Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition*

Feng Liu[1], Ronghang Zhu[1], Dan Zeng[1], Qijun Zhao[1,†], and Xiaoming Liu[2]
[1]College of Computer Science, Sichuan University
[2]Department of Computer Science and Engineering, Michigan State University

Abstract

This paper proposes an encoder-decoder network to disentangle shape features during 3D face reconstruction from single 2D images, such that the tasks of reconstructing accurate 3D face shapes and learning discriminative shape features for face recognition can be accomplished simultaneously. Unlike existing 3D face reconstruction methods, our proposed method directly regresses dense 3D face shapes from single 2D images, and tackles identity and residual (i.e., non-identity) components in 3D face shapes explicitly and separately based on a composite 3D face shape model with latent representations. We devise a training process for the proposed network with a joint loss measuring both face identification error and 3D face shape reconstruction error. To construct training data we develop a method for fitting 3D morphable model (3DMM) to multiple 2D images of a subject. Comprehensive experiments have been done on MICC, BU3DFE, LFW and YTF databases. The results show that our method expands the capacity of 3DMM for capturing discriminative shape features and facial detail, and thus outperforms existing methods both in 3D face reconstruction accuracy and in face recognition accuracy.

1. Introduction

3D face shapes reconstructed from 2D images have been proven to benefit many tasks, e.g., face alignment or facial landmark localization [18, 41], face animation [9, 13], and face recognition [5, 12]. Many prior work have been devoted to reconstructing 3D face shapes from a single 2D image, including shape from shading (SFS)-based methods [14, 20], 3D morphable model (3DMM) fitting-based methods [4, 5], and recently proposed regression-based methods [22, 23]. These methods mostly aim to recover 3D face shapes that are loyal to the input 2D images or retain as much facial detail as possible (see Fig. 1). Few of them explicitly consider the identity-sensitive and identity-irrelevant features in the reconstructed 3D faces. Consequently, very few studies have been reported about recognizing faces using the reconstructed 3D face either by itself or by fusing with legacy 2D face recognition [5, 33].

Using real 3D face shapes acquired by 3D face scanners

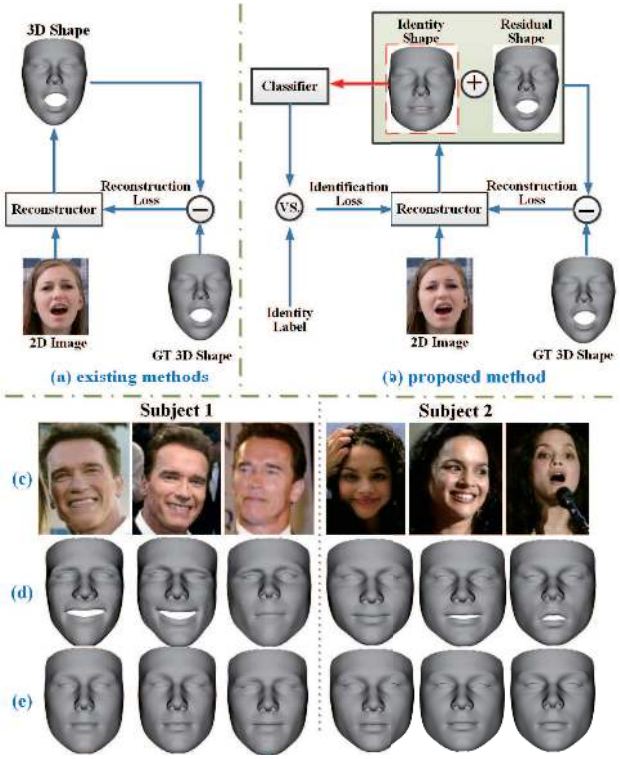


Figure 1. Comparison between the learning process of (a) existing methods and (b) our proposed method. GT denotes Ground Truth. (d) and (e) are 3D face shapes and disentangled identity shapes reconstructed by our method for the images in (c) from LFW [15].

*This work is supported by the National Key Research and Development Program of China (2017YFB0802300) and the National Natural Science Foundation of China (61773270, 61703077).

†Corresponding author. Email: qjzhao@scu.edu.cn.

for face recognition, on the other hand, has been extensively studied, and promising recognition accuracy has been achieved [6, 11]. Apple recently claims to use 3D face matching in its iPhone X for cellphone unlock [1]. All of these prove the discriminative power of 3D face shapes. Such a big performance gap between the reconstructed 3D face shapes and the real 3D face shapes, in our opinion, demonstrates that existing 3D face reconstruction methods seriously undervalue the identity features in 3D face shapes. Taking the widely used 3DMM fitting based methods as example, their reconstructed 3D faces are constrained in the limited shape space spanned by the pre-determined bases of 3DMM, and thus perform poorly in capturing the features unique to different individuals [39].

Inspired by the latest development in disentangling feature learning for 2D face recognition [26, 34], we propose to disentangle the identity and non-identity components of 3D face shapes, and more importantly, fulfill *reconstructing accurate 3D face shapes* loyal to input 2D images and *learning discriminative shape features* effective for face recognition in a *joint* manner. These two tasks, at the first glance, seem to contradict each other. On one hand, face recognition prefers identity-sensitive features, but not every detail on faces; on the other hand, 3D reconstruction attempts to recover as much facial detail as possible, regardless whether the detail benefits or distracts facial identity recognition. In this paper, however, we will show that by exploiting the 'contradictory' objectives of recognition and reconstruction, we are able to *disentangle identity-sensitive features from identity-irrelevant features in 3D face shapes*, and thus simultaneously robustly recognize faces with identity-sensitive features and accurately reconstruct 3D face shapes with both features (see Fig. 1).

