

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

Feng Liu, Member, IEEE, 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 faces and cascaded regression in 2D and 3D shape spaces, iteratively and alternately applies two cascaded regressors, one for updating 2D landmarks and the other for 3D shape. The 3D 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 faces and localize both visible and invisible 2D landmarks. Based on the PEN 3D faces, we devise a method to enhance face recognition accuracy across poses and expressions. Both linear and nonlinear implementations of the proposed method are presented and evaluated in this paper. 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.

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 relationship between 2D landmarks and 3D faces. That is, the visibility and position of landmarks in the projected 2D image are determined by four factors: the 3D shape, the deformation due to expression and pose, and the camera

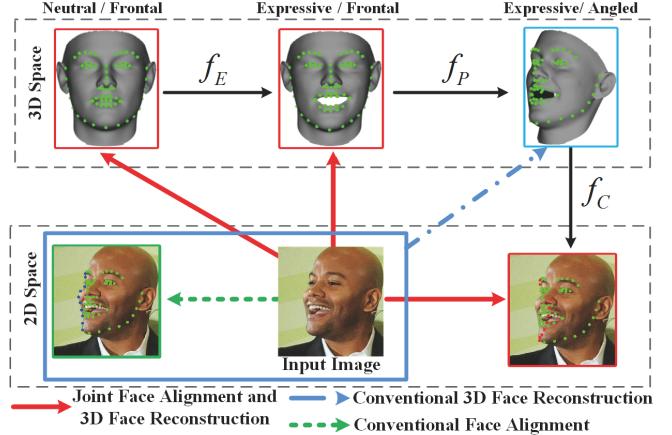


Fig. 1. We view 2D landmarks are generated from a 3D face through 3D expression (f_E) and pose (f_P) deformation, and camera projection (f_C). While conventional face alignment and 3D face reconstruction are two *separated* tasks and the latter requires the former as input, this paper performs these 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.

projection parameters. *Given such a clear correlation between 2D landmarks and 3D shape, it is evident that ideally they should be solved jointly, instead of separately as in prior works - indeed this is the core of this work.*

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

- 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.

Based on the texture features around landmarks on a face image, one set of regressors (called landmark regressors) gradually move the landmarks towards their true positions. By utilizing the facial landmarks as clues, the other set of regressors (called shape regressors) gradually improve the reconstructed 3D face. 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 shape adjustment via the shape regressors. The 3D-to-2D mapping is then computed based on the adjusted 3D shape and 2D landmarks, and it further refines the landmarks.

A preliminary version of this work was published in the 14th European Conference on Computer Vision (ECCV) 2016 [8]. We further extend the work from four aspects. (i) We explicitly reconstruct expression deformation of 3D faces, so that both PEN (pose and expression normalized) and expressive 3D faces can be reconstructed. (ii) We implement the proposed method in both linear and nonlinear regressions. (iii) We present in detail the application of the proposed method to face recognition. (iv) We carry out a more extensive evaluation with comparisons to state-of-the-art methods. In summary, this paper makes the following contributions.

- We present a novel cascaded coupled-regressor based method with linear and non-linear regressions 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 localize 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 faces 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 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 [12] and CFP [13] databases.

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 and two alternative implementations. 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, e.g., Active Shape Model (ASM) [14], [15] or Active Appearance Model (AAM) [16],

[17], [18], [19], search for landmarks based on global shape models and texture models. Constrained Local Model (CLM) [20] also utilizes global shape models to regularize the landmark locations, but it employs discriminative local texture models. Regression based methods [21], [22], [23], [24] have been recently proposed to directly estimate landmark locations by applying cascaded regressors to an input image. These methods mostly do not consider the visibility of 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 [25], [26] 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 cannot 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], [27], [28], [29], [30], [31] use 3D morphable models (3DMM) [32] — either a simplified one with a sparse set of landmarks [10], [28] or a relatively dense one [27]. They estimate the 3DMM parameters by using cascaded regressors with texture features as the input. In [28], the visibility of landmarks is explicitly computed, and the method can cope with face of yaw angles ranging from -90° to 90° , whereas the method in [27] does not work properly for faces of yaw angles beyond 60° . In [33], Tulyakov and Sebe propose to directly estimate the 3D landmark locations via texture-feature-based regressors for faces of yaw angles up to 50° .

These existing 3D model based methods regress 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 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 [34], [35] 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 could be weak to discriminate among different individuals. 3DMM [1], [32], [36], [37], [38] 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 image and the image 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. [39], Qu et al. [40] and Liu et al. [41] 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. [42] 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 [43]. 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 shown to be useful for face recognition.

2.3 Unconstrained Face Recognition

Face recognition has been developed rapidly in the past decade, especially since the emergence of deep learning techniques. Although automated methods [44], [45], [46] outperform humans in face recognition accuracy on the labelled faces in the wild (LFW) benchmark database, it is still very challenging to recognize faces in unconstrained images with large poses or intensive expressions [47], [48]. Potential reasons for degraded accuracy on off-angle and expressive faces include (i) off-angle faces usually have less discriminative texture information for identification than frontal ones, resulting in small inter-class differences, (ii) cross-view faces (e.g., frontal and profile faces) may have very limited features in common, leading to large intra-class differences, and (iii) pose and expression variations could cause substantial deformation to faces.

Existing methods recognize off-angle and expressive faces either by extracting invariant features or by normalizing out the pose or expression deformation. Yi et al. [49] fitted a 3D face mesh to an arbitrary-view face, and extracted pose-invariant features based on the 3D face mesh adaptively deformed to the input face. In DeepFace [50], the input face was first aligned to the frontal view with assistance of a generic 3D face model, and then recognized utilizing a deep network. Zhu et al. [3] proposed to generate frontal and neutral face images from the input images by using 3DMM [32] and deep convolutional neural networks. Very recently, generative adversarial networks (GAN) have been explored by Tran et al. [48], [51] for unconstrained face recognition. They devised a novel network, namely DR-GAN, which simultaneously synthesizes frontal faces and

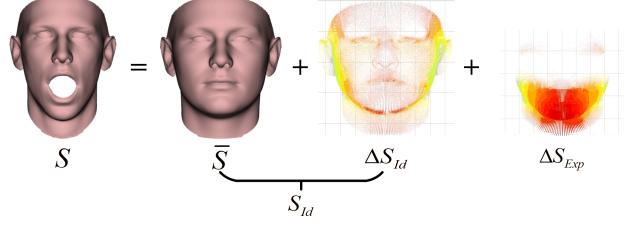


Fig. 2. A 3D face shape of a subject (S) is represented as summation of the mean pose-and-expression-normalized (PEN) 3D face shape (\bar{S}), the difference between the subject's PEN 3D shape and the mean PEN 3D shape (ΔS_{Id}), and the expression deformation (ΔS_{Exp}).

learn pose-invariant feature representations. Hu et al. [52] proposed to directly transform a non-frontal face into frontal face 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 faces and utilize merely 2D features for recognition. In this paper, on the contrary, we generate pose and expression normalized 3D faces from the input 2D images, and use these resultant 3D faces to improve the unconstrained face recognition accuracy.

