Joint Face Alignment and 3D Face Reconstruction with Application to Face Recognition

Feng Liu, Qijun Zhao, *Member, IEEE*, Xiaoming Liu, *Member, IEEE* and Dan Zeng

Abstract—Face alignment and 3D face reconstruction are traditionally accomplished as separated tasks. By exploring the strong correlation between 2D landmarks and 3D shapes, in contrast, we propose a joint face alignment and 3D face reconstruction method to simultaneously solve these two problems for 2D face images of arbitrary poses and expressions. This method, based on a summation model of 3D face shapes and cascaded regression in 2D and 3D face shape spaces, iteratively and alternately applies two cascaded regressors, one for updating 2D landmarks and the other for 3D face shape.The 3D face shape and the landmarks are correlated via a 3D-to-2D mapping matrix, which is updated in each iteration to refine the location and visibility of 2D landmarks. Unlike existing methods, the proposed method can fully automatically generate both pose-and-expression-normalized (PEN) and expressive 3D face shapes and localize both visible and invisible 2D landmarks. Based on the PEN 3D face shapes, we devise a method to enhance face recognition accuracy across poses and expressions. Extensive experiments show that the proposed method can achieve the state-of-the-art accuracy in both face alignment and 3D face reconstruction, and benefit face recognition owing to its reconstructed PEN 3D face shapes.

Index Terms—3D face reconstruction; face alignment; cascaded regression; pose and expression normalization; face recognition.

1 Introduction

THREE-dimensional (3D) face models have recently been employed to assist pose or expression invariant face recognition and achieve state-of-the-art performance [1], [2], [3]. A crucial step in these 3D face assisted face recognition methods is to reconstruct the 3D face model from a two-dimensional (2D) face image. Besides its applications in face recognition, 3D face reconstruction is also useful in other face-related tasks, e.g., facial expression analysis [4], [5] and facial animation [6], [7]. While many 3D face reconstruction methods are available, they mostly require landmarks on the face image as input, and are difficult to handle large-pose faces that have invisible landmarks due to self-occlusion.

Existing studies tackle the problems of facial landmark localization (or face alignment) and 3D face reconstruction *separately*. However, these two problems are chicken-and-egg problems. On one hand, 2D face images are projections of 3D faces onto the 2D plane. Given a 3D face and a 3D-to-2D mapping function, it is easy to compute the visibility and position of 2D landmarks. On the other hand, the landmarks provide rich information about facial geometry, which is the basis of 3D face reconstruction. Figure 1 illustrates the correlation between 2D landmarks and 3D faces. That is, the visibility and position of landmarks in the projected 2D image are determined by three factors: the 3D face shape, the deformation due to expression and pose, and the camera projection parameters. *Given such a clear correlation between 2D landmarks and 3D shape, it is evident that they should ideally be solved jointly, instead of separately as in prior works - indeed this is the core of this work.*

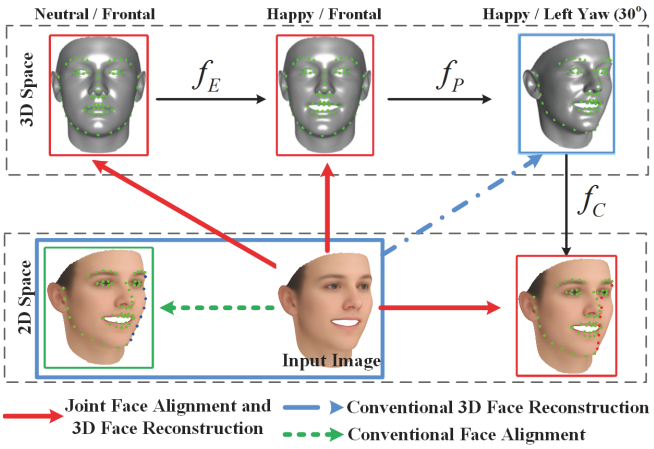

Fig. 1. We view 2D landmarks as being generated from a 3D face through 3D expression (f_E) and pose (f_P) deformation, and camera projection (f_C). While conventional face alignment and 3D face reconstruction are two *separated* tasks and the latter requires the former as the input, this paper performs these two tasks *jointly*, i.e., reconstructing a 3D face and estimating visible/invisible landmarks (green/red points) from a 2D face image with arbitrary poses and expressions.

Motivated by the aforementioned observation, this paper proposes a unified framework to simultaneously solve the two problems of face alignment and 3D face shape reconstruction. Two sets of regressors are jointly learned from a training set of pairing annotated 2D face images and 3D face

- *Feng Liu, Dan Zeng and Qijun Zhao are with the National Key Laboratory of Fundamental Science on Synthetic Vision, College of Computer Science, Sichuan University, Chengdu, Sichuan 610065, P. R. China. Qijun Zhao is the corresponding author, reachable at qjzhao@scu.edu.cn.*
- *Xiaoming Liu is with the Dept. of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, U.S.A.*

- *Manuscript received June 25, 2017.*

shapes. Based on the texture features around landmarks on a 2D face image, one set of regressors (called landmark regressors) gradually move the landmarks towards their true positions. By utilizing the landmarks on the 2D face image as clues, the other set of regressors (called shape regressors) gradually improve the reconstructed 3D face shape. These two sets of regressors are alternately and iteratively applied. Specifically, in each iteration, adjustment to the landmarks is firstly estimated via the landmark regressors, and this landmark adjustment is also used to estimate 3D face shape adjustment via the shape regressors. The 3D-to-2D mapping is then computed based on the adjusted 3D face shape and 2D landmarks, and it further refines the 2D landmarks.

A preliminary version of this work was published in the 14th European Conference on Computer Vision (ECCV2016) [8]. We further extend the work from three aspects. (i) We explicitly reconstruct expression deformation of 3D faces, so that both PEN (pose and expression normalized) and expressive 3D face shapes can be reconstructed. (ii) We present in detail the application of the proposed method to face recognition. (iii) We carry out more extensive evaluation with comparison to state-of-the-art methods. In summary, this paper makes the following contributions.

- We present a novel cascaded coupled-regressor based method for joint face alignment and 3D face reconstruction from a single 2D image of arbitrary pose and expression.
- By integrating 3D shape information, the proposed method can more accurately locate landmarks on images of arbitrary view angles in $[-90^\circ, 90^\circ]$.
- We explicitly deal with expression deformation of 3D faces, so that both PEN and expressive 3D face shapes can be reconstructed at a high accuracy.
- We propose a 3D-enhanced approach to improve face recognition accuracy on off-angle and expressive face images based on the reconstructed PEN 3D faces.
- We achieve state-of-the-art 3D face reconstruction and face alignment performance on BU3DFE [5], AFLW [9], and AFLW2000 3D [10] databases. We investigate the other-race effect on 3D face reconstruction of the proposed method on FRGC v2.0 database [11]. We demonstrate the effectiveness of our proposed 3D-enhanced face recognition method in improving state-of-the-art deep learning based face matchers on Multi-PIE database [12].

The rest of this paper is organized as follows. Section 2 briefly reviews related work in the literature. Section 3 introduces in detail the proposed joint face alignment and 3D face reconstruction method. Section 4 shows its application to face recognition. Section 5 reports the experimental results. Section 6 concludes the paper.

2 Prior Work

2.1 Face Alignment

Classical face alignment methods, including Active Shape Model (ASM) [13], [14] or Active Appearance Model (AAM) [15], [16], search for landmarks based on global shape models and generative texture models. Constrained Local Model (CLM) [17] also utilizes global shape models to regularize the landmark locations, but it employs discriminative local texture models. Regression based methods [18], [19], [20], [21] have been recently proposed to directly estimate landmark locations by applying cascaded regressors to an input 2D face image. These methods mostly do not consider the visibility of facial landmarks under different view angles. Consequently, their performance degrades substantially for non-frontal faces, and their detected landmarks could be ambiguous because the anatomically correct landmarks might be invisible due to self-occlusion (see Fig. 1).

A few methods focused on large-pose face alignment, which can be roughly divided into two categories: multi-view based and 3D model based. Multi-view based methods [22], [23] define different sets of landmarks as templates, one for each view range. Given an input image, they fit the multi-view templates to it and choose the best fitted one as the final result. These methods are usually complicated to apply, and can not detect invisible self-occluded landmarks. 3D model based methods, in contrast, can better handle self-occluded landmarks with the assistance of 3D face models. Their basic idea is to fit a 3D face model to the input image to recover the 3D landmark locations. Most of these methods [10], [24], [25], [26], [27] use 3D morphable models (3DMM) [28] – either a simplified one with a sparse set of landmarks [10], [25] or a relatively dense one [24]. They estimate the 3DMM parameters by using cascaded regressors with texture features as the input. In [25], the visibility of landmarks is explicitly computed, and the method can cope with face images of yaw angles ranging from -90° to 90°, whereas the method in [24] does not work properly for faces of yaw angles beyond 60°. In [29], Tulyakov and Sebe propose to directly estimate the 3D landmark locations via texture-feature-based regressors for faces of yaw angles upto 50°.

These existing 3D model based methods establish regressions between 2D image features and 3D landmark locations (or indirectly, 3DMM parameters). While our proposed approach is also based on 3D model, unlike existing methods, it carries out regressions both on 2D images and in the 3D space. Regressions on 2D images predict 2D landmarks, while regressions in the 3D space predict 3D landmarks coordinates. By integrating both regressions, our proposed method can more accurately estimate or localize landmarks, and better handle self-occluded landmarks. It thus works well for images of arbitrary view angles in $[-90^\circ, 90^\circ]$.