Specifically, we represent 3D face shapes with a composite model, in which identity and residual (i.e., non-identity) shape components are represented with separate latent variables. Based on the composite model, we propose a joint learning pipeline that is implemented as an encoder-decoder network to disentangle shape features during reconstructing 3D face shapes. The encoder network converts the input 2D face image to identity and residual latent representations, from which the decoder network recovers its 3D face shape. The learning process is supervised by both reconstruction loss and identification loss, and based on a set of 2D face images with labelled identity information and corresponding 3D face shapes that are obtained by an adapted multi-image 3DMM fitting method. Comprehensive evaluation experiments prove the superiority of the proposed method over existing baseline methods in both 3D face reconstruction accuracy and face recognition accuracy. Our main contributions are summarized below.

(i) We propose a method which for the first time explicitly optimizes face recognition and 3D face reconstruction simultaneously. The method achieves state-of-the-art 3D face reconstruction accuracy via joint discriminative feature learning and 3D face reconstruction.

(ii) We devise an effective training process for the proposed network that can disentangle identity and non-identity features in reconstructed 3D face shapes. The network, while being pre-trained by 3DMM-generated data, can surmount the limited 3D shape space determined by the 3DMM bases, in the sense that it better captures identity-sensitive and identity-irrelevant features in 3D face shapes.

(iii) We leverage the effectiveness of disentangled identity features in reconstructed 3D face shapes for improving face recognition accuracy, as being demonstrated by our experimental results. This further expands the application scope of 3D face reconstruction.

2. Related Work

In this section, we review existing work that is closely related to our work from two aspects: 3D face reconstruction for recognition and Convolutional Neural Network (CNN) based 3D face reconstruction.

3D Face Reconstruction for Recognition. 3D face reconstruction was first introduced for recognition by Blanz and Vetter [5]. They reconstructed 3D faces by fitting 3DMM to 2D face images, and used the obtained 3DMM parameters as features for face recognition. Their employed 3DMM fitting method is essentially an image-based analysis-by-synthesis approach, which does not consider the features unique to different individuals. This method was recently improved by Tran et al. [33] via pooling the 3DMM parameters of the images of the same subject and using a CNN to regress the pooled parameters. They experimentally proved the improved discriminative power of their obtained 3DMM parameters.

Instead of using 3DMM parameters for recognition, Liu et al. [23] proposed to recover pose and expression normalized 3D face shapes directly from 2D face landmarks via cascaded regressors and match the reconstructed 3D face shapes via the iterative closest point algorithm for face recognition. Other researchers [31, 36] utilized the reconstructed 3D face shapes for face alignment to assist extracting pose-robust features.

To summarize, *existing methods, when reconstructing 3D face shapes, do not explicitly consider recognition performance*. In [23] and [33], even though the identity of 3D face shapes in the training data is stressed, respectively, by pooling 3DMM parameters and by normalizing pose and expression, their methods of learning mapping from 2D images to 3D face shapes are *unsupervised* in the sense of utilizing identity labels of the training data (see Fig. 1).

CNN-based 3D Face Reconstruction. Existing CNN-based 3D face reconstruction methods can be divided into two categories according to the way of representing 3D

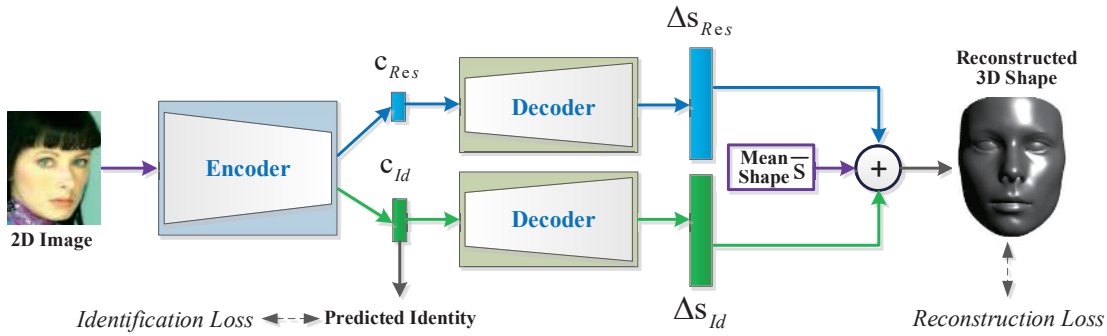


Figure 2. Overview of the proposed encoder-decoder based joint learning pipeline for face recognition and 3D shape reconstruction.

faces. Methods in the first category use 3DMM parameters [10, 27, 30, 32, 33, 41], while methods in the second category use 3D volumetric representations. Jourabloo and Liu [17–19] first employed CNN to regress 3DMM parameters from 2D images for the purpose of large-pose face alignment. In [41], a cascaded CNN pipeline was proposed to exploit the intermediate reconstructed 3D face shapes for better face alignment. Recently, Richardson et al. [27] used two CNNs to reconstruct detailed 3D faces in a coarse-to-fine approach. Although they showed visually more plausible 3D shapes, it is not clear how beneficial the reconstructed 3D facial details are to face recognition.

Jackson et al. [16] proposed to represent 3D face shapes by 3D volumetric coordinates, and train a CNN to directly regress the coordinates from the input 2D face image. Considering the high dimensionality of original 3D face point clouds, as a compromise, they employed 3D volumetric representations. In consequence, the 3D face shapes generated by their method are of low resolution, which are apparently not favorable for face recognition.