3 PROPOSED METHOD

In this section, we introduce the proposed joint face alignment and 3D face reconstruction method and its implementations in detail. We start by defining the 3D face model with separable identity and expression components, and based on this model formulate the problem of interest. We then provide the overall procedure of the proposed method. Afterwards, the preparation of training data is presented, followed by the 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. Finally, a deep learning based nonlinear implementation of the proposed method is given.

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 shape, ΔS_{Id} is the difference between the subject's PEN 3D shape (denoted as S_{Id}) and \bar{S} , and ΔS_{Exp} is the expression-induced deformation in S w.r.t. S_{Id} (Fig. 2).

We use S_L to denote a subset of S with columns corresponding to l landmarks. The projections of these

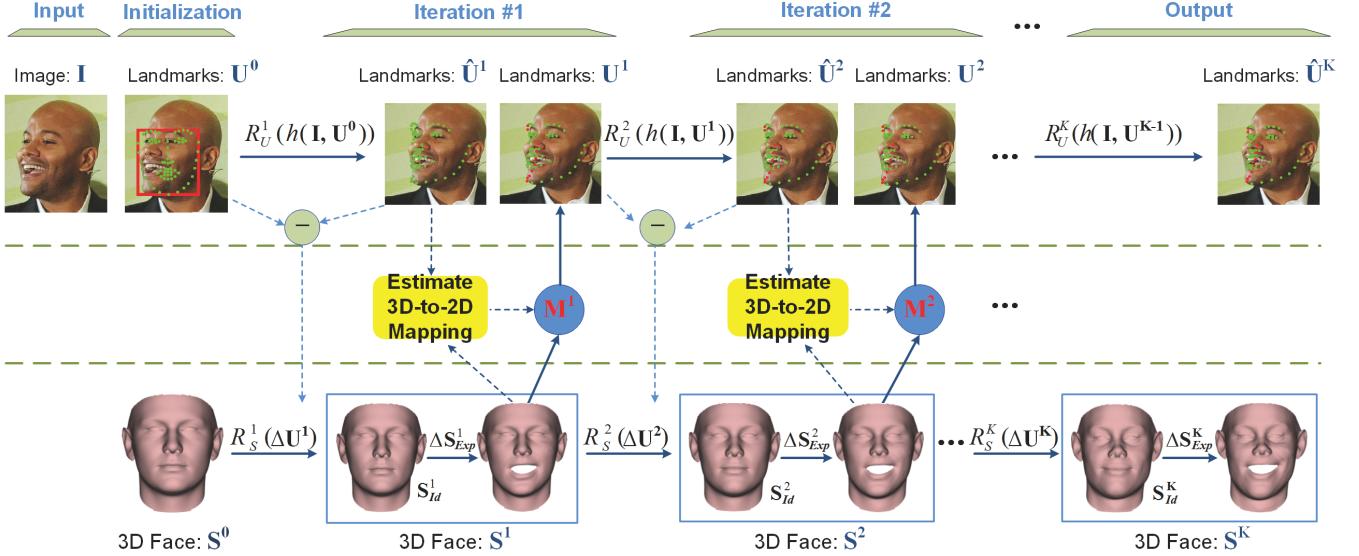


Fig. 3. Flowchart of the proposed joint face alignment and 3D face reconstruction method.

landmarks onto an image \mathbf{I} of the subject with arbitrary view are represented by

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

where f_C and f_P are, respectively, camera projection and pose-induced deformation. In this work, 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 a face image \mathbf{I} , our goal is to simultaneously estimate its landmarks \mathbf{U} , PEN 3D shape S_{Id} , and expression deformation ΔS_{Exp} . Note that, in some context, we also write the 3D shape and landmarks as column vectors: $\mathbf{S} = (x_1, y_1, z_1, \dots, x_n, y_n, z_n)^T$, and $\mathbf{U} = (u_1, v_1, \dots, u_l, v_l)^T$, where ' T ' is transpose operator.

3.2 The Overall Procedure

Figure 3 shows the flowchart of the proposed method. Given an image \mathbf{I} , its 3D 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 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 image. This is similar to the conventional cascaded regressor based 2D face alignment [21]. 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,

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

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 in linear case will be introduced in Sec. 3.4.

Updating 3D face shape In this step, the aforementioned landmark location adjustment is used to estimate the adjustment of the 3D 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.,

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

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 in linear case 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 with the assistance of the 3D-to-2D mapping matrix. We estimate \mathbf{M}^k based on \mathbf{S}^k and $\hat{\mathbf{U}}^k$. The refined landmarks \mathbf{U}^k can be obtained by projecting \mathbf{S}^k onto the image via \mathbf{M}^k according to Eq. (3). In this process, the landmark visibility is also recomputed. Details of this step will be given in Sec. 3.6.

3.3 Training Data Preparation

Before we provide details of the three steps, we first introduce the training data needed for learning the landmark and shape regressors, which will also facilitate the understanding of our algorithms. Since the purpose of these regressors is to gradually adjust the estimated landmark and shape towards their ground truth, we need a sufficient number of triplet data $\{(\mathbf{I}_i, \mathbf{S}_i^*, \mathbf{U}_i^*)\}_{i=1}^N$, 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 shapes have established dense correspondences among their vertices; i.e., they have the same number of vertices, and vertices of the same index in the 3D shapes have the same semantic meaning. Here, each of the ground

truth 3D shapes includes two parts, the PEN 3D 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 (e.g., the red points in Fig. 1).

Obviously, to enable regressors to cope with expression and pose variations, the training data should contain faces of these variations. It is, however, difficult to find in the public domain such data sets of 3D faces and corresponding annotated 2D images with various expressions/poses. Thus, we construct two training sets by ourselves: one based on BU3DFE [5], and the other based on 300W-LP [10], [53].

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 intensity levels. These 3D 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 [54] to establish dense correspondence of the 3D scans of 5,996 vertices. With the registered 3D scans, we compute the mean PEN 3D face shape by averaging all the subjects' PEN 3D shapes, which are defined by their 3D scans of frontal pose and neutral expression. All the images of one subject share the same PEN 3D shape of that subject, while their expression shapes can be obtained by first subtracting from their corresponding 3D scans, their PEN 3D face shape, and then adding the mean PEN 3D shape.

300W-LP database [10] is created based on 300W [53] database, which integrates multiple face alignment benchmark datasets (i.e., AFW [25], LFPW [55], HELEN [56], IBUG [53] and XM2VTS [57]). It includes 122,450 in-the-wild images of a wide variety of poses and expressions. For each image, its corresponding registered PEN 3D shape and expression shape are estimated by using the method in [3] based on BFM [58] and FaceWarehouse [59]. The obtained 3D faces have 53,215 vertices. Figure 4 and 5 shows example images and corresponding PEN 3D shapes and expression shapes in our training sets.

