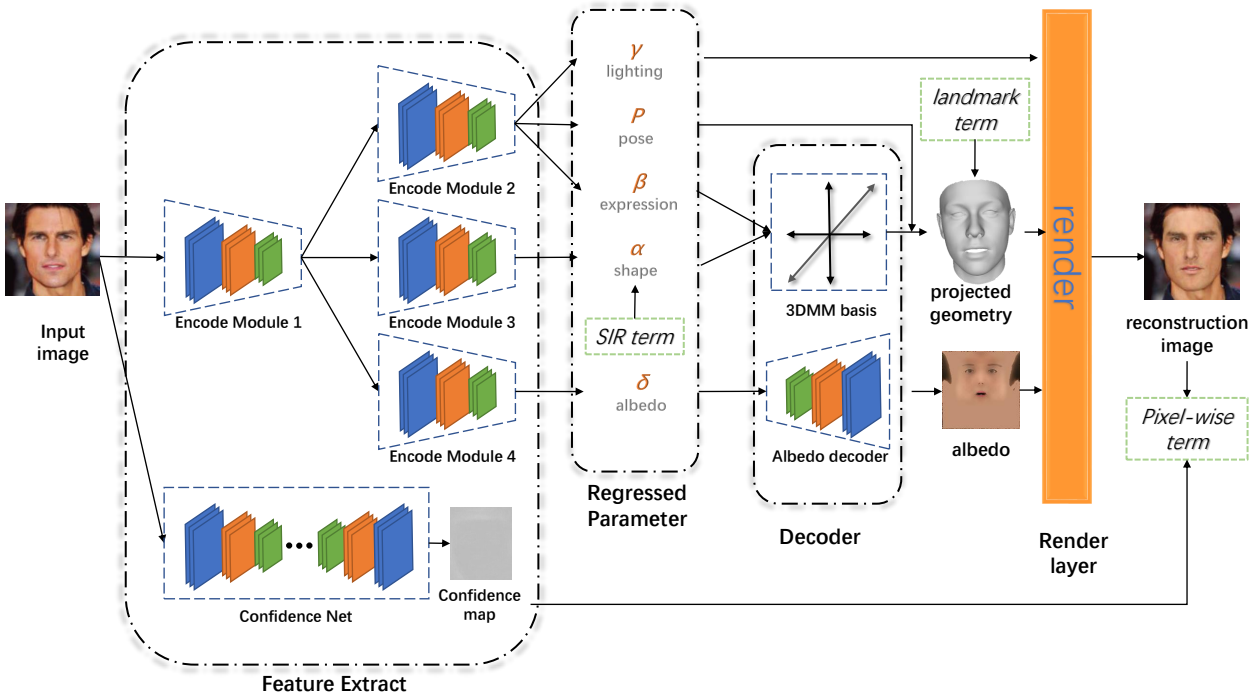

Figure 2. **The framework of our method.** Our network contains a feature extraction module (Encoder Module 1) followed by three encoders (Encoder Modules 2, 3 and 4) sharing the same weights of Encoder Module 1. The confident network estimates the confidence map, which represents the aleatoric uncertainty of the model during training. By switching the SIR item and the landmark item, even incompletely labeled data can effectively train the network.

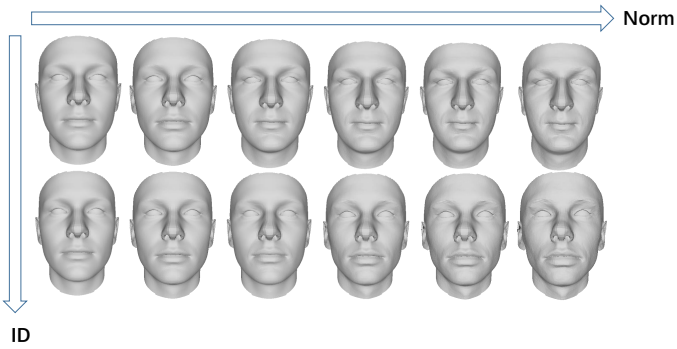

Figure 3. From left to right in each row, the shape parameter is multiplied by 0.1, 0.4, 0.6, 0.8, 1.0 and 1.2. The face shapes in different rows represent different people.

Considering that $\mathbf{A}$ is the orthonormal basis, the relationship between E and e is as shown in Equation(6), which means that geometric Euclidean distance is proportional to the parameter Euclidean distance.

$$E = \frac{1}{n}\|x - y\|_2^2 = \frac{m}{n}e \tag{6}$$

This equation proves that when we minimize (maximize) the Euclidean distance between the parameters of face shapes of the same person (different people), their corresponding geometric distances are smaller (larger) as well, which suggests that it is feasible to formulate recognition errors using the Euclidean distance in addition to the cosine distance (e.g. softmax-like loss functions).

From shape geometry discrimination to visual discrimination. As mentioned above, in Euclidean space, separation of the shape parameters ensures separation of the shape geometries. However, even though the shape geometries can be separated numerically, we cannot ensure that the separation is visually recognizable, since people usually fail to perceive small differences between meshes. As shown in 3, in each column, shape parameters are multiplied by different factors, which does not influence the separation of parameters and geometries numerically. However, even though the shapes in each column have the same numerical separation of geometry, we find it difficult to visually distinguish face shapes when the parameter norms are relatively small, implying that the same geometry discrimination could have various degrees of visual discrimination. We find that when only the above two losses occur, the network has a high probability of falling into a local minimum and resembling an average face, when the norm of regressed shape parameters is minimal. This observation suggests that in addition to the individual shape geometry, we also need to pay attention to adding additional constraints to make the shape geometries visually distinguishable. According to 4 and 6, the norm of the shape parameter $\|\alpha\|_2$ is proportional to $\|\mathbf{A}\alpha\|_2$, the residual between facial geometry S and mean face $\bar{S}$. Therefore, by constraining the parameters to fit an appropriate distribution, the geometric residuals of faces can be sufficiently large to make the shapes intuitively distinguishable.