3. Proposed Method

In this section, we first introduce a composite 3D face shape model with latent representations, based on which our method is devised. We then present the proposed encoder-decoder based joint learning pipeline. We finally give the implementation detail of our proposed method, including network structure, training data, and training process.

3.1. A Composite 3D Face Shape Model

In this paper, 3D face shapes are densely aligned, and each 3D face shape is represented by the concatenation of its vertex coordinates as

$$\mathbf{s} = [x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_n, y_n, z_n]^T, \quad (1)$$

where n is the number of vertices in the point cloud of the 3D face, and 'T' means transpose. Based on the assumption that 3D face shapes are composed by identity-sensitive and identity-irrelevant parts, we re-write the 3D face shape $\mathbf{s}$ of a subject as

$$\mathbf{s} = \bar{\mathbf{s}} + \Delta\mathbf{s}_{Id} + \Delta\mathbf{s}_{Res}, \quad (2)$$

where $\bar{\mathbf{s}}$ is the mean 3D face shape (computed across all training samples with neutral expression), $\Delta\mathbf{s}_{Id}$ is the identity-sensitive difference between $\mathbf{s}$ and $\bar{\mathbf{s}}$, and $\Delta\mathbf{s}_{Res}$ denotes the residual difference. A variety of sources could lead to the residual difference, for example, expression-induced deformations and temporary detail.

We further assume that $\Delta\mathbf{s}_{Id}$ and $\Delta\mathbf{s}_{Res}$ can be described by latent representations, $\mathbf{c}_{Id}$ and $\mathbf{c}_{Res}$, respectively. This is formulated by

$$\Delta\mathbf{s}_{Id} = f_{Id}(\mathbf{c}_{Id}; \theta_{Id}),\ \Delta\mathbf{s}_{Res} = f_{Res}(\mathbf{c}_{Res}; \theta_{Res}). \quad (3)$$

Here, f_{Id} (f_{Res}) is the mapping function that generates the corresponding shape component $\Delta\mathbf{s}_{Id}$ ($\Delta\mathbf{s}_{Res}$) from the latent representation, with parameters θ_{Id} (θ_{Res}). The latent representations can be obtained from the input 2D face image $\mathbf{I}$ via another function h:

$$[\mathbf{c}_{Id}, \mathbf{c}_{Res}] = h(\mathbf{I}; \theta), \quad (4)$$

where θ are the parameters involved in h. Usually, the latent representations $\mathbf{c}_{Id}$ and $\mathbf{c}_{Res}$ ($\in \mathbb{R}^{Q\times 1}$) are of much lower dimension than the input 2D face image $\mathbf{I}$ as well as the output 3D face shape point cloud $\mathbf{s}$ (see Fig. 3).

3.2. An Encoder-Decoder Network

The above composite model can be naturally implemented as an encoder-decoder network, in which h serves as an encoder to extract latent representations of 2D face images, and f_{Id} and f_{Res} are decoders to recover the identity and residual shape components. As shown in Fig. 2, the latent representation $\mathbf{c}_{Id}$ is employed as features for face recognition. In order to enhance the discriminative capability of $\mathbf{c}_{Id}$, we impose over $\mathbf{c}_{Id}$ an identification loss that can disentangle identity-sensitive from identity-irrelevant features in 3D face shapes. Meanwhile, a reconstruction loss is applied to the 3D face shapes generated by the

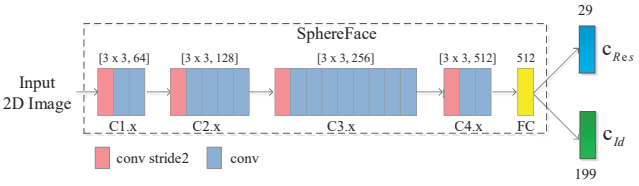


Figure 3. Encoder in the proposed method is implemented based on SphereFace [24]. It converts the input 2D image to latent identity and residual shape feature representations.

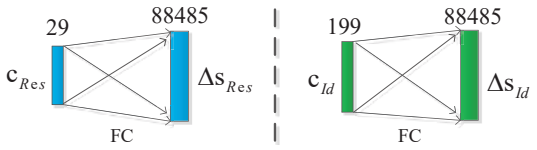


Figure 4. Decoders in the proposed method are implemented as a fully connected (FC) layer. They convert the latent representations to corresponding shape components.

decoders to guide $\mathbf{c}_{Res}$ and f_{Res} to better capture identity-irrelevant shape components. Such an encoder-decoder network enables us to jointly learn accurate 3D face shape reconstructor and discriminative shape features. Next, we detail the implementation of our proposed method.

3.3. Implementation Detail

3.3.1 Network Structure

Encoder Network. The encoder network, aiming at extracting latent identity and residual shape representations of 2D face images, should have good capacity for discriminating different faces as well as capturing abundant detail on faces. Hence, we employ a state-of-the-art face recognition network, i.e., SphereFace [24], as the base encoder network. This network consists of 20 convolutional layers and a fully-connected (FC) layer, and takes the 512-dim output of the FC layer as the feature representation of faces. We append another two parallel FC layers to the base SphereFace network to generate 199-dim identity latent representation and 29-dim residual latent representation, respectively. Fig. 3 depicts the SphereFace-based encoder network. Input 2D face images to the encoder network are pre-processed as in [24]: The face regions are detected by using MTCNN [40], and then cropped and scaled to 112×96 pixels whose values are normalized to the interval from -1 to 1. Each dimension in the output latent representations is also normalized to the interval from -1 to 1.