2.2 3D Face Reconstruction

Estimating the 3D face geometry from a single 2D image is an ill-posed problem. Existing methods, such as Shape from Shading (SFS) and 3DMM, thus heavily depend on priors or constraints. SFS based methods [30], [31] usually utilize an average 3D face model as a reference, and assume the Lambertian lighting model for the 3D face surface. One limitation of SFS methods lies in its assumed connection between 2D texture clues and 3D shape, which is too weak to discriminate among different individuals. 3DMM [1], [28], [32], [33] establishes statistical parametric models for both texture and shape, and represents a 3D face as a linear combination of basis shapes and textures. To recover the

3D face from a 2D image, 3DMM-based methods estimate the combination coefficients by minimizing the discrepancy between the input 2D face image and the one rendered from the reconstructed 3D face. They can better cope with 2D face images of varying illuminations and poses. However, they still suffer from invisible facial landmarks when the input face has large pose angles. To deal with extreme poses, Lee et al. [34], Qu et al. [35] and Liu et al. [36] propose to discard the self-occluded landmarks or treat them as missing data.

All the aforementioned 3D face reconstruction methods require landmarks as input. Consequently, they either manually mark the landmarks, or employ standalone face alignment methods to automatically locate the landmarks. Very recently, Tran et al. [37] propose a convolutional neural network (CNN) based method to estimate discriminative 3DMM parameters directly from single 2D images without requirement of input landmarks. Yet, existing methods always generate 3D faces that have the same pose and expression as the input image, which may not be desired in face recognition due to the challenge of matching 3D faces with expressions [38]. In this paper, we improve 3D face reconstruction by (i) integrating the face alignment step into the 3D face reconstruction procedure, and (ii) reconstructing both expressive and PEN 3D faces, which is believed to be useful for face recognition.

2.3 Unconstrained Face Recognition

Face recognition has been developed rapidly during the past decade, especially since the emergence of deep learning techniques. Automated methods [39], [40], [41] even beat humans at face recognition accuracy on the labelled faces in the wild (LFW) benchmark database. Yet, it is still very challenging to recognize faces in unconstrained 2D images with large pose angles or intensive expressions [42], [43], [44]. Potential reasons for degraded face recognition accuracy on off-angle and expressive face images include (i) off-angle faces usually have less discriminative texture information for identification than frontal faces, resulting in small inter-class differences, (ii) cross-view face images (e.g., frontal and profile face images) may have very limited features in common, leading to large intra-class differences, and (iii) pose and expression variations could cause substantial deformation to face images.

Existing methods recognize off-angle and expressive faces either by extracting invariant features or by normalizing out the deformation caused by pose or expression. Yi et al. [45] fitted a 3D face mesh to the input arbitrary view face images, and extracted pose-invariant features based on the 3D face mesh that was adaptively deformed to the input face images. In DeepFace [46], the input face images were first aligned to the frontal view with assistance of a generic 3D face model, and then recognized based on a deep network. Zhu et al. [3] proposed to generate frontal and neutral face images from the input images by using a method based on 3DMM [28] and deep convolutional neural networks. Very recently, generative adversarial networks (GAN) have been explored by Tran et al. [44] for unconstrained face recognition. They devised a novel network, namely DR-GAN, which can simultaneously synthesize frontal face images and learn pose-invariant feature representations. Hu et al. [47] propose to directly transform a non-frontal face image into frontal one by Learning a Displacement Field network (LDF-Net). LDF-Net achieves state-of-the-art performance for face recognition across poses on Multi-PIE, especially at large poses. To summarize, all these existing methods carry out pose and expression normalization on 2D face images and utilize merely 2D features for recognition. In this paper, on the contrary, we will generate pose and expression normalized 3D faces from the input 2D face images, and use these 3D faces to improve the unconstrained face recognition accuracy.

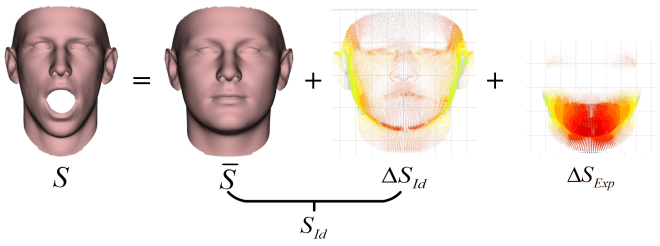

Fig. 2. A 3D face shape of a subject (S) is represented as summation of the mean pose-and-expression-normalized (PEN) 3D face shape ($\bar{S}$), the difference between the subject's PEN 3D face shape and the mean PEN 3D face shape (ΔS_{Id}), and the expression deformation (ΔS_{Exp}).

3 Proposed Method

In this section, we introduce the proposed joint face alignment and 3D face reconstruction method in detail. We start by defining the 3D face model with separable identity and expression components, and based on this model formulating the problem to be solved in this paper. We then give the overall procedure of the proposed method. Afterwards, the preparation of training data is presented, followed by the detailed introduction of key steps in the proposed method, including learning 2D landmark and 3D shape regressors, and estimating 3D-to-2D mapping and landmark visibility.

3.1 Problem Formulation

We denote an n-vertex frontal pose 3D face shape of one subject as

$$S = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{pmatrix} \in \mathbb{R}^{3\times n}, \quad (1)$$

and represent it as a summation of three components:

$$S = S_{Id} + \Delta S_{Exp} = \bar{S} + \Delta S_{Id} + \Delta S_{Exp}, \quad (2)$$

where $\bar{S}$ is the mean of frontal pose and neutral expression 3D face shapes, termed pose-and-expression-normalized (PEN) 3D face shapes in this paper, ΔS_{Id} is the difference between the subject's PEN 3D face shape (denoted as S_{Id}) and $\bar{S}$, and ΔS_{Exp} is the expression-induced deformation in S with respect to S_{Id}. See Fig. 2.

We use S_L to denote a subset of S with columns corresponding to l landmarks. The projections of these landmarks onto a 2D face image $\mathbf{I}$ of the subject with arbitrary view are represented by

$$U = \begin{pmatrix} u_1 & u_2 & \cdots & u_l \\ v_1 & v_2 & \cdots & v_l \end{pmatrix} = f_C \circ f_P(S_L) \in \mathbb{R}^{2\times l}, \quad (3)$$

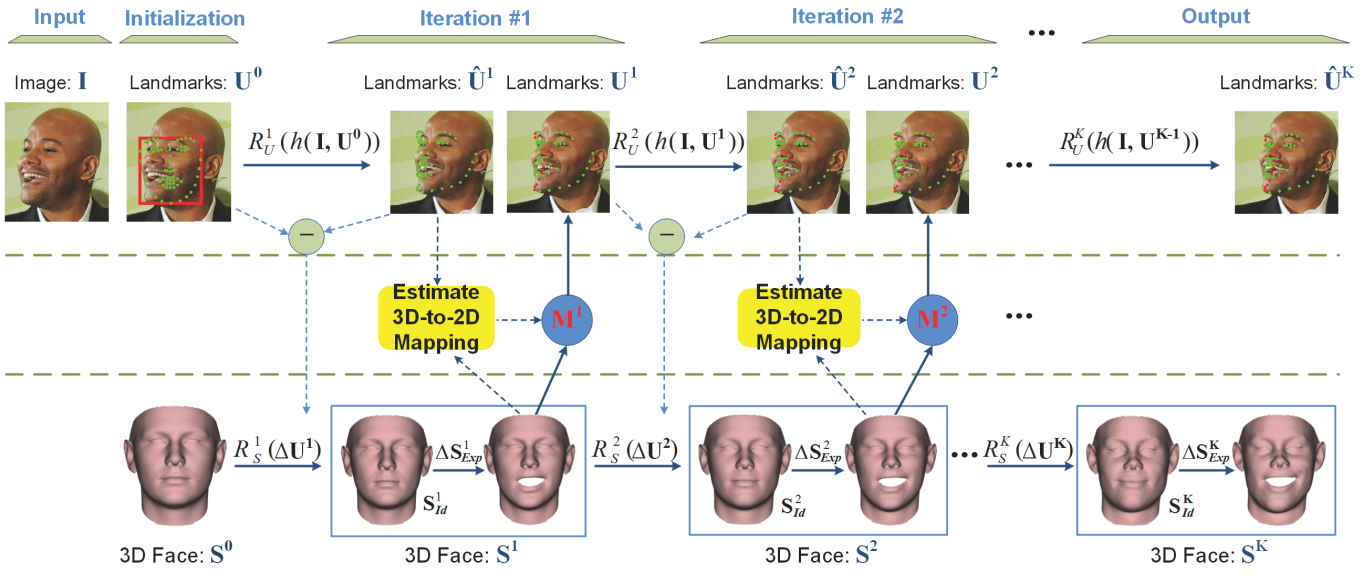

Fig. 3. Flowchart of the proposed joint face alignment and 3D face reconstruction method.

where f_C and f_P are, respectively, camera projection and pose-caused deformation. In this paper, we employ a 3D-to-2D mapping matrix $\mathbf{M} \approx f_C \circ f_P$ to approximate the composite effect of pose induced deformation and camera projection.

Given an input 2D face image $\mathbf{I}$, our goal is to simultaneously localize its landmarks U and estimate its PEN 3D face shape S_{Id} and expression deformation ΔS_{Exp}. Note that, in some context, we also write the 3D face shape and the landmarks as column vectors: $\mathbf{S} = (x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_n, y_n, z_n)^{\mathsf{T}}$, and $\mathbf{U} = (u_1, v_1, u_2, v_2, \cdots, u_l, v_l)^{\mathsf{T}}$, where 'T' is transpose operator.