The 3DMM is based on assuming that the data (face vertices) follow a multivariate normal distribution. This as-

sumption is a necessary prerequisite of principal component analysis (PCA). The multivariate normal distribution of shape parameters is given by[1]:

$$p(\alpha) \sim \exp[-\frac{1}{2}\sum_{i=1}^{m}(\alpha_i/\sigma_i)^2] \tag{7}$$

where σ_i is the eigenvalues of the shape covariance matrix. If our trained model outputs shape parameters that have the same distribution as in Equation(7), the reconstructed face shape should share the same level of visual distinctiveness with the scan data used to build the 3DMM. According to Equation(7), the parameters divided by the eigenvalue follows the standard normal distribution. We minimize the KL divergence between the parameters divided by the eigenvalues and the standard normal distribution function to constrain the parameter to fit the distribution described in Equation(7).

$$\underset{\theta}{\operatorname{argmin}} KL(\boldsymbol{P}(\boldsymbol{\alpha}/\boldsymbol{\sigma} \mid \boldsymbol{I}, \boldsymbol{\theta}) \| \boldsymbol{N}(0,1)) \tag{8}$$

where I is the input image and θ is the weight of the network. Once the shape parameters fit a multivariate Gaussian distribution, the norm is sufficiently large to represent visually discriminative face shapes.

From parameter discrimination to visual discrimination. To reconstruct a visually recognizable face shape, the reconstructed 3D face shapes from images with different lighting, expressions and poses should be the same as the one from the neutral frontal face image of the same person. In the parameter domain, shape parameters of the same identity are tightly grouped around the parameter regressed from the neutral frontal face image. To meet this condition, we modify the center loss[36] to push shape parameters of the same identity towards the center of its class and update the center by assigning higher weights to the samples with smaller expression and pose variance. Therefore, the parameters regressed from the neutral frontal face image have a more significant impact on the class center, and the shape parameters from nonneutral and nonfrontal faces will approach those from the neutral frontal face using the modified center loss. In summary, to transfer parameter discrimination to visual discrimination, the shape parameters should satisfy the following conditions: (1) the shape parameters are discriminative in Euclidean space; (2) shape parameters follow a specific multivariate Gaussian distribution; and (3) the centers of the shape parameters are the parameters regressed from the neutral face image of the identity.

C. Network Structure

Our network has four branches to regress the albedo parameters, shape parameters, confidence map and other non-identity information (expression parameters, camera parameters and illumination parameters). We use the same residual block(Sphere64a) used in SphereFace[37]. As shown in Figure 2, Encode Module 2, 3 and 4 contain last two blocks(Conv3.x,Conv4.x) of Sphere64a, followed by a multilayer perceptron(MLP). They share the weights of Encode Module 1, which contains the first two blocks(Conv0.x,Conv1.x) of Sphere64a. Encode Module 1 serves as the feature extraction module, and Encode Modules 2-4 serve as the feature separation modules. The feature extraction module shares its low-level features with the feature separation modules, reducing the number of parameters of the whole network. Encode Modules 2-4 improve the network ability to separate high-level features. By balancing the depth of the feature extraction module and that of the feature separation module, the network can get obtain better results with fewer network parameters. The MLP adapts the output features to the size of our parameters (199-dim of the shape parameter, 29-dim of the expression parameters, 7-dim of the camera parameters, 27-dim of the illumination parameters and 512-dim of the albedo parameters). We use the same ConfNet structure to generate confidence map as in [38]. The decoder of the albedo consists of a transposed convolution network and regresses the albedo UV map with a resolution of 256×256. We regard the 3DMM basis as the fully connected layer in the neural network. We build the rendering layer based on Pytorch3d implementation, and the illumination model is a spherical harmonic illumination model.

D. Loss function

Our loss function contains three terms: a landmark term, a pixelwise photometric term and a SIR regularization term. According to the existing labels of the training samples, we determine which terms can take effect. For example, if the training sample has identity labels, the SIR term and pixelwise photometric term will be enabled. Otherwise, the face reconstruction term and pixelwise photometric term will take effect. All mentions of εs in this sub-section represent the weight of the loss function.

$$L = \begin{cases} \varepsilon_l L_{land}(F_\theta(I)) + L_{pix}(F_\theta(I)) & I \in S_{recon} \\ \varepsilon_s L_{SIR}(F_\theta(I)) + L_{pix}(F_\theta(I)) & I \in S_{id} \end{cases} \tag{9}$$

In the above equation, L_{land} denotes the landmark term, L_{SIR} denotes the SIR term and L_{pix} denotes the pixelwise photometric term. S_{recon} is the face reconstruction dataset and S_{id} is the face recognition dataset. I, F and θ denote the input image, the network and the weights of the network respectively. In the rest of this section, we introduce the three loss terms in detail.

Landmark term The landmark term L_{proj} simply uses the L_2 loss between projected landmarks $\hat{V}_{2d}$ and ground-truth landmarks V_{2d}.

$$L_{land} = \frac{1}{N}\left\|V_{2d} - \hat{V}_{2d}\right\|_2 \tag{10}$$

where N is the number of landmarks.

Pixelwise photometric term Pixelwise photometric term measures the reconstruction errors by both the pixel loss and perceptual loss on the confidence map[38].