Decoder Network. Taking the identity and residual latent representations as input, the decoder network recovers the identity and residual shape components of 3D face shapes. Since both the input and output of the decoder network are vectors, we use a multilayer perception (MLP) network to implement the decoder. More specifically, we use two FC layers to convert the latent representations to corresponding shape components, one for identity and the other for the residual. Fig. 4 shows the detail of the implemented decoder network. As can be seen, the generated 3D face point clouds have $29,495$ vertices, and the output of the MLP-based decoder network thus is $88,485$-dim. By analogy with the 3DMM of 3D faces, the weights of the connections between one entry in $\mathbf{c}_{Id}$ or $\mathbf{c}_{Res}$ and the output neurons can be considered as one basis of 3DMM. Thanks to the joint training strategy, the capacity of the 'bases' learnt here is much beyond that of the classical 3DMM, as we will show in the experiments.

Loss Functions. We use two loss functions, 3D shape reconstruction error and face identification error, as the supervisory signals during the end-to-end training of the encoder-decoder network. To measure the 3D shape reconstruction error, we use the Euclidean loss, L_R, to evaluate the deviation of the reconstructed 3D face shape from the ground truth one. The reconstructed 3D face shape is obtained according to Eq. (2) based on the decoder network's output $\Delta\mathbf{s}_{Id}$ and $\Delta\mathbf{s}_{Res}$ (see Fig. 2). The face identification error is measured by using the softmax loss, L_C, over the identity latent representation. The overall loss to the proposed encoder-decoder network is defined by

$$L = \lambda_R L_R + L_C, \tag{5}$$

where λ_R is the weight for the reconstruction loss.

3.3.2 Training Data

To train the encoder-decoder network, we need a set of data that contain multiple 2D face images of same subjects with their corresponding 3D face shapes, i.e., $\{\mathbf{I}^i, l^i, \mathbf{s}^i\}_{i=1}^N$. $l^i \in \{1, 2, \cdots, K\}$ is the subject label of the 2D face image $\mathbf{I}^i$ and 3D face $\mathbf{s}^i$. N is the total number of 2D images, and K is the total number of subjects in the training set. However, such a large-scale dataset is not publicly available. Motivated by prior work [33], we construct the training data from CASIA-WebFace [37], a widely-used 2D face recognition database, via a multi-image 3DMM fitting method, which is adapted from the method in [29, 42].

Faces on the images in CASIA-WebFace are detected by using the method in [40], and 68 landmarks are located by the method in [7]. We discard images where either detection or alignment fails, which results in $488,848$ images of $10,575$ different subjects in our training data. On average, each subject has $\sim$ 46 images. Given the face images and their facial landmarks, we apply the following multi-image 3DMM fitting method to estimate for each subject an identity 3D shape component that is common to all its 2D

face images, and different residual 3D shape components that are unique to each of the subject's 2D images.

The 3DMM represents a 3D face shape as

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}, \tag{6}$$

where $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ are, respectively, the identity and expression shape bases, and α_{id} and α_{exp} are the corresponding coefficients. In this paper, we use the shape bases given by the Basel Face Model [25] as $\mathbf{A}_{id}$, and the blendshape bases in FaceWarehouse [8] as $\mathbf{A}_{exp}$.

To fit the 3DMM to M images of a subject, we attempt to minimize the difference between $\mathbf{u}$, the landmarks detected on the images, and $\hat{\mathbf{u}}$, the landmarks obtained by projecting the estimated 3D face shapes onto the images, under the constraint that all the images of the subject share the same α_{id}. $\hat{\mathbf{u}}$ is computed from the estimated 3D face shape $\hat{\mathbf{s}}$ (let $\hat{\mathbf{s}}_U$ denote the vertices in $\hat{\mathbf{s}}$ corresponding to the landmarks) by $\hat{\mathbf{u}} = f \cdot \mathbf{P} \cdot \mathbf{R} \cdot (\hat{\mathbf{s}}_U + \mathbf{t})$, where f is the scale factor, $\mathbf{P}$ is the orthographic projection, $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector in 3D space. Mathematically, our multi image 3DMM fitting optimizes the following objective:

$$\min_{\alpha_{id}, \{f^j, \mathbf{R}^j, \mathbf{t}^j, \alpha^j_{exp}\}_{j=1}^{M}} \sum_{j=1}^{M} \|\mathbf{u}^j - \hat{\mathbf{u}}^j\|_2^2. \tag{7}$$

We solve the optimization problem in Eq. (7) in an alternating way. As an initialization, we set both α_{id} and α_{exp} to zero. We first estimate the projection parameters $\{f^j, \mathbf{R}^j, \mathbf{t}^j\}_{j=1}^{M}$, then expression parameters $\{\alpha^j_{exp}\}_{j=1}^{j=M}$, and lastly identity parameters α_{id}. When estimating one of the three sets of parameters, the rest two sets of parameters are fixed as they are. The optimization is repeated until the objective function value does not change. We have typically found this to converge within seven iterations.

3.3.3 Training Process

With the prepared training data, we train our encoder-decoder network in three phases. In Phase I, we train the encoder by setting the target latent representations as $\mathbf{c}_{Id} = \alpha_{id}$ and $\mathbf{c}_{Res} = \alpha_{exp}$ and using Euclidean loss. In Phase II, we train the decoder for the identity and residual components separately. In Phase III, the end-to-end joint training is conducted based on the pre-trained encoder and decoder. Considering that the network already has good performance in reconstruction after pre-training, we first lay more emphasis on recognition in the joint loss function by setting λ_R to 0.5. When the loss function gets saturated (usually within 10 epochs), we continue the training by updating λ_R to 1.0. The joint training concludes in about another 20 epochs.