3.2 The Overall Procedure

Figure 3 shows the flowchart of the proposed method. For the input 2D face image $\mathbf{I}$, its 3D face shape $\mathbf{S}$ is initialized as the mean PEN 3D shape of training faces (i.e., $\mathbf{S}^0 = \bar{\mathbf{S}}$). Its landmarks $\mathbf{U}$ are initialized by placing the mean landmarks of training frontal and neutral faces into the face region specified by a bounding box in $\mathbf{I}$ via similarity transforms. $\mathbf{U}$ and $\mathbf{S}$ are iteratively updated by applying a series of regressors. Each iteration contains three main steps: (i) updating landmarks, (ii) updating 3D face shape, and (iii) refining landmarks.

Updating landmarks This step updates the landmarks' locations from $\mathbf{U}^{k-1}$ to $\hat{\mathbf{U}}^k$ based on the texture features in the input 2D image. This is similar to the conventional cascaded regressor based 2D face alignment [18]. The adjustment to the landmarks' locations in k^{th} iteration, $\Delta\mathbf{U}^k$ is determined by the local texture feature around $\mathbf{U}^{k-1}$ via a regressor,

$$\Delta\mathbf{U}^k = R_U^k(h(\mathbf{I}, \mathbf{U}^{k-1})), \tag{4}$$

where $h(\mathbf{I}, \mathbf{U})$ denotes the texture feature extracted around the landmarks $\mathbf{U}$ in the image $\mathbf{I}$, and R_U^k is a regression function. The landmarks can then be updated by $\hat{\mathbf{U}}^k = \mathbf{U}^{k-1} + \Delta\mathbf{U}^k$. The method for learning these landmark regressors will be introduced in Sec. 3.4.

Updating 3D face shape In this step, the above-obtained landmark location adjustment is used to estimate the adjustment of the 3D face shape $\Delta\mathbf{S}^k$, which consists of two components, $\Delta\mathbf{S}_{Id}^k$ and $\Delta\mathbf{S}_{Exp}^k$. Specifically, a regression function R_S^k models the correlation between the landmark location adjustment $\Delta\mathbf{U}^k$ and the expected adjustment $\Delta\mathbf{S}_{Id}^k$ and $\Delta\mathbf{S}_{Exp}^k$, i.e.,

$$\Delta\mathbf{S}^k = [\Delta\mathbf{S}_{Id}^k; \Delta\mathbf{S}_{Exp}^k] = R_S^k(\Delta\mathbf{U}^k). \tag{5}$$

The 3D shape can be then updated by $\mathbf{S}^k = \mathbf{S}^{k-1} + \Delta\mathbf{S}_{Id}^k + \Delta\mathbf{S}_{Exp}^k$. The method for learning these shape regressors will be given in Sec. 3.5.

Refining landmarks Once a new estimate of the 3D shape is obtained, the landmarks can be further refined accordingly. For this purpose, the 3D-to-2D mapping matrix is needed. Hence, we estimate $\mathbf{M}^k$ based on $\mathbf{S}^k$ and $\hat{\mathbf{U}}^k$. The refined landmarks $\mathbf{U}^k$ can be then obtained by projecting $\mathbf{S}^k$ onto the image via $\mathbf{M}^k$ according to Eq. (3). During this process, the visibility of the landmarks is also re-computed. Details about this step will be given in Sec. 3.6.

3.3 Training Data Preparation

Before we provide the details about the three steps, we first introduce the training data needed for learning the landmarks and 3D shape regressors, which will also facilitate the understanding of our learning algorithms. Since the purpose of these regressors is to gradually adjust the estimated landmarks and 3D shape towards their ground truth values, we need a sufficient number of triplet data $\{(\mathbf{I}_i, \mathbf{S}_i^*, \mathbf{U}_i^*)|i = 1, 2, \cdots, N\}$, where $\mathbf{S}_i^*$ and $\mathbf{U}_i^*$ are, respectively, the ground truth 3D shape and landmarks for the image $\mathbf{I}_i$, and N is the total number of training samples. All the 3D face shapes have been established dense correspondences among their vertices; in other words, they have the same number of vertices, and vertices of the same index in the 3D face shapes have the same semantic meaning. Here, each of the ground truth 3D face shapes includes two parts, the PEN 3D face

shape $\mathbf{S}^*_{Id}$ and its expression shape $\mathbf{S}^*_{Exp} = \bar{\mathbf{S}} + \Delta\mathbf{S}^*_{Exp}$, i.e., $\mathbf{S}^* = [\mathbf{S}^*_{Id}; \mathbf{S}^*_{Exp}]$. Moreover, both visible and invisible landmarks in $\mathbf{I}_i$ have been annotated and included in $\mathbf{U}^*_i$. For invisible landmarks, the annotated positions should be anatomically correct positions (for example the red points in Fig. 1).

Obviously, to enable the regressors to cope with expression and pose variations, the training data should contain 2D face images of varying expressions and poses. It is, however, difficult to find in the public domain such data sets of 3D face shapes and corresponding annotated 2D images with various expressions/poses. Thus, we construct two sets of training data by ourselves: one based on BU3DFE [5], and the other based on 300W-LP [10], [48].

BU3DFE database contains 3D face scans of 56 females and 44 males, acquired in neutral plus six basic expressions (happiness, disgust, fear, anger, surprise and sadness). All basic expressions are acquired at four levels of intensity. These 3D face scans have been manually annotated with 84 landmarks (83 landmarks provided by the database plus one nose tip marked by ourselves). For each of the 100 subjects, we select the scans of neutral and the level-one intensity of the rest six expressions as the ground truth 3D face shapes. From each of the chosen seven scans of a subject, 19 face images are rendered at different poses ($-90°$ to $90°$ yaw with a $10°$ interval) with landmark locations recorded. As a result, each subject has 133 images of different poses and expressions. We use the method in [49] to establish dense correspondence of the 3D face scans of $n = 5,996$ vertices. With the registered 3D face scans, we compute the mean PEN 3D face shape by averaging all the subjects' PEN 3D face shapes, which are defined by their 3D face scans of frontal pose and neutral expression. All the 2D face images of one subject share the same PEN 3D face shape of the subject, while their expression shapes can be obtained by first subtracting from their corresponding 3D face scans their PEN 3D face shape and then adding the mean PEN 3D face shape.

300W-LP database [10] is created based on 300W [48] database, which is an integration of multiple face alignment benchmark datasets (i.e., AFW [22], LFPW [50], HELEN [51], IBUG [48] and XM2VTS [52]). It includes 122,450 in-the-wild face images of a large variety of poses and expressions. For each image, its corresponding registered PEN 3D face shape and expression shape are estimated by using the method in [3] based on BFM [53] and FaceWarehouse [54]. The obtained 3D face shapes have $n = 53,215$ vertices. Figure 5 shows some example 2D face images and corresponding PEN 3D face shapes and expression shapes in our constructed training datasets.

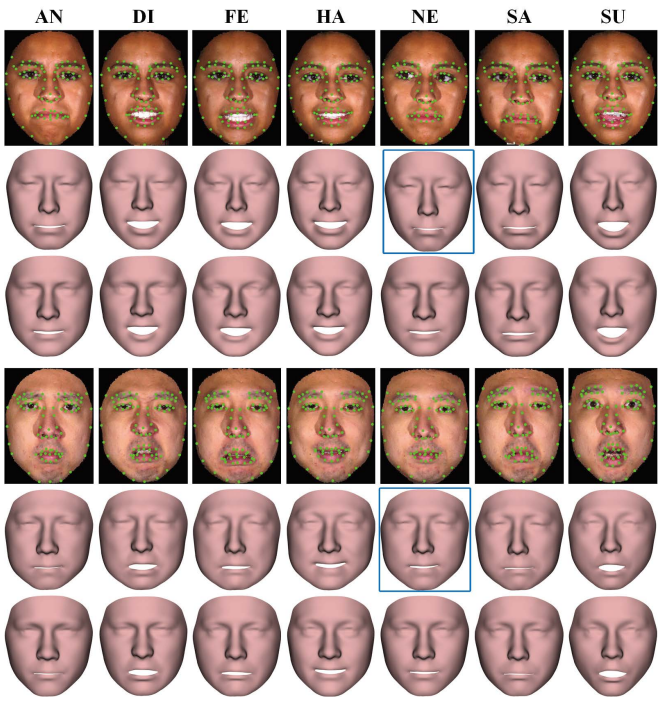

Fig. 4. Example 2D face images with annotated landmarks (1^{st} and 4^{th} rows), their 3D face shapes (2^{nd} and 5^{th} rows) and expression shapes (3^{rd} and 6^{th} rows) from the BU3DFE database. Seven expressions are shown: angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA), and surprise (SU). The 3D face shapes corresponding to the neutral expression are their PEN 3D face shapes, which are highlighted with blue boxes.

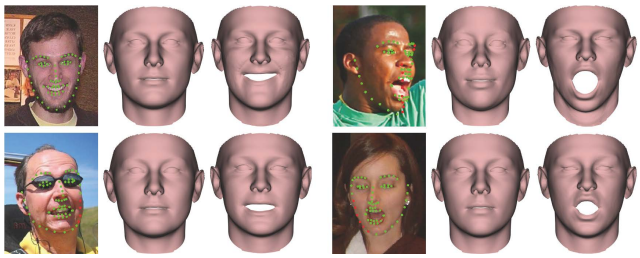

Fig. 5. Example data of four subjects in the 300W-LP database. From left to right: 2D face images with annotated landmarks, PEN 3D face shapes, and expression shapes.