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 ground truth, which, along with landmark visibility, are given by \mathbf{U}^* in training. Therefore, the objective of landmark regressors R_U^k is to better predict the difference between \mathbf{U}^{k-1} and \mathbf{U}^* . In this section, we first implement the proposed method in a linear manner, by optimizing:

$$R_U^k = \arg \min_{R_U^k} \sum_{i=1}^N \| (\mathbf{U}_i^* - \mathbf{U}_i^{k-1}) - R_U^k(h(\mathbf{I}_i, \mathbf{U}_i^{k-1})) \|_2^2, \quad (6)$$



Fig. 4. Example images with annotated landmarks (1st, 4th rows), their 3D faces (2nd, 5th rows) and expression shapes (3rd, 6th 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 shapes corresponding to the neutral expression are their PEN 3D face shapes, which are highlighted in blue boxes.



Fig. 5. Four subjects in 300W-LP. From left to right: images with annotated landmarks, PEN 3D face shapes, and expression shapes.

which has a closed-form least-square solution. Note that, as we will show later, other nonlinear regression schemes, such as CNN [29], can also be adopted in our framework.

We use 128-dim SIFT descriptors [60] 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. Note that the regressors estimate the semantic locations of all landmarks including invisible ones.

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:

$$R_S^k = \arg \min_{R_S^k} \sum_{i=1}^N \| (\mathbf{S}_i^* - \mathbf{S}_i^{k-1}) - R_S^k (\Delta \mathbf{U}_i^k) \|_2^2, \quad (7)$$

with its closed-form solution as

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

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

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

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

In order to refine landmarks with the updated 3D shape, we 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 in Sec. 3.1, the mapping matrix is a composite effect of pose-induced deformation and camera projection. By assuming a weak perspective camera projection as in prior work [28], [61], the mapping matrix \mathbf{M}^k is represented by a 2×4 matrix, and can be estimated as a least-square solution to the following fitting problem:

$$\mathbf{M}^k = \arg \min_{\mathbf{M}^k} \| \hat{\mathbf{U}}^k - \mathbf{M}^k \mathbf{S}_L^k \|_2^2. \quad (9)$$

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

The visibility of the landmarks can be then computed based on the mapping matrix \mathbf{M} using the method in [28]. Suppose the average surface normal around a landmark in the 3D face shape \mathbf{S} is $\vec{\mathbf{n}}$. Its visibility v is measured by

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

where $\text{sgn}()$ 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 top two rows of \mathbf{M} . This rotates the surface normal and validates if it points toward the camera.

Algorithm 1 summarizes the process of learning the cascaded coupled linear regressors. Next, we introduce an alternative implementation of our proposed method by using nonlinear regressors, i.e., neural networks.

3.7 Nonlinear Regressors

In the above linear implementation, linear regressors with hand-crafted features are used. Here, we provide a nonlinear implementation, in which landmark and 3D shape regressors are implemented by deep convolutional neural networks (DCNN) and multiple layer perceptions (MLP), respectively. Figure 6 shows its pipeline.

Given a face image, as in linear implementation, its landmarks and 3D shape are initialized as the average

Algorithm 1 Learning Cascaded Coupled Linear Regressors

Input: Training data $\{(\mathbf{I}_i, \mathbf{S}_i^*, \mathbf{U}_i^*) | i = 1, 2, \dots, N\}$, 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** $k = 1, \dots, K$ **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**
-

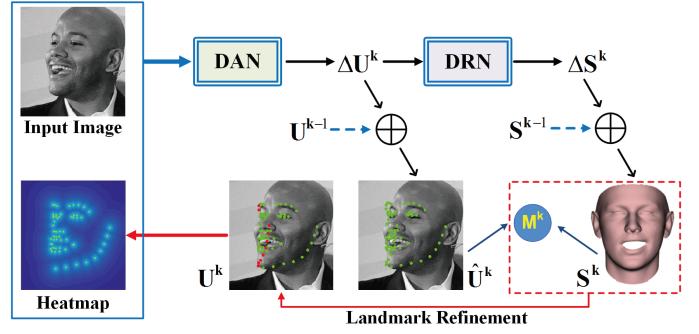


Fig. 6. Diagram of the proposed method implemented with nonlinear regressors. Deep Alignment Network (DAN) denotes the DCNN-based landmark regressors and Deep Reconstruction Network (DRN) denotes the MLP-based 3D shape regressors. Note that the landmark heatmap is not used at the initial stage.

landmarks and the average 3D shape. In every iteration, a landmark heatmap \mathbf{H} , which has the same dimension as the input image, is generated from the current estimated landmarks. The value of pixel (p_u, p_v) in the heatmap is set as the accumulated contributions of the visible landmarks, and the contribution of a landmark U_j is determined by

$$\mathbf{H}(p_u, p_v) = 1 / (1 + \min_{U_j \in U} \| (p_u, p_v) - U_j \|). \quad (11)$$

The heatmap and face image are stacked together as input to the DCNN-based landmark regressor. In this paper, we employ the structure of Deep Alignment Network (DAN) [62], and adapt its output layer so that landmark adjustment is estimated. The obtained landmark adjustment is then fed into the MLP-based 3D shape regressor (Deep Reconstruction Network, or DRN). DRN, consisting of a full-connection layer and a $\tanh()$ activation function, computes the 3D shape adjustment. After updating the 3D shape with the shape adjustment, we further refine the landmarks as in Sec. 3.6.

The DCNN- and MLP-based regressors are learned iteratively. We first train the regressors in prior iteration until convergence, and then move on to the next iteration. We employ the Euclidean loss in training both regressors.

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	$\pm 90^\circ$	$\pm 80^\circ$	$\pm 70^\circ$	$\pm 60^\circ$	$\pm 50^\circ$	$\pm 40^\circ$	$\pm 30^\circ$	$\pm 20^\circ$	$\pm 10^\circ$	0°	Avg.
Zhu et al. [3]	-	-	-	-	-	2.73	2.74	2.56	2.32	2.22	2.51
Tran et al. [42]	-	-	-	-	-	2.26	2.19	2.16	2.08	2.06	2.15
Liu et al. [41]	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 (Linear)	1.85	1.83	1.83	1.83	1.86	1.89	1.90	1.91	1.90	1.91	1.87
Proposed (Nonlinear)	1.92	1.91	1.93	1.92	1.92	1.91	1.92	1.92	1.93	1.93	1.92

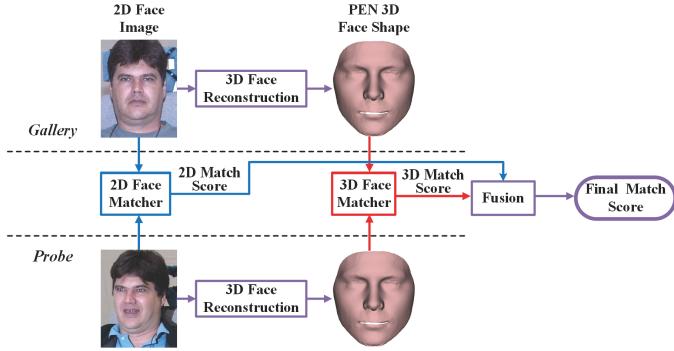


Fig. 7. Block diagram of the proposed 3D-enhanced face recognition.