It is worth mentioning that the recovered 3DMM parameters are directly used as the latent representations during pre-training. This provides a good initialization for the encoder-decoder network, but limits the network to the capacity of the pre-determined 3DMM bases. The joint training in Phase III alleviates such limitation by utilizing the identification loss as a complementary supervisory signal to the reconstruction loss. As a result, the learnt encoder-decoder network can better disentangle identity from non-identity information in 3D face shapes, and thus enhance face recognition accuracy without impairing the 3D face reconstruction accuracy.

4. Experiments

Two sets of experiments have been done to evaluate the effectiveness of the proposed method in 3D face reconstruction and face recognition. The MICC [2] and BU3DFE [38] databases are used for experiments of 3D face reconstruction, and the LFW [15] and YTF [35] databases are used in face recognition experiments. Next, we report the experimental results [1].

4.1. 3D Shape Reconstruction Accuracy

The 3D face reconstruction accuracy is assessed by using 3D Root Mean Square Error (RMSE) [33], defined as $\texttt{RMSE} = \frac{1}{N_T}\sum_{i=1}^{N_T}(\|\mathbf{s}_i^* - \hat{\mathbf{s}}_i\|/n)$, where N_T is the total number of testing samples, $\mathbf{s}_i^*$ and $\hat{\mathbf{s}}_i$ are the ground truth and reconstructed 3D face shape of the $i^{\texttt{th}}$ testing sample. To compute the RMSE, the reconstructed 3D faces are first aligned to ground truth via Procrustes global alignment based on 68 3D landmarks as suggested by [3], and then cropped at a radius of $95mm$ around the nose tip.

We compare our method with four state-of-the-art 3D face reconstruction methods, 3DDFA [42], 3DMM-CNN [33], 3D shape regression based (3DSR) method [23], and VRN [16]. Among them, the first two methods reconstruct 3D face shapes via estimating 3DMM parameters, while the other two directly regress 3D face shapes from either landmarks or 2D images. 3DMM-CNN method is the only existing method that takes into consideration the discriminative power of the estimated 3DMM parameters. 3DSR method generates pose and expression normalized 3D face shapes that are believed to be more beneficial to face recognition. For those methods that need facial landmarks on 2D images, we use the method in [7] to automatically detect the landmarks.

Results on MICC. The MICC database contains three challenging face videos and ground-truth 3D models acquired using a structured-light scanning system for each of 53 subjects. The videos span the range of controlled indoor to unconstrained outdoor settings. The outdoor videos are very challenging due to the uncontrolled lighting conditions. In this experiment, we randomly select $5,000$

[1] More experimental results are provided in the supplementary material.

Table 1. 3D face reconstruction accuracy (RMSE) under different yaw angles on the BU3DFE database.

Method	±90°	±80°	±70°	±60°	±50°	±40°	±30°	±20°	±10°	0°	Avg.
VRN	6.96	6.20	6.14	6.01	5.91	5.50	4.93	3.86	3.70	3.66	5.29
3DDFA	2.90	2.88	2.81	2.82	2.77	2.79	2.76	2.73	2.55	2.48	2.75
3DMM-CNN	-	-	-	-	2.30	2.26	2.23	2.22	2.19	2.17	2.23
3DSR	2.11	2.11	2.12	2.13	2.16	2.14	2.12	2.10	2.10	2.09	2.12
Proposed	**2.09**	**2.04**	**2.03**	**2.03**	**2.00**	**1.99**	**2.03**	**2.01**	**1.97**	**1.93**	**2.01**

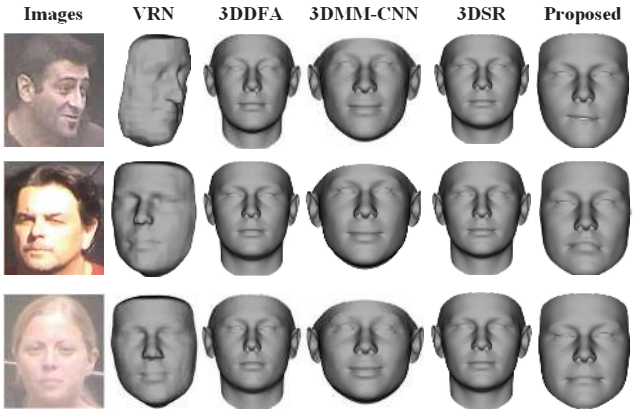


Figure 5. Reconstruction results for three MICC subjects. The first column shows the input images, and the rest columns show the reconstructed 3D shapes that have the same expression as the input images, using the methods of VRN [16], 3DDFA [42], 3DMM-CNN [33], 3DSR [23] and the proposed method.

Table 2. 3D face reconstruction accuracy on the MICC database.

Method	VRN	3DDFA	3DMM-CNN	3DSR	Proposed
RMSE	5.34	2.73	2.20	2.07	**2.00**

images from 31,466 outdoor video frames of 53 subjects. Table 2 shows the 3D face reconstruction error of different methods on the MICC database. As can be seen, our proposed method obtains the best accuracy due to its fine-grained processing of features in 3D face shapes. Note that VRN, the first method in the literature that regresses 3D face shapes directly from 2D images, has relatively high reconstruction error in terms of RMSE, mainly because it generates low-resolution 3D face shapes as volumetric representations. In contrast, we reconstruct high-resolution (dense) 3D face shapes as point clouds with help from low dimensional latent representations.