3.4 Learning Landmark Regressors

According to Eq. (4), landmark regressors estimate the adjustment to $\mathbf{U}^{k-1}$ such that the updated landmarks $\mathbf{U}^k$ are closer to their true positions. In the training phase, the true positions and visibility of the landmarks are given by the ground truth $\mathbf{U}^*$. Therefore, the objective of the landmark regressors R^k_U is to better predict the difference between $\mathbf{U}^{k-1}$ and $\mathbf{U}^*$. In this paper, we employ linear regressors as the landmark regressors, and learn them by fulfilling the following optimization:

$$R^k_U = \arg\min_{R^k_U} \sum_{i=1}^{N} \| \left(\mathbf{U}^*_i - \mathbf{U}^{k-1}_i\right) - R^k_U(h(\mathbf{I}_i, \mathbf{U}^{k-1}_i)) \|^2_2, \tag{6}$$

which has a closed-form least-square solution. Note that other regression schemes, such as CNN [26], can be easily adopted in our framework.

We use 128-dim SIFT descriptors [55] as the local feature. The feature vector of h is a concatenation of the SIFT descriptors at all the l landmarks, i.e., a $128l$-dim vector. If a landmark is invisible, no feature will be extracted, and its corresponding entries of h will be zero. It is worth mentioning that the regressors estimate the semantic positions of all landmarks including invisible landmarks.

3.5 Learning 3D Shape Regressors

The landmark adjustment $\Delta\mathbf{U}^k$ is also used as the input to the 3D shape regressor R_S^k. The objective of R_S^k is to compute an update to the initially estimated 3D shape $\mathbf{S}^{k-1}$ in the k^{th} iteration to minimize the difference between the updated 3D shape and the ground truth. Using similar linear regressors, the 3D shape regressors can be learned by solving the following optimization via least squares:

$$R_S^k = \arg\min_{R_S^k} \sum_{i=1}^{N} \parallel (\mathbf{S}_i^* - \mathbf{S}_i^{k-1}) - R_S^k\left(\Delta\mathbf{U}_i^k\right) \parallel_2^2, \quad (7)$$

with its closed-form solution as

$$R_S^k = \Delta\mathbb{S}^k(\Delta\mathbb{U}^k)^\mathsf{T}(\Delta\mathbb{U}^k(\Delta\mathbb{U}^k)^\mathsf{T})^{-1}, \quad (8)$$

where $\Delta\mathbb{S}^k = \mathbb{S}^* - \mathbb{S}^{k-1}$ and $\Delta\mathbb{U}^k$ are, respectively, the 3D shape and landmark adjustment. $\mathbb{S}$ and $\mathbb{U}$ denote, respectively, the ensemble of 3D face shapes and 2D landmarks of all training samples with each column corresponding to one sample.

Since $\mathbb{S} \in \mathbb{R}^{6n\times N}$ (recall that $\mathbb{S}$ has two parts, PEN shape and expression deformation) and $\mathbb{U} \in \mathbb{R}^{2l\times N}$, it can be mathematically shown that N should be larger than $2l$ so that $\Delta\mathbb{U}^k(\Delta\mathbb{U}^k)^\mathsf{T}$ is invertible. Fortunately, since the set of used landmarks is usually sparse, this requirement can be easily satisfied in real-world applications.

3.6 Estimating 3D-to-2D Mapping and Landmark Visibility

In order to refine the landmarks with the updated 3D face shape, we have to project the 3D shape to the 2D image with a 3D-to-2D mapping matrix. In this paper, we dynamically estimate the mapping matrix based on $\mathbf{S}^k$ and $\hat{\mathbf{U}}^k$. As discussed earlier in Sec. 3.1, the mapping matrix is a composite effect of pose induced deformation and camera projection. Here, we assume a weak perspective projection for the camera projection as in prior work [25], [56]. As a result, the mapping matrix $\mathbf{M}^k$ is represented by a 2×4 matrix, and can be estimated as a least squares solution to the following fitting problem:

$$\mathbf{M}^k = \arg\min_{\mathbf{M}^k} \parallel \hat{U}^k - \mathbf{M}^k \times S_L^k \parallel_2^2 . \quad (9)$$

Once a new mapping matrix is computed, the landmarks can be further refined as $U^k = \mathbf{M}^k \times S_L^k$.

The visibility of the landmarks can be then computed based on the mapping matrix $\mathbf{M}$ using the method in [25]. Suppose the average surface normal around a landmark in the 3D face shape $\mathbf{S}$ is $\overrightarrow{\mathbf{n}}$. Its visibility $\mathbf{v}$ can be measured by

$$\mathbf{v} = \frac{1}{2}\left(1 + sgn\left(\overrightarrow{\mathbf{n}} \cdot \left(\frac{\mathbf{M}_1}{\|\mathbf{M}_1\|} \times \frac{\mathbf{M}_2}{\|\mathbf{M}_2\|}\right)\right)\right), \quad (10)$$

where $sgn()$ is the sign function, '$\cdot$' means dot product and '$\times$' cross-product, and $\mathbf{M}_1$ and $\mathbf{M}_2$ are the left-most three elements at the first and second row of the mapping matrix $\mathbf{M}$. This basically rotates the surface normal and validates if it points toward the camera or not.

The whole process of learning the cascaded coupled landmark and 3D shape regressors is summarized in Algorithm 1.

Algorithm 1 Cascaded Coupled-Regressor Learning.

Input: Training data $\{(\mathbf{I}_i, \mathbf{S}_i^*, \mathbf{U}_i^*)|i = 1, 2, \cdots, N\}$, initial shape $\mathbf{S}_i^0$ & landmarks $\mathbf{U}_i^0$.

Output: Cascaded coupled-regressors $\{R_U^k, R_S^k\}_{k=1}^K$.

1: **for** $k = 1, ..., K$ **do**
2: Estimate R_U^k via Eq. (6), and compute landmark adjustment $\Delta\mathbf{U}_i^k$ via Eq. (4);
3: Update landmarks $\hat{\mathbf{U}}_i^k$ for all images: $\hat{\mathbf{U}}_i^k = \mathbf{U}_i^{k-1} + \Delta\mathbf{U}_i^k$;
4: Estimate R_S^k via Eq. (7), and compute shape adjustment $\Delta\mathbf{S}_i^k$ via Eq. (5);
5: Update 3D face $\mathbf{S}_i^k$: $\mathbf{S}_i^k = \mathbf{S}_i^{k-1} + \Delta\mathbf{S}_i^k$;
6: Estimate the 3D-to-2D mapping matrix $\mathbf{M}_i^k$ via Eq. (9);
7: Compute the refined landmarks $\mathbf{U}_i^k$ via Eq. (3) and their visibility via Eq. (10).
8: **end for**

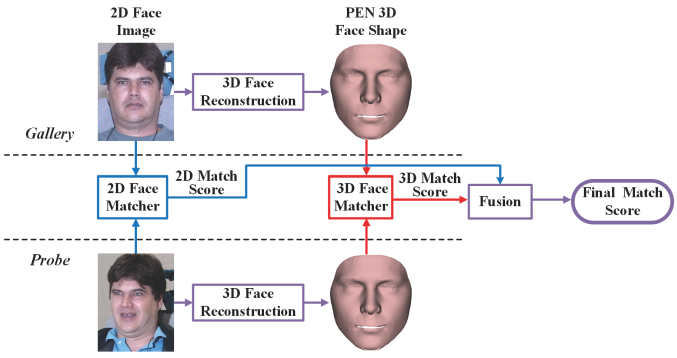

Fig. 6. Block diagram of the proposed 3D-enhanced face recognition method.

4 Application to Face Recognition

In this section we explore the reconstructed 3D faces to improve face recognition accuracy on off-angle and expressive face images. The basic idea is to utilize the additional feature provided by the reconstructed PEN 3D face shapes and fuse it with conventional 2D face matchers. Figure 6 shows the 3D-enhanced face recognition method employed in this paper. As can be seen, 3D face reconstruction methods are applied to both gallery and probe face images to generate PEN 3D face shapes. The iterative closest point (ICP) algorithm [57] is applied to match the reconstructed normalized 3D face shapes. It aligns the 3D shapes reconstructed from probe and gallery images, and computes the distances between them, which are then converted to similarity scores via subtracting them from the maximum distance. These scores are finally normalized to the range of $[0, 1]$ via min-max normalization, and fused with the scores of the the conventional 2D face matcher on the gallery and probe face images (which are within $[0, 1]$ also) by using a sum rule. The recognition result for a probe is defined as the subject whose gallery sample has the highest match score with it. Note that we employ the ICP-based 3D face shape matcher and the sum fusion rule for simplicity sake. Other more elaborated 3D face matchers and fusion rules can be also applied with our proposed method. Thanks to the additional discriminative feature in PEN 3D face shapes