The confidence map aims to achieve robustness when occlusions and other challenging appearance variations exist such as beard and hair. The pixel loss is defined as follows:

$$L_{pixel}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) = -\frac{1}{|\Omega|}\sum_{uv\in\Omega} \ln \frac{1}{\sqrt{2}\sigma_{uv}} \exp -\frac{\sqrt{2}\ell_{1,uv}}{\sigma_{uv}} \tag{11}$$

where $\ell_{1,uv} = \left|\hat{\mathbf{I}}_{uv} - \mathbf{I}_{uv}\right|$ is the L_1 distance between the intensity of input image $\mathbf{I}$ and the reconstruction image $\hat{\mathbf{I}}$ at location (u, v) and $\sigma \in \mathbb{R}_{+}^{W \times H}$ is the confidence map. The perceptual loss mitigates the blurriness in the reconstruction result, which is defined as follows:

$$L_{pr}^{(k)}\left(\hat{\mathbf{I}}, \mathbf{I}, \sigma^{(k)}\right) = -\frac{1}{|\Omega_k|} \sum_{uv \in \Omega_k} \ln \frac{1}{\sqrt{2\pi}\sigma_{uv}^{(k)}} \exp - \frac{\left(\ell_{uv}^{(k)}\right)^2}{2\left(\sigma_{uv}^{(k)}\right)^2} \tag{12}$$

where $\ell_{uv}^{(k)} = \left|e_{uv}^{(k)}(\hat{\mathbf{I}}) - e_{uv}^{(k)}(\mathbf{I})\right|$ is the L_1 distance between feature maps of the k-th layer. $e^{(k)}(\mathbf{I}) \in \mathbb{R}^{C_k \times W_k \times H_k}$ is the k-th layer of an off-the-shelf image encoder e(VGG16[39]) and $\Omega_k = \{0, \ldots, W_k - 1\} \times \{0, \ldots, H_k - 1\}$ is the corresponding spatial domain. $\sigma^{(k)}$ is a confidence map of perceptual loss.

As shown in Equation(13), the pixelwise term consists of two losses: L_{recon} and L_{reg}.

$$L_{pix} = L_{recon} + \varepsilon_{reg} L_{reg} \tag{13}$$

L_{recon} measures the reconstruction errors by both the pixel loss and perceptual loss. L_{reg} avoids overfitting when predicting 3DMM parameters and albedo UV maps, which is defined by:

$$L_{reg} = L_{regp} + \varepsilon_{rega} L_{rega} \tag{14}$$

The regularization term of L_{regp} for 3DMM coefficients is:

$$L_{regp} = \varepsilon_{id} \| \frac{\alpha_{id}}{\sigma_{id}} \|^2 + \varepsilon_{exp} \| \frac{\alpha_{exp}}{\sigma_{exp}} \|^2 \tag{15}$$

where σ_{id} is the eigenvalue of the shape basis and α_{exp} is the eigenvalue of the expression basis. The regularization of albedo UV maps consists of smooth and residual terms, which penalize differences with neighboring pixels and enforce a prior distribution towards the mean albedo to avoid the regressed albedo being away from the mean albedo.

$$L_{rega}(\mathbf{A}) = \sum_{\mathbf{p}_i^{\mathrm{uv}} \in \mathbf{A}^{\mathrm{uv}}} \left\| \mathbf{A}^{\mathrm{uv}}\left(\mathbf{p}_i^{\mathrm{uv}}\right) - \frac{1}{|\mathcal{N}_i|} \sum_{\mathbf{p}_j^{\mathrm{uv}} \in \mathcal{N}_i} \mathbf{A}^{\mathrm{uv}}\left(\mathbf{p}_j^{\mathrm{uv}}\right) \right\|_2 + \varepsilon_{uv} \left\| \mathbf{A}^{\mathrm{uv}} \right\|_2^2 \tag{16}$$

where $\mathbf{A}^{\mathrm{uv}}$ is the albedo UV map and $\mathcal{N}_i$ denotes a set of A 4-pixel neighborhood of pixel $\mathbf{p}_i^{\mathrm{uv}}$.

Shape identity-aware regularization term

The SIR term includes two components as shown in Equation(17), an identification loss and a Kullback-Leibler Loss.

$$L_{id} = L_{recog} + \varepsilon_{kl} L_{kl} \tag{17}$$

To ensure the criteria that shape parameters are discriminative in Euclidean space, the identification loss is defined as in Equation(18). It combines softmax-like loss and center loss[36].

$$L_{recog} = L_{sm} + \varepsilon_c L_c \tag{18}$$

L_{sm} is a softmax-like loss (e.g. softmax, Cosloss[40], A-softmax[37], and Arcloss[41]), which separates parameters and speeds up the convergence and L_c discriminates features in Euclidean space (e.g triplet loss and center loss). We choose Cosloss [40] as our softmax-like loss.

To ensure the condition that the centers of shape parameters of the same identity are the parameters regressed from neutral frontal face images, we first calculate the confidence, which indicates similarity to the neutral frontal face.

$$f = \frac{1}{8}(cos\alpha + 1)(cos\beta + 1)(cos\gamma + 1) \cdot \exp^{-\lambda \|\alpha_{exp}\|_2} \tag{19}$$

where α_{exp} represents the expression parameters. α, β and γ are Euler angles of 3D face poses.

We use the following formula to update the centers. It assigns higher weights with the neutral frontal face.

$$\Delta c_j = \frac{\sum_{i=1}^{n} \delta(y_i = j) \cdot (c_j - x_{id_i})}{1 + \sum_{i=1}^{n} \delta(y_i = j)} \cdot f \tag{20}$$

where n is the number of samples in a mini-batch, $\delta(\cdot) = 1$ if the condition is true and $\delta(\cdot) = 0$ otherwise. y_i is the identity label of the sample, c_j is the shape parameter center of the j-th class, and α_{id} represents the shape parameters.

To ensure that the shape parameter satisfies a specific distribution, we use the Kullback-Leibler loss to constrain the shape parameters to fit a zero-mean multivariate Gaussian distribution with the eigenvalues as its variances.

$$L_{kl} = KL(\boldsymbol{P}(\boldsymbol{\alpha}/\boldsymbol{\sigma} \mid \boldsymbol{I}, \boldsymbol{\theta}) \| \boldsymbol{N}(0, 1)) \tag{21}$$