Results on BU3DFE. The BU3DFE database contains 3D faces of 100 subjects displaying expression of neutral (NE), happiness (HA), disgust (DI), fear (FE), anger (AN), surprise (SU) and sadness (SA). All non-neutral expressions were acquired at four levels of intensity. We select neutral and the first intensity level of the rest six expressions as testing data, resulting in 700 testing samples. Further, we render another set of testing images of neutral expression at different poses, i.e., $-90°$ to $90°$ yaws with a $10°$ interval. These two testing sets evaluate the reconstruction across

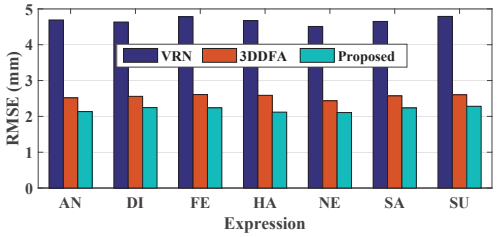


Figure 6. Reconstruction accuracy of 3D face shapes under different expressions on the BU3DFE database. The mean RMSEs of thee methods over all expressions are 4.68, 2.56, and 2.19 respectively.

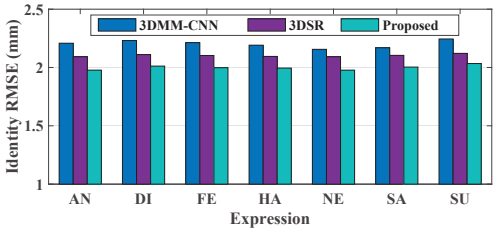


Figure 7. Reconstruction accuracy of the identity component of 3D face shapes under different expressions on the BU3DFE database. The mean RMSEs of thee methods over all expressions are 2.21, 2.10, and 2.00 respectively.

expressions and poses, respectively.

Table 1 shows the *reconstruction error across poses* (i.e., yaw) of different methods. It can be seen that the RMSE of the proposed method is lower than that of baselines. Moreover, as the pose angle becomes large, the error of our method does not increase substantially. This proves the robustness of the proposed method to pose variations. Figure 6 shows the *reconstruction error across expressions* of VRN, 3DDFA, and the proposed method based on their reconstructed 3D face shapes that have the same expression as the input images. Figure 7 compares 3DMM-CNN, 3DSR, and the proposed method in terms of RMSE of their reconstructed identity or expression-normalized 3D face shapes. These results demonstrate the superiority of the proposed method over baselines in handling expressions.

Some example 3D face reconstruction results are shown in Fig. 5 and Fig. 8. From these results, we can clearly see that the proposed method not only performs well in reconstructing accurate 3D face shapes for in-the-wild 2D images, but also disentangles identity and non-identity (e.g.,

Table 3. Face recognition accuracy on the LFW and YTF databases.

Method	Shape	Texture	Accuracy	100%-EER	AUC	TAR-10%	TAR-1%
Labeled Faces in the Wild (LFW)							
3DMM	√	×	66.13 ± 2.79	65.70 ± 2.81	72.24 ± 2.75	35.90 ± 3.74	12.37 ± 4.81
	×	√	74.93 ± 1.14	74.50 ± 1.21	82.94 ± 1.14	60.40 ± 3.15	28.73 ± 7.17
	√	√	75.25 ± 2.12	74.73 ± 2.56	83.21 ± 1.93	59.40 ± 4.64	29.67 ± 4.73
3DDFA	√	×	66.98 ± 2.56	67.13 ± 1.90	73.30 ± 2.49	36.76 ± 6.27	10.00 ± 3.22
	√	×	90.53 ± 1.34	90.63 ± 1.61	96.60 ± 0.79	91.13 ± 2.62	58.20 ± 12.14
3DMM-CNN	×	√	90.60 ± 1.07	90.70 ± 1.17	96.75 ± 0.59	91.23 ± 2.42	52.60 ± 8.14
	√	√	92.35 ± 1.29	92.33 ± 1.33	97.71 ± 0.64	94.20 ± 2.00	65.57 ± 6.93
Proposed	√	×	$\mathbf{94.43 \pm 1.47}$	$\mathbf{94.40 \pm 1.52}$	$\mathbf{98.12 \pm 0.90}$	$\mathbf{95.07 \pm 2.39}$	$\mathbf{74.54 \pm 4.33}$
YouTube Faces (YTF)							
3DMM	√	×	73.26 ± 2.51	73.08 ± 2.65	80.41 ± 2.60	51.36 ± 5.11	24.04 ± 4.56
	×	√	77.34 ± 2.54	76.96 ± 2.64	85.32 ± 2.63	63.16 ± 5.07	31.36 ± 5.21
	√	√	79.56 ± 2.08	79.20 ± 2.07	87.35 ± 1.92	69.08 ± 5.00	34.56 ± 6.89
3DDFA	√	×	68.10 ± 2.93	67.96 ± 3.12	74.95 ± 3.04	40.52 ± 3.65	12.20 ± 2.67
	√	×	88.28 ± 1.84	88.32 ± 2.16	95.95 ± 1.38	86.60 ± 3.95	51.12 ± 8.86
3DMM-CNN	×	√	87.56 ± 2.56	87.68 ± 2.25	94.44 ± 1.38	84.80 ± 4.89	40.92 ± 8.26
	√	√	$\mathbf{88.80 \pm 2.21}$	$\mathbf{88.84 \pm 2.40}$	95.37 ± 1.43	87.92 ± 4.18	46.56 ± 6.20
Proposed	√	×	88.74 ± 1.03	88.70 ± 1.15	$\mathbf{96.28 \pm 0.63}$	$\mathbf{89.00 \pm 2.40}$	$\mathbf{53.44 \pm 4.51}$