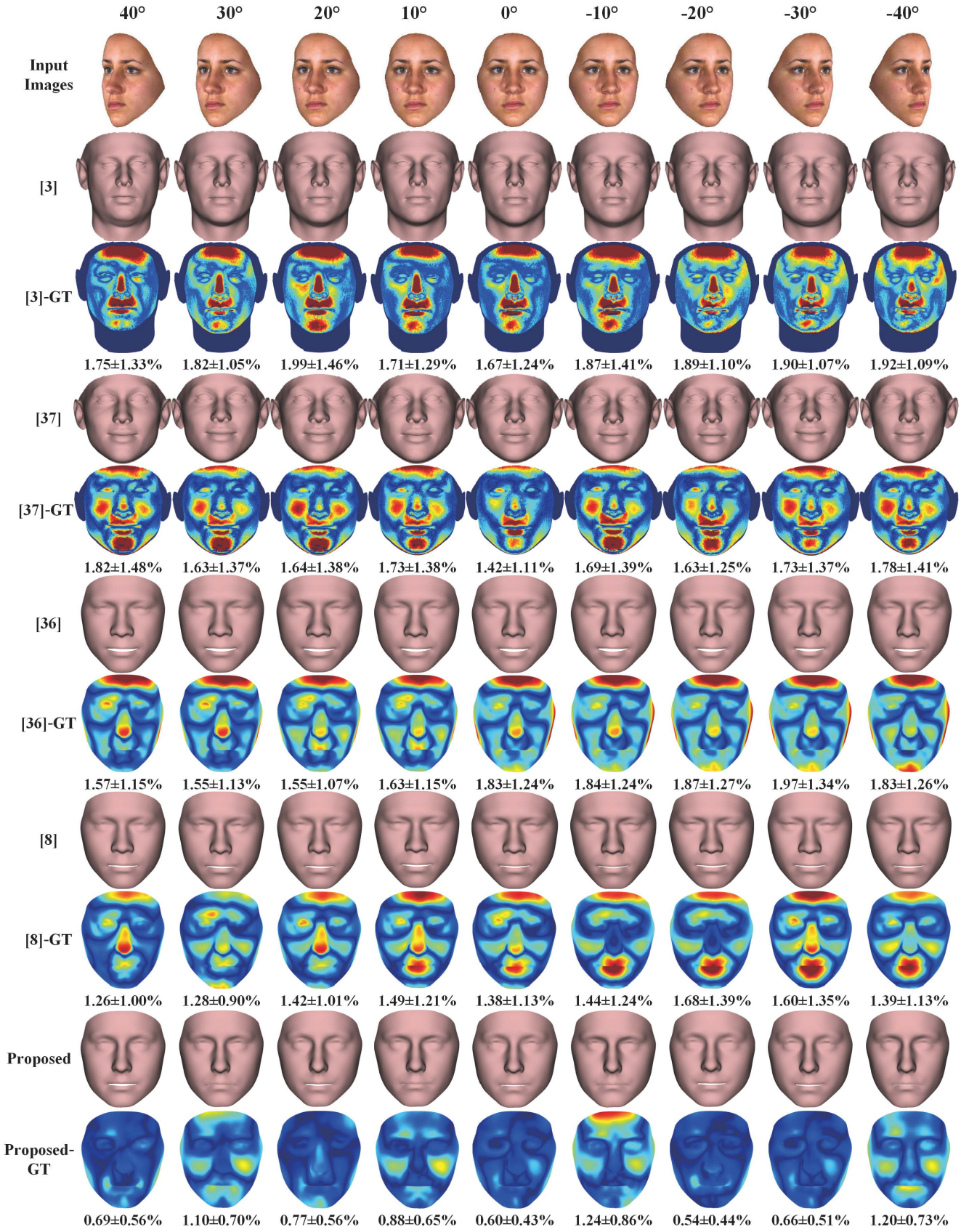

Fig. 7. Reconstruction results for a BU3DFE subject at nine different pose angles. First row: The input images. Second, forth, sixth, eighth and tenth rows: The reconstructed 3D face shapes by [3], [37], [36], [8] and the proposed method. Third, fifth, seventh, ninth and eleventh rows: Their corresponding NPDE maps. The colormap goes from dark blue to dark red (corresponding to an error between 0 and 5). The numbers under each of the error maps represent mean and standard deviation values (in %).

TABLE 1
3D face reconstruction accuracy (MAE) of the proposed method and state-of-the-art methods at different yaw poses on the BU3DFE database.

Method	±90°	±80°	±70°	±60°	±50°	±40°	±30°	±20°	±10°	0°	Avg.
Zhu et al. [3]	-	-	-	-	-	2.73	2.74	2.56	2.32	2.22	2.51
Tran et al. [37]	-	-	-	-	-	2.26	2.19	2.16	2.08	2.06	2.15
Liu et al. [36]	1.95	1.91	1.95	1.96	1.97	1.97	1.96	1.98	2.01	2.03	1.97
Liu et al. [8]	1.92	1.89	1.90	1.93	1.95	1.93	1.93	1.95	1.98	2.01	1.94
Proposed	**1.85**	**1.83**	**1.83**	**1.83**	**1.86**	**1.89**	**1.90**	**1.91**	**1.90**	**1.91**	**1.87**

and its robustness to pose and expression variations, the accuracy of conventional 2D face matchers on off-angle and expressive face images can be effectively improved after fusion with the PEN 3D face shape based matcher. In the next Section, we will experimentally demonstrate this.

5 Experiments

We conduct three sets of experiments to evaluate the proposed method in 3D face reconstruction, face alignment, and benefits to face recognition.

5.1 3D Face Reconstruction Accuracy

To evaluate the 3D shape reconstruction accuracy, a 10-fold cross validation is applied to split the BU3DFE data into training and testing subsets, resulting in 11,970 training samples and 1,330 testing samples. We compare the proposed method with its preliminary version in [8] and another three state-of-the-art methods in [36], [3] and [37]. The methods in [8] and [37] reconstruct PEN 3D face shapes only, while the methods in [36] and [3] reconstruct 3D face shapes that have the same pose and expression as the input images. Moreover, the method in [36] requires that visible landmarks are available together with the input images. In the following experiments, we use the visible landmarks projected from ground truth 3D face shapes for the method in [36]. For the methods of [3] and [37], we use the implementation provided by the authors. In the implementation, these two methods are based on the 68 landmarks that are detected by using the method in [58]. As a result, they can not be applied to face images of extreme pose angles (i.e., beyond 40 degrees).

Two metrics are used to evaluate the 3D face shape reconstruction accuracy: Mean Absolute Error (MAE) and Normalized Per-vertex Depth Error (NPDE). MAE is defined as [59]

$$\text{MAE} = \frac{1}{N_T}\sum_{i=1}^{N_T}(\|\mathbf{S}_i^* - \hat{\mathbf{S}}_i\|/n), \tag{11}$$

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^{th} testing sample.

NPDE measures the depth error at the j^{th} vertex in a testing sample as [30]

$$\text{NPDE}(x_j, y_j) = \left(|z_j^* - \hat{z}_j|\right) / \left(z_{max}^* - z_{min}^*\right), \tag{12}$$

where z_{max}^* and z_{min}^* are the maximum and minimum depth values in the ground truth 3D shape of the testing sample, and z_j^* and $\hat{z}_j$ are the ground truth and reconstructed depth values at the j^{th} vertex.

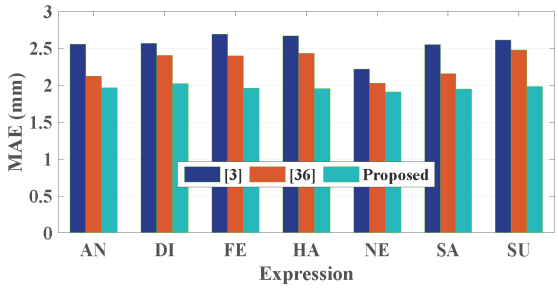

Fig. 8. 3D face reconstruction accuracy (MAE) of the proposed method, [36] and [3] under different expressions. i.e., angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA) and surprise (SU).

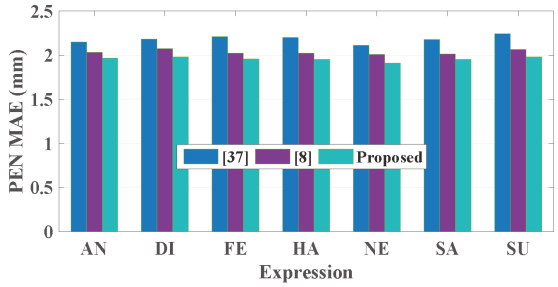

Fig. 9. PEN 3D face reconstruction accuracy (MAE) of the proposed method, [8] and [37] under different expressions. i.e., angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA) and surprise (SU).

Reconstruction accuracy across poses Table 1 shows the average MAE of the proposed method under different pose angles of the input 2D images. For a fair comparison with the counterpart methods, we only compute the reconstruction error of neutral testing images. To compute MAE, the reconstructed 3D faces should be first aligned to the ground truth. Since the results of [8], [36] and our proposed method already have the same number of vertices as the ground truth, we employ Procrustes alignment for these methods as being suggested by [60]. For the results of [3] and [37], however, the number of vertices is different from the ground truth. Hence, we align them by using rigid iterative closest point method as [37] does. It can be seen from Table 1 that the average MAE of the proposed method is lower than that of counterpart methods. Moreover, as the pose angle becomes large, the error does not increase substantially. This proves the effectiveness of the proposed method in handling arbitrary view face images. Figure 7