E. Training strategy

The public databases usually contain either face labels or landmark labels. Some existing works need to use face detectors or optimization-based methods to generate the annotation needed for face reconstruction. However, these annotation is unsatisfactory in challenging examples and limits the performance of models by those algorithms. Based on the above considerations, we choose to build a new dataset with a mixture of face recognition and facial landmark data. Directly training our network on this mixed dataset with the different labels results in tricky convergence for the following reasons: (1) The numbers of samples of face recognition and face alignment are unbalanced. (2) Incomplete labels can result in an oscillating learning process. (3)The objective function is complicated, making our network easily fall into a local minima without good initialization. Therefore, one must important to warm up the network and maintain a balanced proportion of face recognition and face reconstruction data in the mixed database. The warming-up stage consists of two steps. First, we train our network on the 300W-LP[4] database without SIR loss. Second, we train the whole network on the mixed database and add the SIR loss. The mixed database consists of VGGFace2[42] and the 300W-LP[4]. 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. We use only the 300W-LP landmark label because the synthetic face shape is not precise. Considering taht the sample numbers of the two databases are extremely unbalanced, we design a sampling

scheme in which the probability of selecting samples from the face recognition database is given by:

$$P = \frac{N_{recon}}{N_{recog} + N_{recon}} \tag{22}$$

where N_{recon} is the number of samples in the face reconstruction dataset and N_{recog} is that in the face recognition dataset. The probability of selecting samples from the face reconstruction database is $1 - P$.

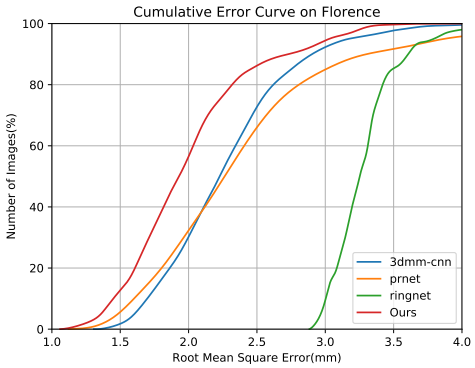

Figure 4. Cumulative error distribution(CED) curve on Florence dataset. We compare our method with Tran et al.[5], Ringnet[6] and MGCNet[44].

Losses L_{sm}	L_c	L_{wc}	LFW	CFP-FP	YTF
			67.67	54.16	66.46
✓			90.55	70.77	81.48
✓	✓		**95.23**	**83.45**	**89.10**
✓		✓	94.47	80.73	86.40

Table II
THE FACE VERIFICATION ACCURACY(%) ON LFW, CFP-FP AND YTF WITH FOR DIFFERENT LOSSES OF THE SHAPE IDENTITY-AWARE REGULARIZATION TERM. L_{sm} IS THE COSLOSS[40]. L_c IS THE CENTER LOSS[36], AND L_{mc} IS THE WEIGHTED CENTER LOSS.

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
Ours	**95.36**	**83.34**	**89.07**
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
Ours	**94.47**	**80.78**	**86.40**

Table III
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[5], LIU ET AL.[7], D3FR[32], TDDFA[48], MGCNET[44] AND RINGNET[6].

IV. EXPERIMENTAL RESULTS

To evaluate the effectiveness of SIR losses, we first compare our method with previous methods by measuring their face recognition accuracies with the generated shape parameters. We then test their face reconstruction errors. Finally, we qualitatively evaluate the visualized reconstruction results.

A. Face Recognition Performance

We design an ablation study to investigate the impact of SIR losses on face recognition performance and compare it with other methods based on the 3DMM models. First, we introduce the test datasets and evaluation method.

Testing benchmarks We use the following datasets: (1) LFW[45], a standard face verification testing dataset. It contains 13,233 labeled face images for 5,749 different individuals with a total of 6000 defined pairs. (2) CFP, the Celebrities Frontal-Profile dataset[46]. It is aimed at evaluating face identification with frontal and profile pairs and has approximately 7,000 pairs of matches defined by 3,500 same pairs and 3,500 not-same pairs for approximately 500 different subjects. (3)YTF, Youtube face dataset[47]. It contains 3,425 videos of 1,595 individuals. We follow the verification protocol and report the result on 5000 video pairs.

Evaluation method In the methods designed for the face recognition task, the identity of a subject in an image can be represented as learned latent codes. The similarity between two identity representations (usually based on the cosine distance or Euclidean distance) is calculated to determine whether the images are of the same person. In our evaluation, the shape parameters can be used for identity representation, similar to the latent codes in other face recognition methods. This evaluation method is not suitable for evaluating shape parameter discrimination because the Euclidean distance between parameters can reflect the separation of the face geometry while the cosine distance cannot. Therefore, we directly use the Euclidean distance between 3DMM shape parameters to measure the similarity between two faces. To be fair, we also show the results of using the cosine distance.

Results on testing benchmarks As mentioned above, the shape parameters must be able to minimize the intraclass distance and maximize the interclass distance in Euclidean space. To evaluate the effectiveness of SIR loss when learning the discriminative shape parameters, we test the face recognition performance of the 3DMM shape parameters on LFW, CFP-FP and YTF.

Ablation study To validate the efficiency of each loss in our proposed SIR term, we test face verification under various loss combinations. As shown in Table II, the face verification accuracy is very low without any SIR loss terms. When we added center loss in addition to the cosloss, the accuracy increases significantly because our evaluation method is based on the Euclidean distance between parameters, while the center loss can reduce the Euclidean distance between parameters belonging to the same class. The difference in face verification accuracy between the center loss and weighted center loss is subtle because the weighted center loss makes only the class center closer to its neutral frontal face parameters. It updates the center by assigning higher weights to the neutral frontal face parameters. This operation does not significantly impact

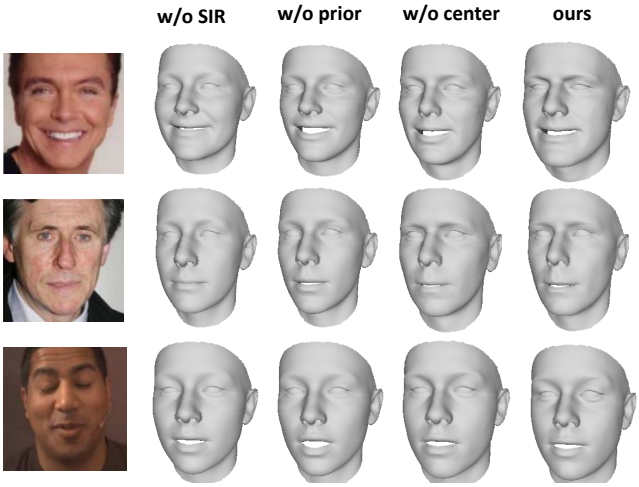

Figure 5. The ablation study of SIR loss terms. 'w/o' SIR means that we do not use the SIR term in training. 'w/o' prior means that we do not use the KL loss in training. 'w/o' center means that we do not use the weighted center loss in training.