4 APPLICATION TO FACE RECOGNITION

In this section we apply the reconstructed 3D faces to improve face recognition accuracy on off-angle and expressive faces. The basic idea is to utilize the additional feature provided by the reconstructed PEN 3D faces and fuse it with conventional 2D face matchers. Figure 7 shows the proposed 3D-enhanced face recognition method. As can be seen, 3D face reconstruction methods are applied to both gallery and probe faces to generate PEN 3D faces. The iterative closest point (ICP) algorithm [63] is applied to match the reconstructed normalized 3D face shapes. It aligns the 3D shapes reconstructed from probe and gallery images, and computes their distances, 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 conventional 2D face matcher (which are within $[0, 1]$ also) by 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 matcher and the sum fusion rule for simplicity. Other more elaborated 3D face matchers and fusion rules can also be applied with our proposed method. Thanks to the additional discriminative feature in PEN 3D face shapes 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 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 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 and 1,330 testing samples. We compare the proposed method with its preliminary version in [8] and three state-of-the-art methods in [3], [41], [42]. The methods in [8], [42] reconstruct PEN 3D faces only, while the methods in [3], [41] reconstruct 3D faces that have the same pose and expression as the input images. Moreover, the method in [41] 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 faces for [41]. For the methods of [3], [42], 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 [64]. As a result, they cannot be applied to faces of large poses (i.e., beyond 40 degrees).

We use two metrics to evaluate the 3D face reconstruction accuracy: Mean Absolute Error (MAE) and Normalized Per-vertex Depth Error (NPDE). MAE is defined as [65]:

$$\text{MAE} = \frac{1}{N_T} \sum_{i=1}^{N_T} (\|\mathbf{S}_i^* - \hat{\mathbf{S}}_i\|/n), \quad (12)$$

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 [34]:

$$\text{NPDE}(x_j, y_j) = (|z_j^* - \hat{z}_j|) / (z_{\max}^* - z_{\min}^*), \quad (13)$$

where z_{\max}^* and z_{\min}^* are the maximum and minimum depth values in the ground truth 3D face of testing samples, and z_j^* and \hat{z}_j are the ground truth and reconstructed depth values at the j^{th} vertex. We first report the results of our linear implementation, and then those of the nonlinear one. Note that when we mention the proposed method, the linear implementation is referred unless specified.

Reconstruction accuracy across poses Table 1 shows the average MAE of the proposed method under different poses of the input faces. 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], [41] and our proposed method already have the same number of vertices as the ground truth, we employ Procrustes alignment for these methods as suggested by [66]. For the results of [3], [42], however, the

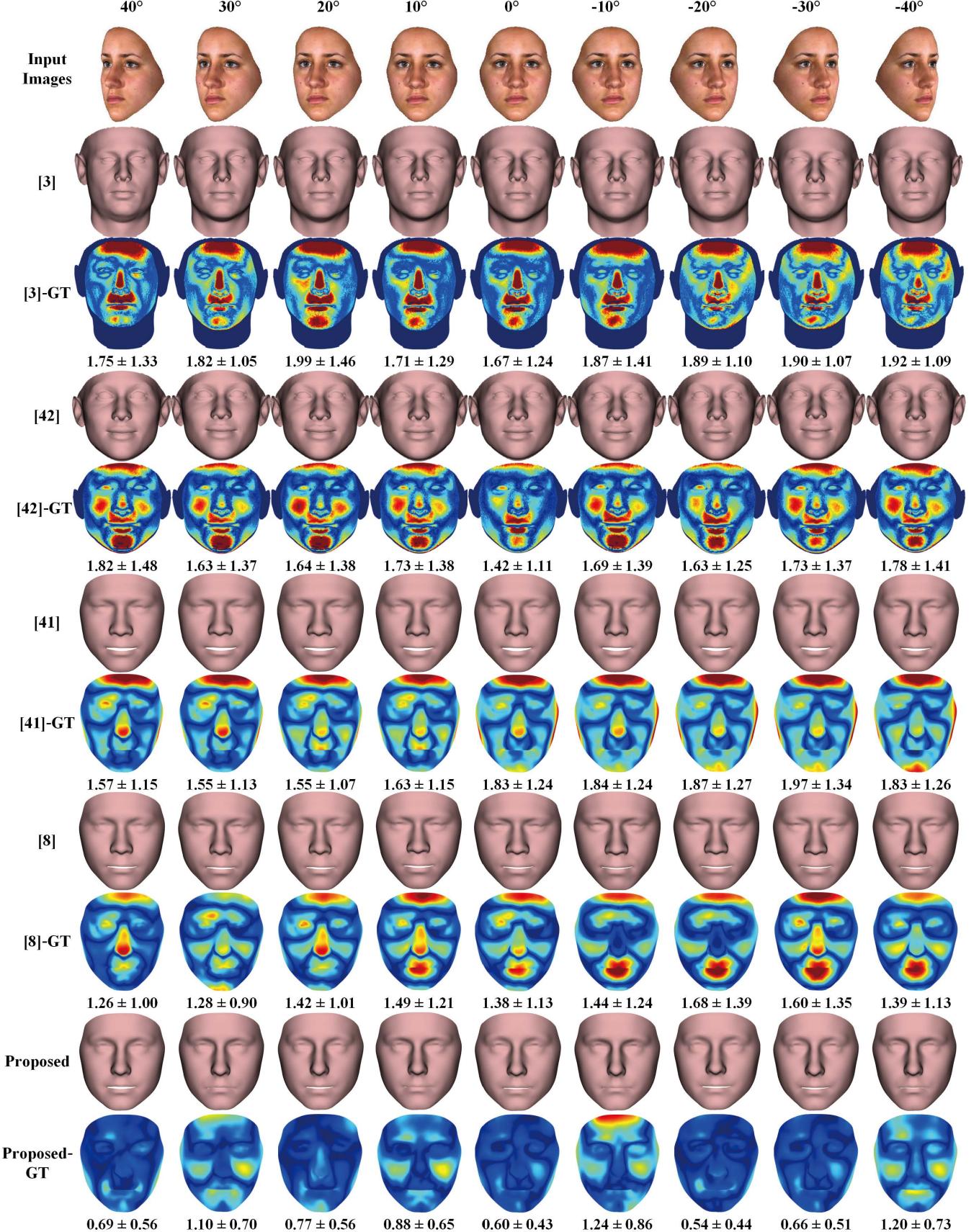


Fig. 8. Reconstruction results for a BU3DFE subject at nine poses. The even rows show the reconstructed 3D faces by [3], [42], [41], [8] and the proposed method. Except the first row, the odd rows show their corresponding NPDE maps. The colormap goes from dark blue to dark red (corresponding to errors between 0 and 5). The numbers under each error map represent the mean and standard deviation (in %).

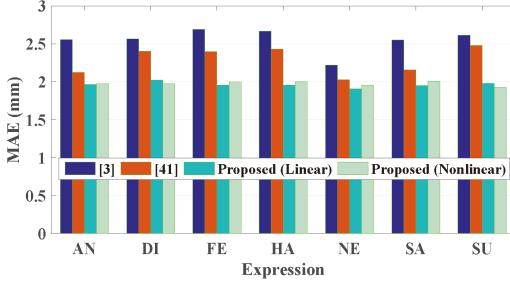


Fig. 9. 3D face reconstruction accuracy (MAE) of the proposed method, [41] and [3] under different expressions: angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA) and surprise (SU).