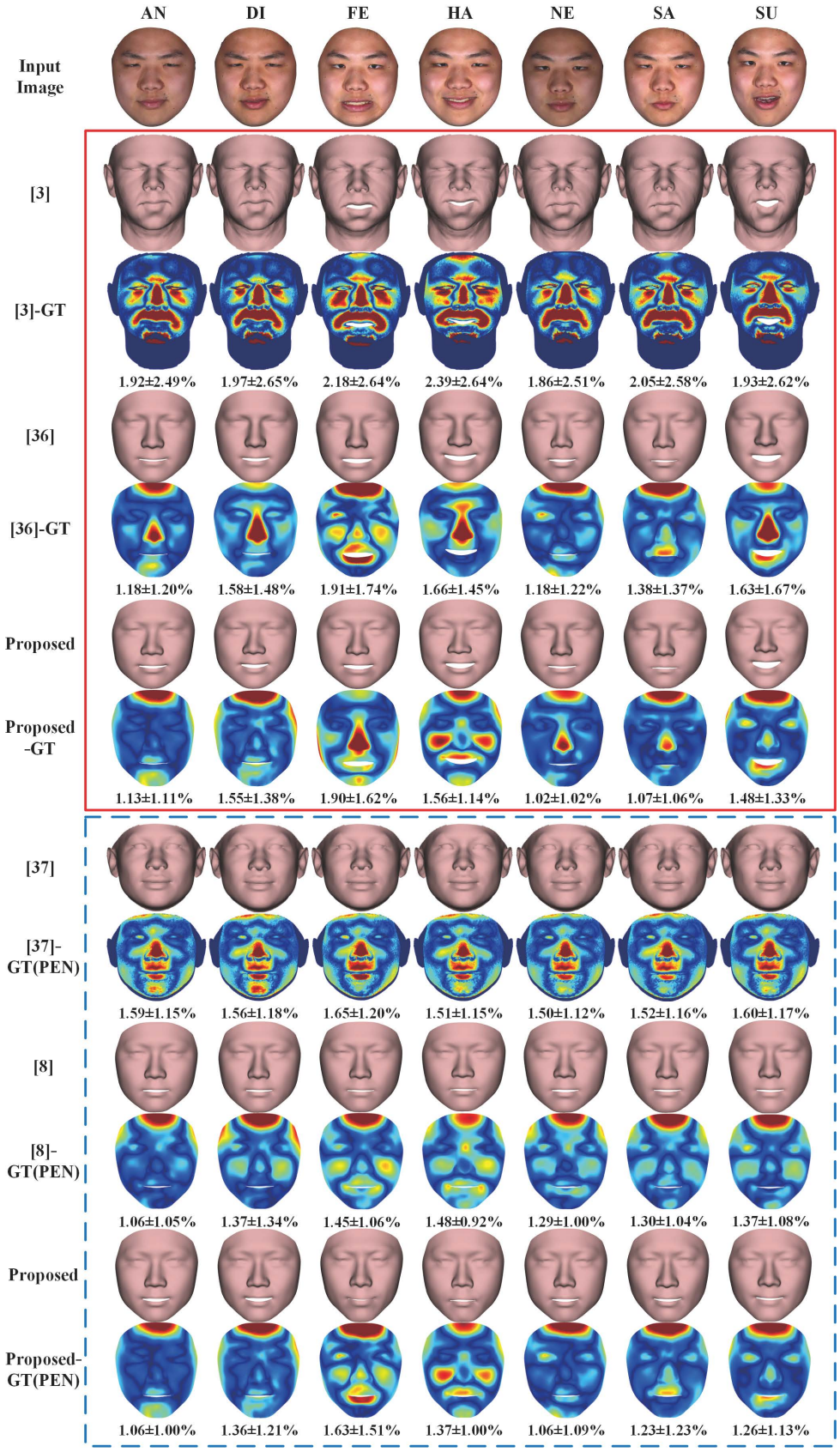

Fig. 10. Reconstruction results for a BU3DFE subject under seven different expressions. The first row shows the input images. In the red box, we show the reconstructed 3D shapes that have the same expression as the input images, using the methods of [36], [3] and the proposed method. In the blue box, we show the reconstructed PEN 3D shapes obtained by [8], [37] and the proposed method. The NPDE maps of these results are shown below the reconstructed 3D faces, which go from dark blue to dark red (corresponding to an error between 0 and 5). The numbers under each of the error maps represent mean and standard deviation values (in %).

TABLE 2
Number and percentage of subjects of different genders and races in the FRGC v2.0 database.

	Asian	Black	Hispanic	White	Unknown	Total
Female	55 (11.8%)	2 (0.4%)	5 (1.1%)	134 (28.8%)	6 (1.3%)	202 (43.3%)
Male	57 (12.2%)	4 (0.9%)	8 (1.7%)	185 (39.7%)	10 (2.1%)	264 (56.7%)
Total	112 (24.0%)	6 (1.3%)	13 (2.8%)	319 (68.5%)	16 (3.4%)	466 (100%)

shows the reconstruction results of one subject.

Reconstruction accuracy across expressions Figure 8 shows the average MAE of the proposed method and the methods in [36] and [3] across expressions based on their reconstructed 3D face shapes that have the same pose and expression as the input images. The proposed method overwhelms its counterpart method for all different expressions. Moreover, as expressions change, the maximum MAE variance of the methods in [3] and [36] are 15.7% and 17.9%, whereas that of the proposed method is 3.4%. This proves the superior robustness of the proposed method to expression variations.

Figure 9 compares the average MAE of the proposed method and the methods in [8] and [37] across expressions based on their reconstructed PEN 3D face shapes. Again, the proposed method shows superior performance both in terms of MAE under all different expressions and robustness across expressions. We believe that the improvement achieved by the proposed method is owed to its explicit modelling of expression deformation. Figure 10 shows the reconstruction results for a subject under seven expressions.

Reconstruction accuracy across races It is well known that people from different races (e.g., Asian and Western people) show different characteristics in their facial shapes. Such other-race effect has been reported in the face recognition literature [61]. In this experiment, we study the impact of races on the 3D face reconstruction accuracy based on the FRGC v2.0 database [11]. FRGC v2.0 contains 3D face models and 2D face images of 466 subjects who are from different ethnic groups (see Table 2). Since these face data have no variation in expression, the expression shape component in our proposed model is set to zero. We use the method in [49] to establish dense correspondence of the 3D faces of $n = 5,996$ vertices. We conduct three series of experiments: (i) training using 100 Asian samples (denoted as **Setting I**), (ii) training using 100 White samples (denoted as **Setting II**), and (iii) training using 100 Asian and 100 White samples (denoted as **Setting III**). The testing set contains the samples of the remaining subjects in FRGC v2.0, including 12 Asian, 6 Black, 13 Hispanic, 19 White and 16 Unknown races.

Figure 11 compares the 3D face reconstruction accuracy (MAE) across different ethnic groups. Not surprisingly, training for one ethnic group can yield better accuracy on the same ethnic testing samples. As for the other-race effect, the model trained on White achieves comparable accuracy on White and Hispanic, but much worse accuracy on the other races (and worst on Asian). On the other hand, the model trained on Asian performs much worse on all the other races compared with on its own race, and obtains the worse accuracy on Black. These results reveal the variations in the facial shapes of people from different races. Furthermore, by combining the training data of Asian and White (Setting III), comparable reconstruction accuracy is achieved for both Asian and White, which is also comparable to those in Setting I and Setting II. This proves the capacity of the proposed method in handling the 3D facial shape variations among people from different ethnic groups.

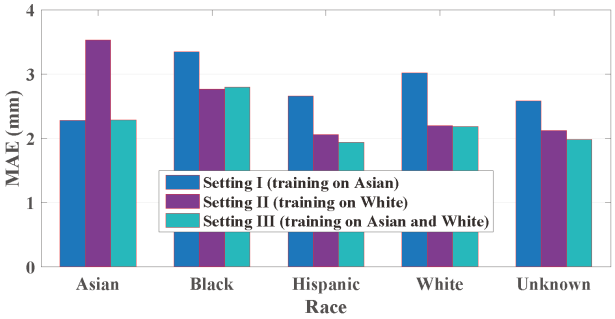

Fig. 11. 3D face reconstruction accuracy (MAE) of the proposed method across different ethnic groups.

5.2 Face Alignment Accuracy

In face alignment accuracy evaluation, several state-of-the-art face alignment methods are considered for comparison to the proposed method, including RCPR [62], ESR [19], SDM [18], 3DDFA and 3DDFA+SDM [10]. The dataset constructed from 300W-LP is used for training, the AFLW [9] and AFLW2000-3D [10] are used for testing. AFLW contains 25,993 in-the-wild faces with large-pose variations (yaw from -90° to 90°). Each image is annotated with up to 21 visible landmarks. For a fair comparison to [10], we use the same 21,080 samples as our testing set, and divide the testing set into three subsets according to the absolute yaw angles of the testing images: $[0^\circ, 30^\circ)$, $[30^\circ, 60^\circ)$ and $[60^\circ, 90^\circ]$. The resulting three subsets have 11,596, 5,457 and 4,027 samples, respectively. AFLW2000-3D contains the ground truth 3D faces and the corresponding 68 landmarks of the first 2,000 AFLW samples. There are 1,306 samples in $[0^\circ, 30^\circ)$, 462 samples in $[30^\circ, 60^\circ)$ and 232 samples in $[60^\circ, 90^\circ]$. The bounding boxes provided by AFLW are used in the AFLW testing experiment, while the ground truth bounding boxes enclosing all 68 landmarks are used for the AFLW2000-3D testing experiment.

Normalized Mean Error (NME) [25] is employed to measure the face alignment accuracy. It is defined as the mean of the normalized estimation error of visible landmarks for all testing samples:

$$\text{NME} = \frac{1}{N_T}\sum_{i=1}^{N_T}\left(\frac{1}{d_i}\frac{1}{N_i^{\mathbf{v}}}\sum_{j=1}^{l}\mathbf{v}_{ij}||(\hat{u}_{ij},\hat{v}_{ij}) - (u_{ij}^*, v_{ij}^*)||\right), \tag{13}$$

where d_i is the square root of the face bounding box area of the i^{th} testing sample, $N_i^{\mathbf{v}}$ is the number of visible landmarks in it, (u_{ij}^*, v_{ij}^*) and $(\hat{u}_{ij}, \hat{v}_{ij})$ are, respectively, the ground truth and estimated coordinates of its j^{th} landmark.

Table 3 provides the face alignment accuracy of different methods on the AFLW and AFLW2000-3D datasets. As can

TABLE 3
The face alignment accuracy (NME, %) of the proposed method and existing state-of-the-art methods on AFLW and AFLW2000-3D databases.