Representation	LFW	CFP	YTF
Parameter	94.47	80.73	86.40
Vertices	94.72	80.71	86.40

Table IV
THE RESULTS OF FACE VERIFICATION(%) USING DIFFERENT IDENTITY REPRESENTATIONS. WE USE SHAPE PARAMETERS AND SHAPE VERTICES AS THE IDENTITY REPRESENTATION.

the effect of face recognition but benefits face reconstruction, as shown in Table V.

Table III shows a comparison between our results and those of other methods on LFW, CFP-FP and YTF. Note that the other methods may use the cosine distance to measure the parameter similarity. However, the Euclidean distance between parameters can better reflect the difference between geometries and is thus more appropriate to use Euclidean distance. For a fair comparison, we also show the result of our method with the cosine distance.

To demonstrate that the SIR loss can transfer parameter discrimination to geometric discrimination, we compare with face verification on LFW, CFP and YTF with the shape parameters and reconstructed vertices as the identity representation. Table IV shows that the discriminative property successfully transfers from the parameter space to the geometric space. However, we use the BFM model provided by 3DDFA, which deletes some vertices on the original BFM model from PCA. Thus, the shape basis is not strictly orthonormal and the face verification accuracy of vertices has little difference with the parameter.

B. Quantitative Results on Shape Reconstruction

To evaluate the stability of our algorithm and the precision of the reconstructed 3D face shapes, we calculate the RMSE between neutral frontal 3D scans and 3D shape faces regressed from images under various conditions, including illumination, head pose, expression, and occlusion, on the Florence dataset[49].

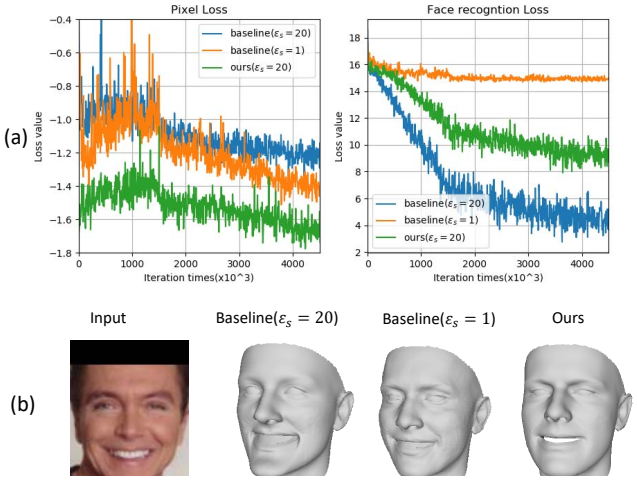

Figure 6. Comparisons to baseline models for feature separation and training convergence. (a) shows that how the pixel loss(Equation(11)) and face recognition loss(Equation(18)) change with the number of training iterations in the second stage of training. ε_s is used in Equation (9). The weights of the other losses remain the same. (b) shows the visualized reconstructed faces.

Test on the Florence dataset The MICC dataset contains the 2D/3D faces of 53 subjects, including two indoor videos, one outdoor video and the faces' 3D models. The 3D face model of each person includes one or two frontal faces with a neutral expression. Unconstrained outdoor videos are recorded under natural lighting conditions, which are more challenging. In our experiment, we choose the outdoor video frames as the input and randomly select 100 frames of each subject to form a test dataset that contains 5,300 face images. In our preprocessing stage, the face and its landmarks are detected by the MTCNN[50]. Afterward, the faces are aligned using similarity transformation and cropped to 112×96 in the RGB format. The ground-truth scans are cropped at a radius of 95 mm around the nose tip, and the meshes generated are aligned to the ground truth using ICP with an isotropic scale.

When testing, we reconstruct only the shape of the face without expressions. Table V reports the RMSE of the point-to-plane distance between the ground-truth face shape and reconstructed face shape after ICP on an isotropic scale. Figure 4 shows the cumulative error distribution(CED) curves of different methods.

As shown in V, the reconstruction error of the weighted center loss is lower than that of the center loss. The reason is that the weighted center loss pushes the regressed shape parameter to the values regressed from the corresponding neutral frontal face image. The parameters obtained from neutral frontal face images are more accurate than those of profiles and faces with extreme expressions. Therefore, using neutral frontal face parameters as the class center can improve the accuracy of reconstruction accordingly. The RMSE of our method on the Florence dataset is also lower than that of other state-of-the-art methods. Note that some results are inconsistent with the results reported in their paper, since they ran their methods