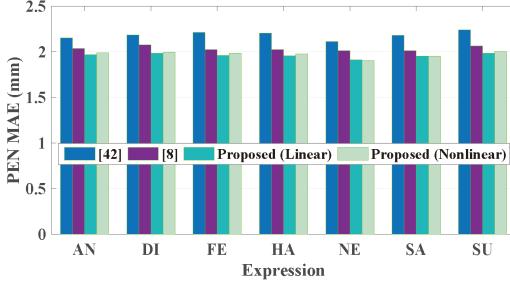


Fig. 10. PEN 3D face reconstruction accuracy (MAE) of the proposed method, [8] and [42] under different expressions.

number of vertices is different from the ground truth. Hence, we align them using rigid ICP method as [42] does. It can be seen from Table 1 that the average MAE of the proposed method (either linear or nonlinear implementation) is lower than that of counterpart methods. Moreover, as the pose becomes large, the error of the proposed method does not increase substantially. This proves the effectiveness of the proposed method in handling arbitrary view faces. Figure 8 shows the reconstruction results of one subject.

Reconstruction accuracy across expressions Figure 9 shows the average MAE of the proposed method and [3], [41] across expressions, based on their reconstructed 3D faces that have the same pose and expression as the input. The proposed method overwhelms its counterpart for all expressions. Moreover, as expressions change, the MAE standard deviation of [3], [41] are 0.157mm and 0.179mm , whereas that of the proposed method is 0.034mm in linear implementation and 0.029mm in nonlinear implementation. This proves the superior robustness of the proposed method to expression variations.

Figure 10 compares the average MAE of the proposed method and [8], [42] across expressions, based on their reconstructed PEN 3D faces. Again, the proposed method shows superiority in both MAE under all expressions and robustness across expressions. We believe that such superiority is owing to its explicit modeling of expression deformation. Figure 11 shows the reconstruction results for one subject under seven expressions.

Reconstruction accuracy across races It is well known that people from different races (e.g., Asian and Caucasian) show different characteristics in facial shapes. Such other-race effect has been reported in face recognition literature [67]. In this experiment, we study the impact of races on

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

	Asian	African	Hispanic	Caucasian	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%)

3D face reconstruction using the FRGC v2.0 database [11]. FRGC v2.0 contains 3D faces and images of 466 subjects with different ethnic groups (Table 2). Since these faces have no expression variation, the expression shape component in our proposed model is set to zero. We use the method in [54] to establish dense correspondence of the 3D faces of 5,996 vertices. We conduct three experiments: (i) training with 100 Asian samples (denoted as **Setting I**), (ii) training with 100 Caucasian samples (**Setting II**), and (iii) training with 100 Asian and 100 Caucasian samples (**Setting III**). The testing set contains samples of remaining subjects in FRGC v2.0, including 12 Asian, 6 African, 13 Hispanic, 19 Caucasian and 16 Unknown races.

Figure 12 compares the 3D face reconstruction accuracy (MAE) across different ethnic groups. Not surprisingly, training for one ethnic group can yield higher accuracy on testing of the same ethnic. As for the other-race effect, the model trained on Caucasian achieves comparable accuracy on Caucasian and Hispanic, but much worse on the other races (and worst on Asian). On the other hand, the model trained on Asian performs much worse on all other races compared to on its own race, and the worst on African. These results reveal the variations in the facial shapes of people from different races. Further, by combining training data of Asian and Caucasian (Setting III), comparable reconstruction accuracy is achieved for both Asian and Caucasian, which is also comparable to those in Setting I and II. This proves the capability of the proposed method in handling the 3D shape variations among *all* ethnic groups.

5.2 Face Alignment Accuracy

In evaluating face alignment, several state-of-the-art face alignment methods are considered for comparison to the proposed method, including RCPR [68], ESR [22], SDM [21], 3DDFA and 3DDFA+SDM [10]. The dataset constructed from 300W-LP is used for training, the AFLW [9] and AFLW2000-3D [10] are for testing. AFLW contains 25,993 in-the-wild faces with large poses (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 angle of the testing image: $[0^\circ, 30^\circ]$, $[30^\circ, 60^\circ]$ 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 in $[30^\circ, 60^\circ]$ and 232 in $[60^\circ, 90^\circ]$. The bounding boxes provided by AFLW are used in the AFLW