Method	AFLW Database (21 points)					AFLW2000-3D Database (68 points)				
	$[0°, 30°)$	$[30°, 60°)$	$[60°, 90°]$	Mean	Std	$[0°, 30°)$	$[30°, 60°)$	$[60°, 90°]$	Mean	Std
RCPR [62]	5.43	6.58	11.53	7.85	3.24	4.26	5.96	13.18	7.80	4.74
ESR [19]	5.66	7.12	11.94	8.24	3.29	4.60	6.70	12.67	7.99	4.19
SDM [18]	4.75	5.55	9.34	6.55	2.45	3.67	4.94	9.76	6.12	3.21
3DDFA [10]	5.00	5.06	6.74	5.60	0.99	3.78	4.54	7.93	5.42	2.21
3DDFA+SDM [10]	4.75	4.83	6.38	5.32	0.92	3.43	4.24	7.17	4.94	1.97
Proposed	**3.75**	**4.33**	**5.39**	**4.49**	**0.83**	**3.25**	**3.95**	**6.42**	**4.61**	**1.78**

Fig. 12. The 68 landmarks detected by the proposed method for images in AFLW. Green/red points denote visible/invisible landmarks.

be seen, the proposed method achieves the best accuracy among all the considered methods for all poses and on both datasets. In order to assess the robustness of different methods to pose variations, we also report the standard deviations of the NME of different methods in Table 3. The results again demonstrate the superiority of the proposed method over the counterpart methods. Figure 12 shows the landmarks detected by proposed method on some example images in AFLW.

5.3 Benefits to Face Recognition

While there are many recent face alignment and reconstruction work [63], [64], [65], [66], [67], few work takes one step further to evaluate the contribution of alignment or reconstruction to subsequent tasks. In contrast, we quantitatively evaluate the effect of the reconstructed pose-expression-normalized (PEN) 3D face shapes on face recognition by performing direct 3D to 3D shape matching and fuse it with conventional 2D face recognition. Refer to Sec. 4 for details of the PEN 3D face shapes enhanced face recognition method.

In this evaluation, we use the BU3DFE (13,300 images of 100 subjects; refer to Sec. 3.3) and MICC [68] databases as training data, and the CMU Multi-PIE database [12] as test data. MICC contains 3D face scans and video clips (indoor, outdoor and cooperative head rotations environments) of 53 subjects. We randomly select face images with different poses from the cooperative environment videos, resulting in 11,788 images of 53 subjects and their corresponding neutral 3D face shapes (whose expression shape components are thus set to zero). The 3D faces are processed by the method in [49] to establish dense correspondence with $n = 5,996$ vertices.

CMU Multi-PIE is a widely used benchmark database for evaluating face recognition accuracy under pose, illumination and expression variations. It contains 2D face images of 337 subjects collected under various views, expressions and lighting conditions. Here, we consider pose and expression variations, and conduct two experiments. In the first experiment, following the setting of [3], [69], probe images consist of the images of all the 337 subjects at 12 poses ($\pm 90°$, $\pm 75°$, $\pm 60°$, $\pm 45°$, $\pm 30°$, $\pm 15°$) with neutral expression and frontal illumination. In the second experiment, instead of the neutral expression images, all the images with smile, surprise, squint, disgust and scream expressions at the 12 poses and under frontal illumination are used as probe images. This protocols is a extended and modified version of [4] and [3] by using more large pose images ($\pm 60°$, $\pm 75°$, $\pm 90°$). In both experiments, the frontal images of the subjects captured in the first session are used as gallery. And four state-of-the-art deep learning based (DL-based) face matchers are used as baseline 2D face matchers, i.e., VGG [70], Lightened CNN [71], CenterLoss [72] and LDF-Net [47]. The first three matchers are publicly available. We evaluate them with all the 337 subjects in Multi-PIE. The last matcher, LDF-Net, is a latest matcher specially designed for pose-invariant face recognition. It uses the first 229 subjects for training and the remaining 108 subjects for testing. Since it is not publicly available, we request the match scores from the authors, and fuse our 3D shape match scores with theirs. Note that given the good performance of LDF-Net, we assign higher weight (i.e., 0.7) to it, whereas the weight for all the other three baseline matchers is set as 0.5.

TABLE 4
Recognition accuracy in the first experiment on Multi-PIE by the four state-of-the-art DL-based face matchers before (indicated by suffix "2D") and after (indicated by suffix "Fusion") the enhancement by our proposed method. Avg. is the average accuracy.

Method	$\pm 90°$	$\pm 75°$	$\pm 60°$	$\pm 45°$	$\pm 30°$	$\pm 15°$	**Avg.**
VGG-2D	36.2%	66.9%	83.5%	93.8%	97.7%	98.6%	79.5%
Lightened CNN-2D	7.5%	31.5%	78.6%	96.3%	99.1%	99.8%	68.8%
CenterLoss-2D	48.2%	72.7%	92.6%	98.8%	99.6%	99.7%	85.3%
LDF-Net-2D	65.3%	86.2%	93.7%	98.4%	98.9%	98.6%	90.2%
VGG-Fusion	52.6%	75.2%	90.5%	96.8%	98.5%	99.4%	85.5%
Lightened CNN-Fusion	23.6%	45.3%	84.6%	97.6%	99.6%	99.9%	75.1%
CenterLoss-Fusion	63.7%	76.7%	92.5%	97.8%	98.4%	98.7%	88.0%
LDF-Net-Fusion	70.4%	87.6%	93.4%	98.1%	97.9%	97.7%	90.9%

TABLE 5
Recognition accuracy of the CenterLoss matcher in the second experiment on Multi-PIE. The results shown in brackets are obtained by using the original CenterLoss matcher without enhancement by our reconstructed 3D shapes.

Pose \ Expression	Smile	Surprise	Squint	Disgust	Scream	**Avg.**
$\pm 90°$	51.4%(36.9%)	46.1%(35.7%)	58.8%(38.7%)	42.0%(24.9%)	63.6%(52.4%)	52.4%(37.7%)
$\pm 75°$	73.1%(67.0%)	56.6%(53.0%)	72.6%(67.8%)	52.5%(43.4%)	75.1%(71.6%)	66.0%(60.4%)
$\pm 60°$	88.6%(89.8%)	80.2%(80.7%)	91.6%(88.2%)	74.6%(69.8%)	91.8%(92.7%)	85.4%(84.2%)
$\pm 45°$	95.9%(97.6%)	89.4%(95.1%)	95.6%(97.8%)	86.7%(83.5%)	97.3%(98.7%)	93.0%(94.5%)
$\pm 30°$	97.8%(99.1%)	93.1%(97.0%)	96.8%(99.3%)	90.4%(91.5%)	98.5%(99.8%)	95.3%(97.3%)
$\pm 15°$	98.5%(99.6%)	95.6%(97.3%)	97.5%(100%)	92.6%(93.5%)	98.1%(99.2%)	96.5%(97.9%)
Avg.	84.2%(81.7%)	76.8%(76.5%)	85.5%(82.0%)	73.1%(67.8%)	87.4%(85.7%)	81.4%(78.7%)

Table 4 reports the rank-1 recognition accuracy of the baseline face matchers in the first experiment. According to the results in Table 4, the baseline matchers are all further improved with our proposed method. Specifically, VGG and Lightened CNN are consistently improved across different pose angles when fused with 3D cues, while CenterLoss gains substantial improvement at large pose angles (15.5% at $\pm 90°$ and 4.0% at $\pm 75°$). Even for the best LDF-Net method, the recognition accuracy is improved by 5.1% at $\pm 90°$ and 1.4% at $\pm 75°$. For all the baseline matchers, the larger the yaw angle is, the more evident the accuracy improvement. This proves the effectiveness of the proposed method in dealing with pose variations, as well as in reconstructing individual 3D face shapes with discriminative details that are beneficial to face recognition.

Given its best performance among the three publicly available baseline matchers, we employ the CenterLoss matcher in the second experiment. The results are shown in Table 5. As can be seen, the compound impact of pose and expression variations makes the face recognition more challenging, resulting in obviously lower accuracy compared with the results in Table 4. Yet, our proposed method still improves the overall accuracy of the baseline matcher, especially for the probe face images of large pose or disgust expression. We believe that such performance gain in recognizing non-frontal and expressive faces is owing to the capability of the proposed method in providing complementary pose-and-expression-invariant discriminative features in 3D face shape space.

5.4 Convergence

The proposed method has two alternate optimization processes, one in 2D space for face alignment and the other in 3D space for 3D shape reconstruction. We experimentally investigate the convergence of these two processes when training the proposed method on the BU3DFE database. We conduct ten-fold cross-validation experiments, and compute the objective function values through ten iterations. The average results are shown in Fig. 13. It can be seen that both optimization processes converge in about five iterations. Hence, in our experiments, we set the number of iterations as $K = 5$.

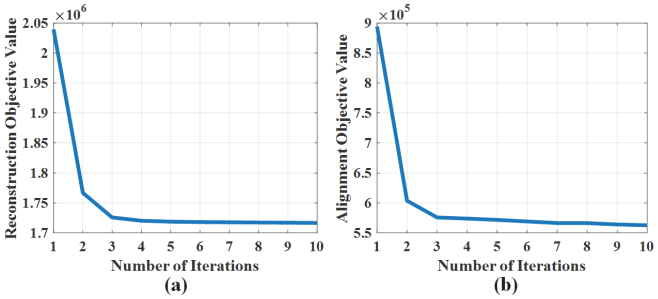

Fig. 13. The reconstruction (a) and alignment (b) objective function values of the proposed method as iteration proceeds, when trained on the BU3DFE database.