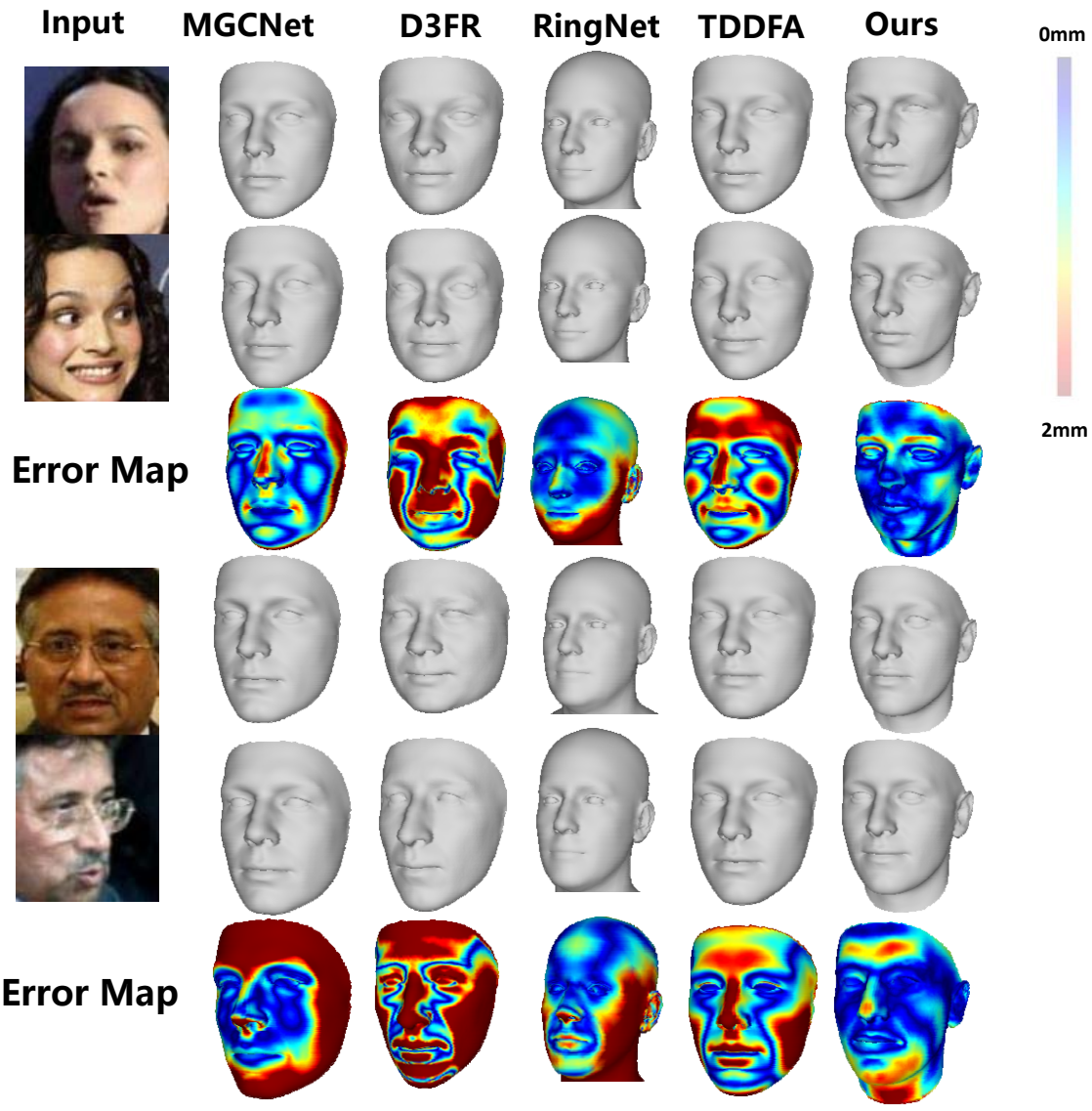

Figure 7. Comparison of our qualitative results under various levels of illumination, various facial expressions, large poses, and occlusion with MGCNet[44], D3FR[32], TDDFA[48] and RingNet[6] on the LFW dataset. We use only shape parameters to reconstruct the face geometries; thus, normalization occurs without expression and pose effects. The error maps reveal the Euclidean distance between two map shapes.

on each frame of Florence's videos and averaged each video's results to obtain a single reconstruction. Since the averaging operation prevents the results from showing the reconstruction stability, we evaluate those methods using the same approach as presented above.

C. Qualitative Results on Shape Reconstruction

Ablation Study Figure 5 shows reconstructed face shapes with different combinations of losses. As presented, some identity details are missing without the SIR item. If we use identity loss without a distribution prior, the identity discrimination in the parameter space cannot be effectually transferred to the appearance space since the norm the shape parameters would be small, leading to the reconstruction of a mean face. The weighted center loss aggregates shape parameters to the frontal face in Euclidean space. It improves the reconstruction robustness to different face poses and facial expressions. The last column of Figure 5 shows that the weighted center loss helps strengthen the identity discrimination.

We compare our method with the baseline model, in which the original structure of SphereFace[37] is utilized. As shown in Figure 6, the face recognition loss of the baseline model drops faster than ours using the same loss weights, but their pixel loss is much higher than ours. The visualization results of the reconstructed faces are not satisfactory. If we reduce the weight of the SIR loss to improve the face reconstruction performance, the face recognition loss of the baseline model does not converge, and the reconstructed face recognizability is also weakened.

To evaluate our reconstructed face's stability and visual identifiability under challenging conditions such as various types of illumination, large poses, various expressions and occlusion, we compare the qualitative results of estimating the face shape from a single image in the LFW database obtained

Method	Tran et al.	Lui et al.	MGCNet	D3FR
RMSE	2.27	2.00	1.94	1.82
Method	RingNet	TDDFA	Ours-lc	Ours-lwc
RMSE	1.84	2.01	1.82	**1.80**

Table V
THE FACE RECONSTRUCTION ERROR OF FLORENCE DATASET. WE COMPARE OUR METHOD WITH TRAN ET AL.[5], LUI ET AL.[7] AND MGCNET[44], D3FR[32], TDDFA[48] AND RINGNET[6]. OUR-LC REPRESENTS THAT WE USE CENTER LOSS AND OUR-LWC REPRESENTS WEIGHTED CENTER LOSS.

with MGCNet[44], D3FR[32], TDDFA[48] and RingNet[6]. As Figure 7 shows, we choose images under four conditions: various levels of illumination, various facial expressions, large poses and occlusion.

Occlusion: As shown at the top of Figure 7, the woman's hair occludes her cheeks. Thus, we cannot directly infer the shape around the cheeks from the picture. Therefore, different orbital geometries can easily be regressed if the constraints of face recognition are not used, such as the result of MGCNet[44]. The SIR term can aggregate the same person's features and infer the geometric information of the occluded part.

Expression: The final geometry is affected by both of the face shape and its expression, which means that the same face geometry can be determined by different combinations of face shapes and expressions. The same person with different expressions may regress varying face shapes. As shown at the top of Figure 7, for the case of smiling, the other methods have some errors in the mouth area, while our reconstruction results remain stable.