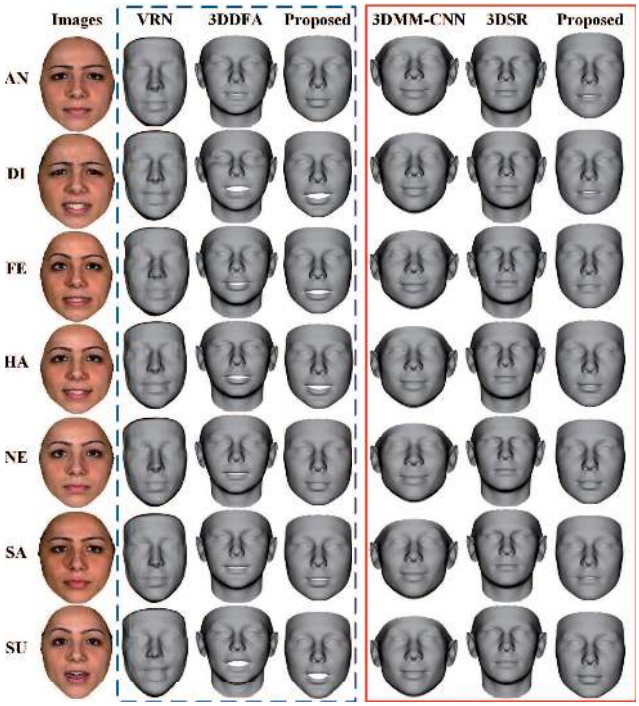


Figure 8. Reconstruction results for an BU3DFE subject under seven different expressions. The first column shows the input images. In the blue box, we show the reconstructed 3D shapes that have the same expression as the input images, using the methods of VRN [16], 3DDFA [42] and the proposed method. In the red box, we show the reconstructed *identity* 3D shapes obtained by 3DMM-CNN [33], 3DSR [23] and the proposed method. Our composite 3D shape model enables us to generate two types of 3D shapes.

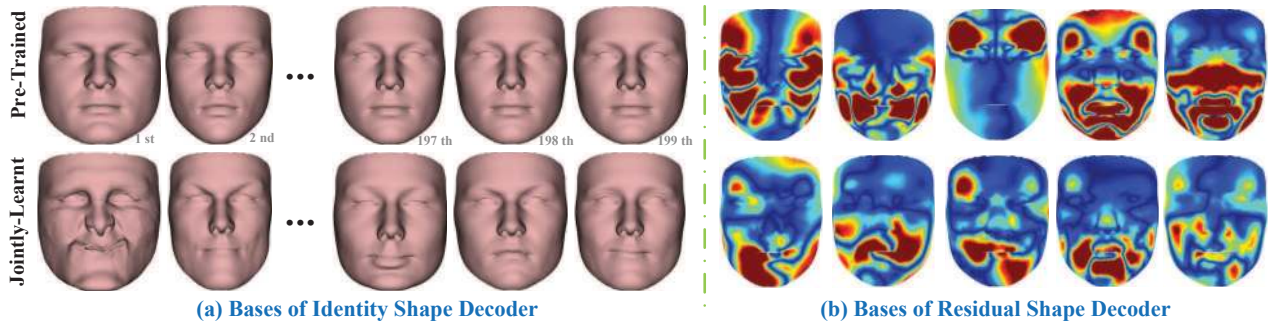
expression) components in 3D face shapes. As we will show in the following face recognition experiments, the disentangled shape features contribute to face recognition.

4.2. Face Recognition Accuracy

To evaluate the effectiveness of our shape features (i.e., the identity representations) to face recognition, we compute the similarity of two faces using the cosine distances between their shape features extracted by the encoder of our method. To investigate the complementarity between our learnt shape features and existing texture features, we also fuse our method with existing methods via summation at the score level [21]. The counterpart methods we consider here include 3DMM [28], 3DDFA [42], 3DMM-CNN [33], and SphereFace [24]. We compare the methods in terms of verification accuracy, 100%-EER (Equal Error Rate), AUC (Area Under Curve) of ROC (Receiver Operating Characteristic) curves, and TAR (True Acceptance Rate) at FAR (False Acceptance Rate) of 10% and 1%.

Results on LFW. The Labeled Faces in the Wild (LFW) benchmark dataset contains $13,323$ images collected from Internet. The verification set consists of 10 folders, each with 300 same-person pairs and 300 different-person pairs. The recognition accuracy of different methods on LFW is listed in Tab. 3. Among all the 3D face reconstruction methods, when using only shape features, our proposed method achieves the highest accuracy, improving TAR@1% FAR from 58.20% to 74.54% with respect to the latest 3DMM-based method [33].

Results on YTF. The YouTube Faces (YTF) database contains $3,425$ videos of $1,595$ individuals. Face images (video frames) in YTF have lower quality than those in LFW, due to larger variations in pose, illumination and expression, and low resolution as well. Table 3 summarizes the recognition accuracy of different methods on YTF. Despite the low-quality face images, our proposed method still outperforms the baseline methods in the sense of extracting discriminative shape features. By fusing with one of the state-of-the-art texture-based face recognition methods (i.e.,



(a) Bases of Identity Shape Decoder (b) Bases of Residual Shape Decoder

Figure 9. Comparing the pre-trained 3DMM-like and our jointly-learnt bases defined by the weights of identity and residual shape decoders. (a) For the bases of identity shape decoder, the weights associated with each entry in $\mathbf{c}_{Id}$ are added to the mean shape, reshaped to a point cloud ($\in \mathbb{R}^{3\times n}$), and shown as polygon meshes. (b) For the bases of residual shape decoder, the weights associated with each entry in $\mathbf{c}_{Res}$ are reshaped to a point cloud ($\in \mathbb{R}^{3\times n}$), and shown as a heat map that measures the norm value of each vertex (i.e., the deviation from the identity shape). Red colors in the heat maps indicate larger deviations. It is important to note that the conventional 3DMM bases are trained from 3D face scans, while our bases are learnt from 2D images.