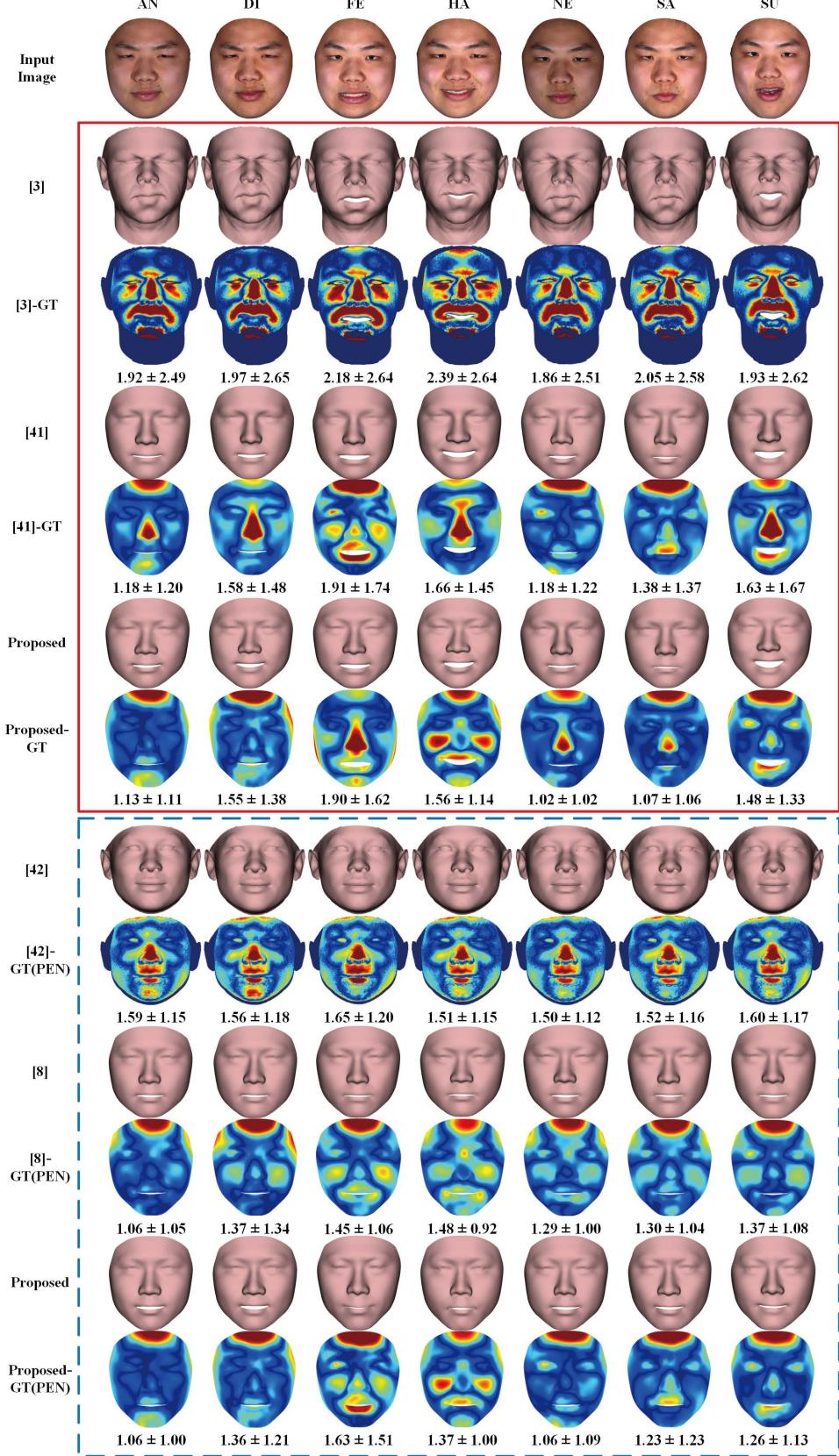


Fig. 11. Reconstruction results for a BU3DFE subject in seven expressions. The first row shows the input images. The red box shows the reconstructed 3D faces with the same expression as the input, using [41], [3] and the proposed method. The blue box shows the reconstructed PEN 3D faces by [8], [42] and the proposed method.

TABLE 3

The face alignment accuracy (NME) of the proposed method and 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 [68]	5.43	6.58	11.53	7.85	3.24	4.26	5.96	13.18	7.80	4.74
ESR [22]	5.66	7.12	11.94	8.24	3.29	4.60	6.70	12.67	7.99	4.19
SDM [21]	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 (Linear)	3.75	4.33	5.39	4.49	0.83	3.25	3.95	6.42	4.61	1.78
Proposed (Nonlinear)	3.22	4.13	5.13	4.16	0.96	2.72	4.06	5.81	4.20	1.55

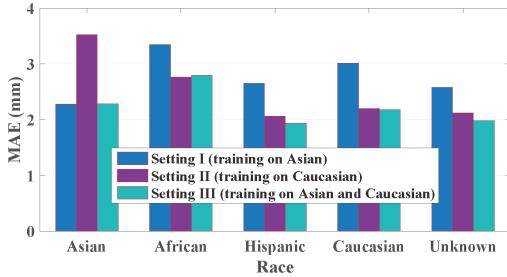


Fig. 12. 3D face reconstruction accuracy (MAE) of the proposed method across different ethnic groups.

testing, while the ground truth bounding boxes enclosing all 68 landmarks are used for the AFLW2000-3D testing.

Normalized Mean Error (NME) [28] 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:

$$\text{NME} = \frac{1}{N_T} \sum_{i=1}^{N_T} \left(\frac{1}{d_i} \frac{1}{N_i^v} \sum_{j=1}^l \mathbf{v}_{ij} \|(\hat{u}_{ij}, \hat{v}_{ij}) - (u_{ij}^*, v_{ij}^*)\| \right), \quad (14)$$

where d_i is the square root of the bounding box area of the i^{th} testing sample, N_i^v is the number of its visible landmarks, (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 compares the face alignment accuracy on the AFLW and AFLW2000-3D datasets. As can be seen, the proposed method achieves the best accuracy for all poses and on both datasets. In order to assess the robustness of different methods to pose variations, we also report their standard deviations of the NME in Table 3. The results again demonstrate the superiority of the proposed method over the counterpart. Figure 13 shows the landmarks detected by the proposed method on some AFLW images.

Moreover, for the proposed method, the nonlinear regression implementation is better than the linear one. CNN feature is more powerful and robust than the handcrafted SIFT feature for the face alignment task. In contrast, in the experiments of 3D face reconstruction on BU3DFE database (see Section 5.1), the reconstruction error of linear regressors is lower than that of nonlinear regressors. This might be because MLP-based nonlinear regressors for 3D face reconstruction need more training samples.

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”) our 3D enhancement.

Method	±90°	±75°	±60°	±45°	±30°	±15°	Avg.
VGG-2D	36.2	66.9	83.5	93.8	97.7	98.6	79.5
LightenedCNN-2D	7.50	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
ICP-3D	31.8	30.6	34.3	32.8	34.7	44.3	33.0
VGG-Fusion	52.6	75.2	90.5	96.8	98.5	99.4	85.5
LightenedCNN-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

5.3 Face Recognition

While there are many recent face alignment and reconstruction works [69], [70], [71], [72], [73], few works take one step further to evaluate the contribution of alignment or reconstruction to subsequent tasks, such as face recognition. In contrast, we quantitatively evaluate the contribution of the reconstructed pose-expression-normalized (PEN) 3D faces to face recognition by directly matching 3D to 3D shape and fusing it with conventional 2D face recognition. Refer to Sec. 4 for details of the PEN 3D faces enhanced face recognition method.

In this evaluation, we employ the linear implementation, and use the BU3DFE (13,300 images of 100 subjects; refer to Sec. 3.3) and MICC [74] databases as training data, the CMU Multi-PIE database [12] and the Celebrities in Frontal-Profile (CFP) database [13] as test data. MICC contains 3D face scans and video clips (indoor, outdoor and cooperative head rotations environments) of 53 subjects. We randomly select faces 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 [54] to establish dense correspondence with $n = 5,996$ vertices.

5.3.1 Face Identification on Multi-PIE Database

CMU Multi-PIE is a widely used benchmark database for face recognition, with faces 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], [75], probe images consist of the images of all 337



Fig. 13. The 68 landmarks detected by the proposed method for AFLW data. Green/red points denote visible/invisible landmarks.

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 faces.

Pose \ Expression	Smile	Surprise	Squint	Disgust	Scream	Avg.
$\pm 90^\circ$	51.4(36.9)	46.1(35.7)	58.8(38.7)	42.0(24.9)	63.6(52.4)	52.4(37.7)
$\pm 75^\circ$	73.1(67.0)	56.6(53.0)	72.6(67.8)	52.5(43.4)	75.1(71.6)	66.0(60.4)
$\pm 60^\circ$	88.6(89.8)	80.2(80.7)	91.6(88.2)	74.6(69.8)	91.8(92.7)	85.4(84.2)
$\pm 45^\circ$	95.9(97.6)	89.4(95.1)	95.6(97.8)	86.7(83.5)	97.3(98.7)	93.0(94.5)
$\pm 30^\circ$	97.8(99.1)	93.1(97.0)	96.8(99.3)	90.4(91.5)	98.5(99.8)	95.3(97.3)
$\pm 15^\circ$	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)

subjects at 12 poses ($\pm 90^\circ$, $\pm 75^\circ$, $\pm 60^\circ$, $\pm 45^\circ$, $\pm 30^\circ$, $\pm 15^\circ$) with neutral expression and frontal illumination. In the second experiment, instead of neutral expression, all images with smile, surprise, squint, disgust and scream expressions at the 12 poses and under frontal illumination are the probe images. This protocol is an extended version of [3], [4] by adding large-pose images ($\pm 60^\circ$, $\pm 75^\circ$, $\pm 90^\circ$). In both experiments, the frontal images captured in the first session are the gallery. And four state-of-the-art deep learning based (DL-based) face matchers are used as baselines, i.e., VGG [76], Lightened CNN [77], CenterLoss [78] and LDF-Net [52]. The first three matchers are publicly available. We evaluate them with all 337 subjects in Multi-PIE. The last matcher, LDF-Net, is a latest one 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 a higher weight (i.e., 0.7) to it, whereas the weights for all the other three baseline matchers are set to 0.5.