Pose: Large poses result in some information loss regarding the face shape due to self-occlusion. As shown at the bottom of Figure 7, the face contour is difficult to estimate. However, the SIR term can push the regression parameters from the profile to the parameters regressed from the frontal face and infer the missing information.

Illumination: As shown in the top row of Figure 7, the contour of a face could be unclear due to inappropriate illumination. Similar effects could be caused by some special poses or occlusion, where some shape information is lost. SIR loss can infer the lost information for the reason described above.

In addition, the reconstructed face shapes of the same person in different environments should be the same. The face shapes reconstructed for different people, however, need to differ from each other. Figure 8 shows the difference in the reconstructed shapes from different people. It shows that the face shapes reconstructed by other methods are similar. In contrast, our result presents the expected differences.

Figure 9 shows a qualitative comparison between our method and other state-of-the-art methods. Different from RingNet[6], TDDFA[48], our results maintain the identity feature of rhe input images. MGCNet[44] and D3FR[32] use photometric metrics and can effectively capture the identity feature of the input image. However, they produce poor results when the faces have extreme expressions and large poses as shown in the second and fourth columns in Figure 9). In contrast, our method can capture identity features under challenging conditions due to the benefits from our identity losses.

V. CONCLUSIONS

Our research started from the observation that despite the high face recognition accuracy obtained using the 3DMM shape parameter, the reconstructed 3D face shapes are lack of significant visual discrimination. We first explored the relationship between the between 3DMM parameter space and 3D geometric space, and propose SIR losses that explicitly enforce shape consistency in the shape parameter space while implicitly guiding reconstructed face shapes to be visually discriminative. In detail, the identification loss explicitly maximizes the interclass and minimizes the intraclass Euclidean distance of shape parameters while it implicitly maximizes/minimizes the MSE of the shape geometry of different people/the same person. Kullback–Leibler losses are also utilized to explicitly constrain the shape parameters to follow a particular distribution and implicitly let them to share the same visual distinction as the shapes used to train the 3DMMs. We build a neural network and an associated training strategy to cope with the lack of such a dataset that contains both identity and 3D geometry annotations, which can quickly converge under our training strategy. The experiments show that our results outperform those of the state-of-the-art methods in terms of the reconstruction error, visual discrimination, and face recognition accuracy.

REFERENCES

[1] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." 1999.

[2] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[3] 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*. Ieee, 2009, pp. 296–301.

[4] 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*, 2015, pp. 787–796.

[5] 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*, 2017, pp. 5163–5172.

[6] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7763–7772.

[7] 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*, 2018, pp. 5216–5225.

[8] V. Kazemi, C. Keskin, J. Taylor, P. Kohli, and S. Izadi, "Real-time face reconstruction from a single depth image," in *2014 2nd International Conference on 3D Vision*, vol. 1. IEEE, 2014, pp. 369–376.

[9] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3d face rigs from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, p. 28, 2016.

[10] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: face model learning from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 812–10 822.

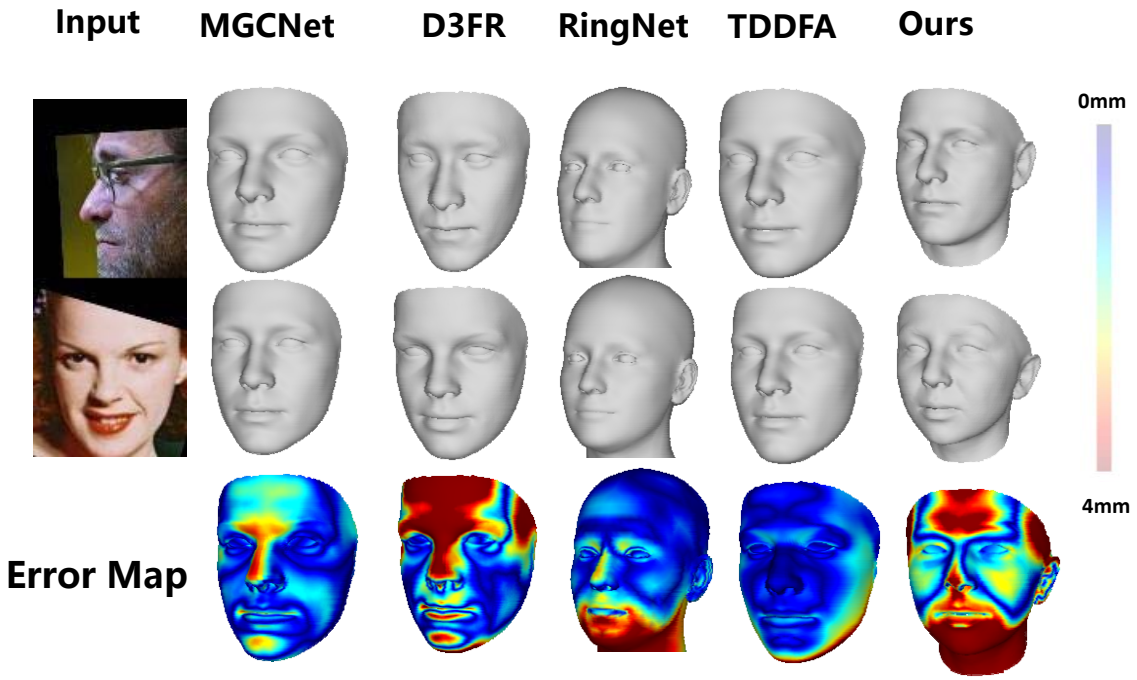

Figure 8. Comparison of differences between two face shapes, which are regressed from images of different people. The last row presents the error maps, which reveal the difference between the two face shapes by the MGCNet[44], D3FR[32], TDDFA[48] and RingNet[6]. We use only shape parameters to reconstruct the face geometries; thus, normalization occurs without expressions and pose effects.

[11] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on graphics (TOG)*, vol. 33, no. 4, p. 43, 2014.