5.5 Computational Complexity

According to our experiments on a PC with i7-4790 CPU and 32 GB memory, the Matlab implementation of the proposed method runs at $\sim$ 26 FPS ($K = 5$ and $n = 5,996$). This indicates that the proposed method can detect landmarks and reconstruct 3D face shape in real time. We also report the consumed time of every step in Table 6, and comparison with other existing methods in Table 7.

6 Conclusion

In this paper, we present a novel regression based method for joint face alignment and 3D face reconstruction from sin-

TABLE 6
The time efficiency (in milliseconds or ms) of the proposed method.

Step	Updating landmarks	Updating shape	Refining landmarks	Total
Time (ms)	14.9	15.3	8.7	38.9

TABLE 7
Efficiency comparison of different reconstruction methods. For the methods [3], [36] , and [37], stand-alone landmark detection is required, but the time here does not include the landmark detection time.

Method	[3]	[37]	[36]	[8]	Proposed
Time (ms)	56.3	88	12.6	32.8	38.9

gle 2D images of arbitrary poses and expressions. It utilizes landmarks on a 2D face image as clues for reconstructing 3D shapes, and uses the reconstructed 3D shapes to refine landmarks. By alternately applying cascaded landmark regressors and 3D shape regressors, the proposed method can effectively accomplish the two tasks simultaneously in real time. Unlike existing 3D face reconstruction methods, the proposed method does not require additional face alignment methods, but can fully automatically reconstruct both pose-and-expression-normalized and expressive 3D shapes from a single face image of arbitrary poses and expressions. Compared with existing face alignment methods, the proposed method can effectively handle invisible and expression-deformed landmarks with the assistance of 3D face models. Extensive experiments with comparison to state-of-the-art methods demonstrate the effectiveness and superiority of the proposed method in both face alignment and 3D face shape reconstruction, and in facilitating cross-view and cross-expression face recognition as well.

REFERENCES

[1] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *TPAMI*, vol. 25, no. 9, pp. 1063–1074, 2003.

[2] H. Han and A. K. Jain, "3D face texture modeling from uncalibrated frontal and profile images," in *BTAS*, 2012, pp. 223–230.

[3] 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*, 2015, pp. 787–796.

[4] B. Chu, S. Romdhani, and L. Chen, "3D-aided face recognition robust to expression and pose variations," in *CVPR*, 2014, pp. 1907–1914.

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

[6] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3D shape regression for real-time facial animation," *TOG*, vol. 32, no. 4, p. 41, 2013.

[7] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, "Real-time facial animation with image-based dynamic avatars," *TOG*, vol. 35, no. 4, pp. 126:1–126:12, 2016.

[8] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Joint face alignment and 3D face reconstruction," in *ECCV*, 2016, pp. 545–560.

[9] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *ICCVW*, 2011, pp. 2144–2151.

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

[11] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *CVPR*, vol. 1, 2005, pp. 947–954.

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *IVC*, vol. 28, no. 5, pp. 807–813, 2010.

[13] T. F. Cootes and A. Lanitis, "Active shape models: Evaluation of a multi-resolution method for improving image search," in *BMVC*, 1994, pp. 327–338.

[14] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models." in *BMVC*, 2007, pp. 1–10.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *TPAMI*, no. 6, pp. 681–685, 2001.

[16] I. Matthews and S. Baker, "Active appearance models revisited," *IJCV*, vol. 60, no. 2, pp. 135–164, 2004.

[17] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[18] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[19] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[20] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *CVPR*, 2014, pp. 1685–1692.

[21] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *CVPR*, 2015, pp. 4998–5006.

[22] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[23] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *ICCV*, 2013, pp. 1944–1951.

[24] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *FG*, vol. 1, 2015, pp. 1–8.

[25] A. Jourabloo and X. Liu, "Pose-invariant 3D face alignment," in *ICCV*, 2015, pp. 3694–3702.

[26] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *CVPR*, 2016, pp. 4188–4196.

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

[28] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194.

[29] S. Tulyakov and N. Sebe, "Regressing a 3D face shape from a single image," in *ICCV*, 2015, pp. 3748–3755.

[30] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *TPAMI*, vol. 33, no. 2, pp. 394–405, 2011.

[31] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total moving face reconstruction," in *ECCV*, 2014, pp. 796–812.

[32] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *CVPR*, vol. 2, 2005, pp. 986–993.

[33] G. Hu, F. Yan, J. Kittler, W. Christmas, C. H. Chan, Z. Feng, and P. Huber, "Efficient 3D morphable face model fitting," *PR*, vol. 67, pp. 366–379, 2017.

[34] Y. J. Lee, S. J. Lee, K. R. Park, J. Jo, and J. Kim, "Single view-based 3D face reconstruction robust to self-occlusion," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–20, 2012.

[35] C. Qu, E. Monari, T. Schuchert, and J. Beyerer, "Fast, robust and automatic 3D face model reconstruction from videos," in *AVSS*, 2014, pp. 113–118.

[36] 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.

[37] 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, in press*, 2017.

[38] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3D face recognition under expressions, occlusions, and pose variations," *TPAMI*, vol. 35, no. 9, pp. 2270–2283, 2013.

[39] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[40] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," *arXiv:1502.00873*, 2015.

[41] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" *arXiv:1501.04690*, 2015.

[42] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on rgb, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications." *TPAMI*, vol. 38, no. 8, pp. 1–1, 2016.

[43] 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 *WACV*, 2016, pp. 1–9.

[44] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *CVPR, in press*, 2017.

[45] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *CVPR*, 2013, pp. 3539–3545.

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

[47] L. Hu, M. Kan, S. Shan, X. Song, and X. Chen, "LDF-Net: Learning a displacement field network for face recognition across pose," in *FG*, in press, 2017.

[48] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *ICCVW*, 2013, pp. 397–403.

[49] T. Bolkart and S. Wuhrer, "3D faces in motion: Fully automatic registration and statistical analysis," *CVIU*, vol. 131, pp. 100–115, 2015.

[50] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *CVPR*, 2011, pp. 545–552.

[51] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *ICCVW*, 2013, pp. 386–391.

[52] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *AVBPA*, vol. 964, 1999, pp. 965–966.

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

[54] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3D facial expression database for visual computing," *TVCG*, vol. 20, no. 3, pp. 413–425, 2014.

[55] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[56] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, "3D shape estimation from 2D landmarks: A convex relaxation approach," in *CVPR*, 2015, pp. 4447–4455.

[57] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," *IVC*, pp. 2724–2729, 1991.

[58] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.

[59] Z. Lei, Q. Bai, R. He, and S. Z. Li, "Face shape recovery from a single image using cca mapping between tensor spaces," in *CVPR*, 2008, pp. 1–7.

[60] 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*, 2016, pp. 377–391.

[61] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Tran. on Applied Perception (TAP)*, vol. 8, no. 2, p. 14, 2011.

[62] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *ICCV*, 2013, pp. 1513–1520.

[63] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "DenseReg: Fully convolutional dense shape regression in-the-wild," in *CVPR*, in press, 2017.

[64] O. Tuzel, T. K. Marks, and S. Tambe, "Robust face alignment using a mixture of invariant experts," in *ECCV*, 2016, pp. 825–841.

[65] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *ECCV*, 2016, pp. 38–56.

[66] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3D morphable model learnt from 10,000 faces," in *CVPR*, 2016, pp. 5543–5552.

[67] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in press, 2017.

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

[69] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *NIPS*, 2014, pp. 217–225.

[70] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[71] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *arXiv:1511.02683*, 2015.

[72] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

Feng Liu received the M.Sc. degrees in computer science in 2014. He is currently a Ph.D. candidate at the National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, China. His main research interests focus on computer vision and pattern recognition, specifically for face modeling, 2D and 3D face recognition. He is a student member of the IEEE.

Qijun Zhao obtained B.Sc. and M.Sc. degrees both from Shanghai Jiao Tong University, and Ph.D. degree from the Hong Kong Polytechnic University. He worked as a post-doc researcher in the Pattern Recognition and Image Processing lab at Michigan State University from 2010 to 2012. He is currently an associate professor in College of Computer Science at Sichuan University. His research interests lie in biometrics, particularly, face recognition, fingerprint recognition, and affective computing. Dr. Zhao has published more than 50 papers in academic journals and conferences, and participated in many research projects either as principal investigators or as primary researchers. He is a program committee co-chair of the 2016 Chinese Conference on Biometric Recognition and the 2018 IEEE International Conference on Identity, Security and Behavior Analysis.

Xiaoming Liu is an Assistant Professor at the Department of Computer Science and Engineering of Michigan State University. He received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2004. Before joining MSU in Fall 2012, he was a research scientist at General Electric (GE) Global Research. His research interests include computer vision, machine learning, and biometrics. As a co-author, he is a recipient of Best Industry Related Paper Award runner-up at ICPR 2014, Best Student Paper Award at WACV 2012 and 2014, and Best Poster Award at BMVC 2015. He has been the Area Chair for numerous conferences, including FG, ICPR, WACV, ICIP, and CVPR. He is the program chair of WACV 2018. He is an Associate Editor of Neurocomputing journal. He has authored more than 100 scientific publications, and has filed 22 U.S. patents.

Dan Zeng receieved the B.Sc. degree in College of Computer Science from Sichuan University, China, in 2013. Since 2012, she paticipated in '3+2+3' successive graduate, postgraduate and doctoral program of Sichuan University. Currently, she is studying in the SCS group, uTwente, Netherlands as a visiting PhD student. Her research area is biometrics, especially low resolution as well as pose problems in face recognition.