Table 4 reports the rank-1 accuracy of the baseline face matchers in the first experiment, where the baseline matchers are all further improved by our proposed method. Specifically, VGG and Lightened CNN are consistently improved across different poses when fused with 3D, while CenterLoss gains substantial improvement at large poses (15.5% at $\pm 90^\circ$ and 4.0% at $\pm 75^\circ$). Even for the latest LDF-Net, the recognition accuracy is improved by 5.1% at $\pm 90^\circ$ and 1.4% at $\pm 75^\circ$. For all the baseline matchers, the larger the yaw angle is, the more evident the accuracy improvement. Table 4 also gives the recognition accuracy of

using only the reconstructed 3D faces, at the row headed by "ICP-3D". Although its average accuracy is much worse compared with its 2D counterparts, it fluctuates more gently as probe faces rotate from frontal to profile. These results prove the effectiveness of the proposed method in dealing with pose variations, as well as in reconstructing individual 3D faces with discriminative details that are complementary to 2D face recognition.

Given its best performance among 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 those in Table 4. Yet, our proposed method still improves the overall accuracy of the baseline, especially for probe faces 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 the 3D face shape space.

5.3.2 Face Verification on CFP Database

We further evaluate our reconstructed PEN 3D faces on a more challenging unconstrained face recognition setting by using the CFP database, which has 500 subjects each with 10 frontal and 4 profile images. The evaluation includes frontal-frontal (FF) and frontal-profile (FP) face verification, each having 10 folders with 350 same-person and 350 different-person pairs. Table 6 reports the average results with standard deviations in terms of Accuracy, Equal Error Rate (EER), and Area Under the Curve (AUC).

TABLE 6

Verification accuracy on CFP by the CenterLoss face matchers before (indicated by suffix “2D”) and after (indicated by suffix “Fusion”) the enhancement by our proposed method.

Method	CenterLoss-2D	ICP-3D	CenterLoss-Fusion	
FF	Accuracy (%) EER (%) AUC (%)	86.43 ± 3.10 14.20 ± 3.58 93.38 ± 2.18	74.83 ± 3.85 27.65 ± 3.80 78.41 ± 4.11	89.21 ± 2.88 11.37 ± 2.94 94.24 ± 2.67
FP	Accuracy (%) EER (%) AUC (%)	69.27 ± 2.33 31.63 ± 2.36 74.61 ± 2.54	65.74 ± 2.47 36.26 ± 2.76 69.02 ± 3.69	72.99 ± 1.90 27.91 ± 2.06 78.64 ± 2.43

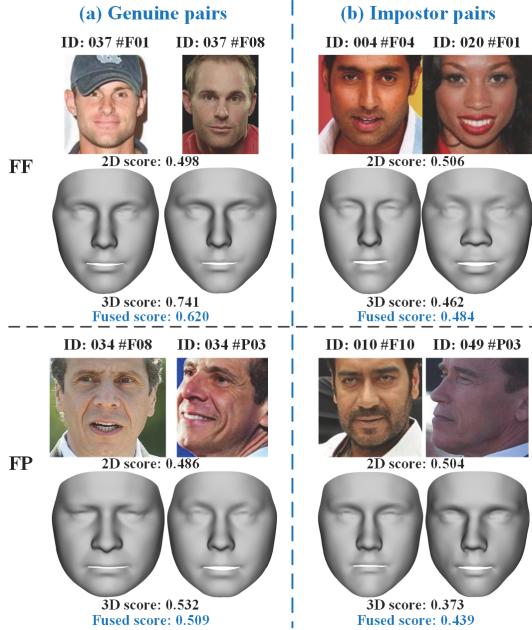


Fig. 14. Example (a) genuine pairs and (b) imposter pairs in CFP and corresponding PEN 3D faces, for which the CenterLoss method fails, whereas its fusion with our proposed method succeeds. Note that the operational threshold in our experiments is empirically set to 0.502.

Given its best performance on Multi-PIE database, we employ the CenterLoss matcher in this experiment. We also report the recognition accuracy of reconstructed PEN 3D faces (see “ICP-3D”). Although its average accuracy is much worse compared with the baseline, it further improves the performance of CenterLoss in both frontal-frontal (FF) and frontal-profile (FP) face verification. These results prove the effectiveness of the proposed method in dealing with pose variations, as well as the ability in providing complementary discriminative features in unconstrained environment. Figure 14 shows some example genuine and imposter pairs in CFP, which are incorrectly recognized by CenterLoss, but correctly recognized by fusion of CenterLoss and our proposed method.

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 linear and nonlinear implemen-

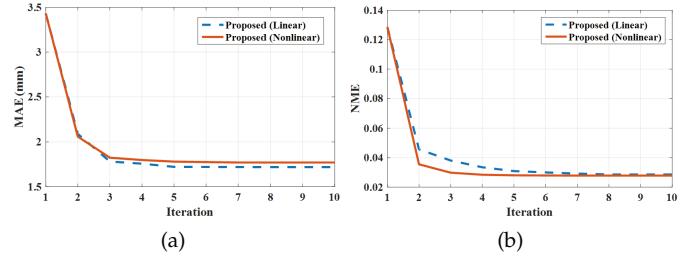


Fig. 15. (a) and (b) show the reconstruction errors (MAE) and alignment errors (NME) during the training of proposed method as iteration proceeds, when trained on the BU3DFE database.

TABLE 7
The time efficiency (in milliseconds or *ms*) of the proposed method.

Step	Updating landmarks	Updating shape	Refining landmarks	Total
Linear (<i>ms</i>)	14.93	15.38	8.57	38.88
Nonlinear (<i>ms</i>)	10.22	0.04	9.28	19.32

TABLE 8

Efficiency comparison of different reconstruction methods. For the methods of [3], [41], [42], although stand-alone landmark detection is required, it is not included in the reported times.

Method	[3]	[42]	[41]	[8]	Proposed (Linear)	Proposed (Nonlinear)
Time (<i>ms</i>)	56.3	88.0	12.6	32.8	38.9	19.3

tations on the BU3DFE database. We conduct ten-fold cross-validation experiments, and compute the average errors over the training data through ten iterations. As shown in Fig. 15, the training errors converge in about five iterations in the linear implementation, while in the nonlinear implementation the training errors converge fast after two to three iterations. Hence, we set the number of iterations as $K = 5$ and $K = 3$ in the linear and nonlinear implementations, respectively.