[12] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler, "A multiresolution 3d morphable face model and fitting framework," in *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[13] S. Saito, T. Li, and H. Li, "Real-time facial segmentation and performance capture from rgb input," in *European Conference on Computer Vision*. Springer, 2016, pp. 244–261.

[14] J. Roth, Y. Tong, and X. Liu, "Adaptive 3d face reconstruction from unconstrained photo collections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4197–4206.

[15] M. Piotraschke and V. Blanz, "Automated 3d face reconstruction from multiple images using quality measures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3418–3427.

[16] J. Roth, Y. Tong, and X. Liu, "Unconstrained 3d face reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2606–2615.

[17] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1259–1268.

[18] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1274–1283.

[19] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3d face reconstruction from a single image via direct volumetric cnn regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1031–1039.

[20] 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*, 2018, pp. 2549–2559.

[21] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives." *ACM Trans. Graph.*, vol. 32, no. 4, pp. 42–1, 2013.

[22] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, p. 40, 2013.

[23] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li, "High-fidelity facial reflectance and geometry inference from an unconstrained image," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 162, 2018.

[24] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3d face reconstruction with deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5908–5917.

[25] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[26] Y. Guo, J. Cai, B. Jiang, J. Zheng *et al.*, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 6, pp. 1294–1307, 2018.

[27] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 460–469.

[28] 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*, vol. 41, no. 1, pp. 78–92, 2017.

[29] 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*, 2020, pp. 4958–4967.

[30] S. Liu, T. Li, W. Chen, and H. Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7708–7717.

[31] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in

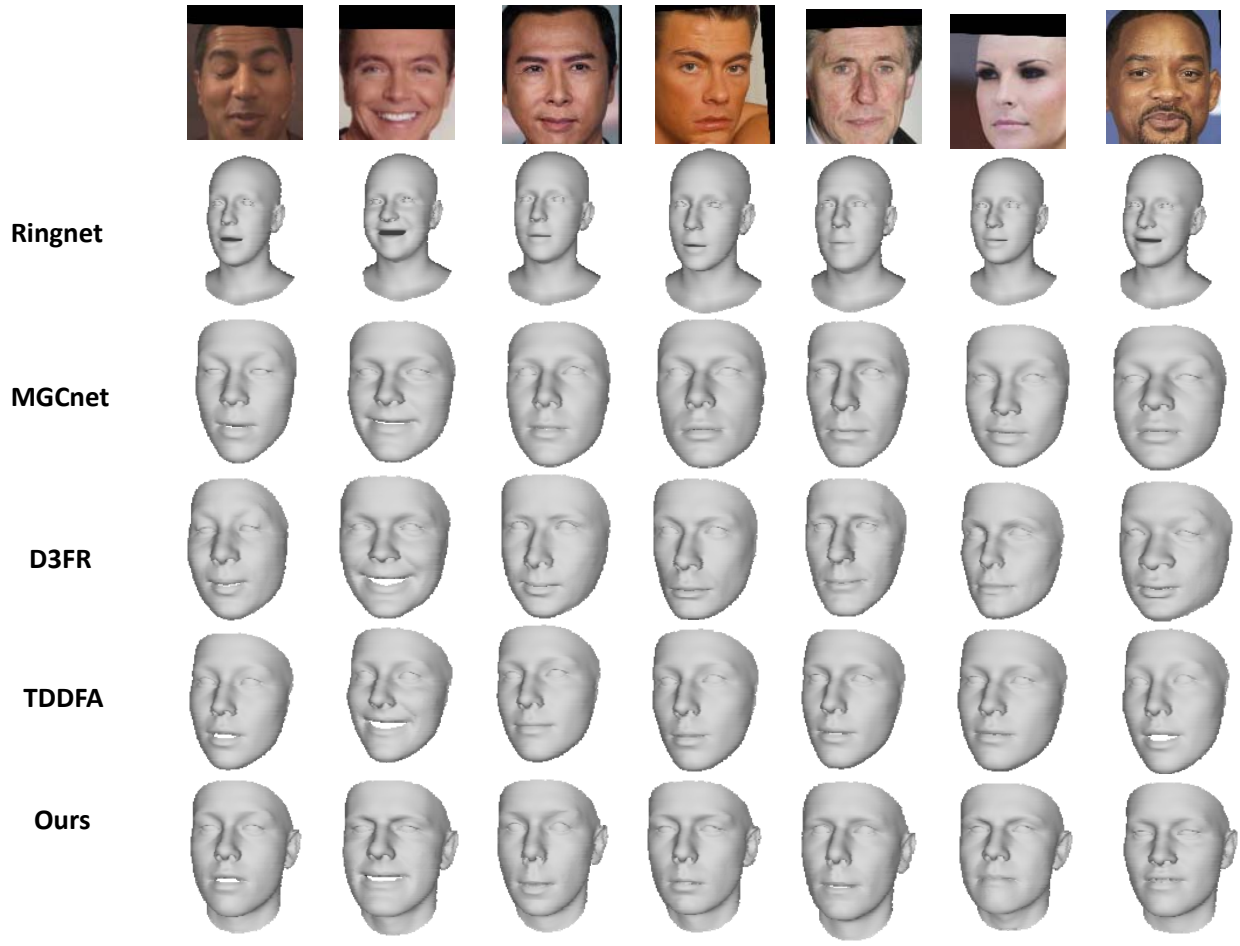

Figure 9. Comparison of our qualitative results with RingNet[6], MGCNet[44], D3FR[32] and TDDFA[48].

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3907–3916.

[32] 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 Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[33] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1126–1135.

[34] 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)*, 2018, pp. 534–551.

[35] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[37] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[38] 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*, 2020, pp. 1–10.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] 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*, 2018, pp. 5265–5274.

[41] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[42] 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)*. IEEE, 2018, pp. 67–74.

[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*, vol. 47, pp. 3–18, 2016.

[44] 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.

[45] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[46] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[47] L. Wolf, T. Hassner, and I. Maoz, *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011.
[48] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[49] 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*. ACM, 2011, pp. 79–80.
[50] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.