Table 4. Efficiency comparison of different methods.

Method	VRN	3DDFA	3DMM-CNN	3DSR	Proposed
Time (ms)	55.68	39.17	30.12	29.80	**4.79**

SphereFace [24]), our proposed method further improves the face recognition accuracy on YTF from 94.78% to 95.18%. This proves the complementarity of *properly reconstructed* shape features to texture features in face recognition. This is a notable result especially considering the 2D face recognition method of SphereFace [24] has already set a very high baseline (i.e., 94.78%).

4.3. Computational Efficiency

To assess the computational efficiency, we run the methods on a PC (with an Intel Core i7-5930K @ 3.5GHz, 32GB RAM and an GeForce GTX 1080) for 700 images, and calculate the average runtime per image in Tab. 4. Note that 3DDFA and 3DMM-CNN estimate the 3DMM parameters in the first step, and we report their runtime of obtaining the final 3D faces. For VRN, 3DDFA and 3DMM-CNN, despite stand-alone landmark detection is required, the reported time does not include the landmark detection time. Our proposed method needs only 4.79 milliseconds (ms) per image, which is an order of magnitude faster than baseline methods. This is owing to the light-weight network in our method. In contrast, baseline methods use either very deep networks [33], or cascade approaches [23, 27].

4.4. Analysis and Discussion

To offer insights into the learnt decoders, we visualize their weight parameters in Fig. 9. The weights associating one entry in the latent representations with all the neurons in the FC layer in the decoders are analogous to a 3DMM basis (see Fig. 4). Both pre-trained bases and jointly-learnt bases are shown for comparison in Fig. 9, from which the following observations can be made.

(i) The pre-trained identity bases approximate the conventional 3DMM bases [4] that are ordered with latter bases capturing less shape variations. In contrast, our jointly-learnt identity bases all describe rich shape variations.

(ii) Some basis shapes in the jointly-learnt bases do not look like regular face shapes. We believe this is due to the employed joint reconstruction and identification loss function. The bases trained from a set of 3D scans as in 3DMM, while optimal for reconstruction, might limit the discriminativeness of shape parameters. Our bases are trained with the classification in mind, which ensures the superior performance of our method in face recognition.

(iii) The pre-trained residual bases, like the expression shape bases [8], appear symmetrical. The jointly-learnt residual bases display more diverse shape deviation patterns. This indicates that the residual shape deformation captured by the jointly-learnt bases is much beyond that caused by expression changes, and proves the effectiveness of our method in disentangling 3D face shape features.

5. Conclusions

We have proposed a novel encoder-decoder-based method for jointly learning discriminative shape features from a 2D face image and reconstructing its dense 3D face shape. To train the encoder-decoder network, we implement a multi-image 3DMM fitting method to construct training data, and develop an effective training scheme with a joint reconstruction and identification loss. We show with comprehensive experimental results that the proposed method can effectively disentangle identity and non-identity features in 3D face shapes and thus achieve state-of-the-art 3D face reconstruction accuracy as well as improved face recognition accuracy.

References

[1] https://support.apple.com/en-us/HT208109. Accessed: 2017-11-15.

[2] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2D/3D hybrid face dataset. In *Workshop on Human gesture and behavior understanding*, pages 79–80. ACM, 2011.

[3] 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 *ACCV*, pages 377–391, 2016.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999.

[5] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *TPAMI*, 25(9):1063–1074, 2003.

[6] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition. *CVIU*, 101(1):1–15, 2006.

[7] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*, 2017.

[8] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014.

[9] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. *TOG*, 35(4):126:1–126:12, 2016.

[10] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *CVPR*, 2017.

[11] M. Emambakhsh and A. Evans. Nasal patches and curves for expression-robust 3D face recognition. *TPAMI*, 39(5):995–1007, 2016.

[12] H. Han and A. K. Jain. 3D face texture modeling from uncalibrated frontal and profile images. In *BTAS*, pages 223–230, 2012.

[13] X. Han, C. Gao, and Y. Yu. Deepsketch2face: A deep learning based sketching system for 3D face and caricature modeling. *TOG*, 36(4), 2017.

[14] B. K. Horn and M. J. Brooks. *Shape from shading*. Cambridge, MA: MIT press, 1989.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[16] 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 *ICCV*, 2017.

[17] A. Jourabloo and X. Liu. Pose-invariant 3D face alignment. In *ICCV*, pages 3694–3702, 2015.

[18] A. Jourabloo and X. Liu. Pose-invariant face alignment via CNN-based dense 3D model fitting. *IJCV*, in press, 2017.

[19] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, 2017.

[20] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *TPAMI*, 33(2):394–405, 2011.

[21] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *TPAMI*, 20(3):226–239, 1998.

[22] F. Liu, D. Zeng, J. Li, and Q. Zhao. Cascaded regressor based 3D face reconstruction from a single arbitrary view image. *arXiv:1509.06161*, 2015.

[23] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, pages 545–560, 2016.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[25] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009.

[26] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017.

[27] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.

[28] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, pages 986–993, 2005.

[29] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. In *CVPR*, pages 4197–4206, 2016.

[30] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017.

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[32] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *CVPR*, 2017.

[33] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR*, 2017.

[34] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1283–1292, 2017.

[35] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.

[36] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *CVPR*, pages 3539–3545, 2013.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014.

[38] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *FG*, pages 211–216, 2006.

[39] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016.

[41] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3D solution. In *CVPR*, pages 146–155, 2016.

[42] 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 *CVPR*, pages 787–796, 2015.