5.5 Computational Complexity

According to our experiments on a PC with i7-4790 CPU and 32 GB memory, the linear implementation of the proposed method runs at ~ 26 FPS, and the nonlinear implementation runs at ~ 52 FPS with a NVIDIA GeForce GTX 1080. This indicates that the proposed method can detect landmarks and reconstruct 3D faces in *real-time*. We also report the efficiency of individual steps in Table 7, and comparison with existing methods in Table 8.

6 CONCLUSION

In this paper, we present a novel regression based method for joint face alignment and 3D face reconstruction from single 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 faces 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 comparisons to state-of-the-art methods demonstrate the effectiveness and superiority of the proposed method in both face alignment and 3D face reconstruction, and in facilitating cross-view and cross-expression face recognition as well.

ACKNOWLEDGMENTS

The authors would like to thank the authors of LDF-Net for sharing us with the match scores of LDF-Net on Multi-PIE. This work is supported by the National Key Research and Development Program of China (2017YFB0802300), the National Natural Science Foundation of China (61773270), and the National Key Scientific Instrument and Equipment Development Projects of China (2013YQ49087904).

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*, 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] S. Sengupta, J. C. Chen, C. Castillo, and V. M. Patel, "Frontal to profile face verification in the wild," in *WACV*, 2016, pp. 1–9.
- [14] 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.
- [15] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models," in *BMVC*, 2007, pp. 1–10.
- [16] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *TPAMI*, no. 6, pp. 681–685, 2001.
- [17] I. Matthews and S. Baker, "Active appearance models revisited," *IJCV*, vol. 60, no. 2, pp. 135–164, 2004.
- [18] X. Liu, P. Tu, and F. Wheeler, "Face model fitting on low resolution images," in *BMVC*, 2006, pp. 1079–1088.
- [19] X. Liu, "Discriminative face alignment," *TPAMI*, vol. 31, no. 11, pp. 1941–1954, 2009.
- [20] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.
- [21] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.
- [22] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.
- [23] 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.
- [24] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *CVPR*, 2015, pp. 4998–5006.
- [25] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.
- [26] 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.
- [27] 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.
- [28] A. Jourabloo and X. Liu, "Pose-invariant 3D face alignment," in *ICCV*, 2015, pp. 3694–3702.
- [29] ———, "Large-pose face alignment via CNN-based dense 3D model fitting," in *CVPR*, 2016, pp. 4188–4196.
- [30] ———, "Pose-invariant face alignment via CNN-based dense 3D model fitting," *IJCV*, vol. 124, no. 2, pp. 187–203, 2017.
- [31] A. Jourabloo, X. Liu, M. Ye, and L. Ren, "Pose-invariant face alignment with a single CNN," in *ICCV*, 2017, pp. 3219–3228.
- [32] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194.
- [33] S. Tulyakov and N. Sebe, "Regressing a 3D face shape from a single image," in *ICCV*, 2015, pp. 3748–3755.
- [34] 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.
- [35] J. Roth, Y. Tong, and X. Liu, "Adaptive 3D face reconstruction from unconstrained photo collections," *TPAMI*, vol. 39, no. 11, pp. 2127–2141, 2017.
- [36] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *CVPR*, 2005, pp. 986–993.
- [37] G. Hu, F. Yan, J. Kittler, W. Christmas, C. H. Chan, Z. Feng, and P. Huber, "Efficient 3D morphable face model fitting," *Pattern Recognition*, vol. 67, pp. 366–379, 2017.
- [38] L. Tran and X. Liu, "Nonlinear 3D face morphable model," in *CVPR*, 2018, pp. 7346–7355.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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, pp. 1493–1502.
- [43] 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.
- [44] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.
- [45] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," *arXiv:1502.00873*, 2015.
- [46] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" *arXiv:1501.04690*, 2015.
- [47] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *TPAMI*, vol. 38, no. 8, pp. 1548–1568, 2016.
- [48] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *CVPR*, 2017, pp. 1283–1292.
- [49] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *CVPR*, 2013, pp. 3539–3545.

- [50] 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.
- [51] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *TPAMI*, 2018.
- [52] 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*, 2017, pp. 9–16.
- [53] 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.
- [54] T. Bolkart and S. Wuhrer, "3D faces in motion: Fully automatic registration and statistical analysis," *CVIU*, vol. 131, pp. 100–115, 2015.
- [55] 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.
- [56] 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.
- [57] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *AVBPA*, vol. 964, 1999, pp. 965–966.
- [58] 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.
- [59] 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.
- [60] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
- [61] 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.
- [62] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *CVPRW*, 2017, pp. 2034–2043.
- [63] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," *IVC*, pp. 2724–2729, 1991.
- [64] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.
- [65] 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.
- [66] 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.
- [67] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Trans. on Applied Perception (TAP)*, vol. 8, no. 2, p. 14, 2011.
- [68] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *ICCV*, 2013, pp. 1513–1520.
- [69] 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*, 2017, pp. 2614–2623.
- [70] O. Tuzel, T. K. Marks, and S. Tambe, "Robust face alignment using a mixture of invariant experts," in *ECCV*, 2016, pp. 825–841.
- [71] 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.
- [72] 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.
- [73] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," pp. 5553–5562, 2017.
- [74] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2D/3D hybrid face dataset," in *ACM J-HGBU*, 2011, pp. 79–80.
- [75] 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.
- [76] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015, pp. 41.1–41.12.
- [77] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *TIFS*, vol. 13, no. 11, pp. 2884–2896, 2018.
- [78] 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 is currently a post-doc researcher in the Computer Vision lab at Michigan State University. He received the Ph.D. degree in Computer Science from Sichuan University in 2018. His main research interests are computer vision and pattern recognition, specifically for face modeling, 2D and 3D face recognition. He is a 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 perception, fingerprint recognition, and affective computing, with applications to forensics, intelligent video surveillance, mobile security, healthcare, and human-computer interactions. Dr. Zhao has published more than 60 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 CCBR 2016 and ISBA 2018, and an area co-chair of BTAS 2018.



Xiaoming Liu is an Associate 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. His research interests include computer vision, pattern recognition, biometrics and machine learning. He is the recipient of 2018 Withrow Distinguished Scholar Award from Michigan State University. 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 an Area Chair for numerous conferences, including FG, ICPR, WACV, ICIP, and CVPR. He is a Co-Program Chair of BTAS 2018 and WACV 2018 conferences. He is an Associate Editor of Neurocomputing journal. He is a guest editor for IJCV Special Issue on Deep Learning for Face Analysis, and ACM TOMM Special Issue on Face Analysis for Applications. He has authored more than 100 scientific publications, and has filed 26 U.S. patents.



Dan Zeng is currently a post-doc researcher in the DMB group, University of Twente. She received the B.Sc. and Ph.D. degrees from Sichuan University in 2013 and 2018. Since 2012, she participated in '3+2+3' successive graduate, postgraduate and doctoral program of Sichuan University. Her main research area is computer vision and biometrics, specifically for challenges in face recognition.