Beyond 3DMM: Learning to Capture High-fidelity 3D Face Shape

Xiangyu Zhu, *Member, IEEE,* Chang Yu, Di Huang, *Senior Member, IEEE,* Zhen Lei, *Senior Member, IEEE,* Hao Wang, and Stan Z. Li, *Fellow, IEEE*

Abstract—3D Morphable Model (3DMM) fitting has widely benefited face analysis due to its strong 3D priori. However, previous reconstructed 3D faces suffer from degraded visual verisimilitude due to the loss of fine-grained geometry, which is attributed to insufficient ground-truth 3D shapes, unreliable training strategies and limited representation power of 3DMM. To alleviate this issue, this paper proposes a complete solution to capture the personalized shape so that the reconstructed shape looks identical to the corresponding person. Specifically, given a 2D image as the input, we virtually render the image in several calibrated views to normalize pose variations while preserving the original image geometry. A many-to-one hourglass network serves as the encode-decoder to fuse multiview features and generate vertex displacements as the fine-grained geometry. Besides, the neural network is trained by directly optimizing the visual effect, where two 3D shapes are compared by measuring the similarity between the multiview images rendered from the shapes. Finally, we propose to generate the ground-truth 3D shapes by registering RGB-D images followed by pose and shape augmentation, providing sufficient data for network training. Experiments on several challenging protocols demonstrate the superior reconstruction accuracy of our proposal on the face shape.

Index Terms—3D face, face reconstruction, 3DMM, fine-grained, personalized, 3D face dataset.

1 Introduction

The core problem addressed in this paper is high-fidelity 3D face reconstruction from a single face image. As a detailed and interpretable description of face images, 3D faces have acted as an essential priori in many face-related tasks, e.g., face recognition [1], [2], face manipulation [3], [4], expression analysis [5], [6] and facial animation [7], [8]. Although there have been significant advances in 3D face reconstruction [9], [10], [11], [12], the limited performance still suffers from insufficient ground-truth 3D face data. There are three processes required in ideal 3D face collection: 1) Instantly capturing the multiview 3D scans of one face. 2) Fusing the scans to a full 3D face with annotated landmarks. 3) Acquiring the topology-uniform 3D mesh by registering the full 3D face to a template. The requirements of expensive devices, fully controlled environment and laborious human labeling limit the large-scale collection of 3D faces for neural network training.

Generally, there are two strategies to tackle the data challenge: one is the 2D-to-3D method and the other is the self-supervised method. The 2D-to-3D method constructs 3D faces by fitting a 3D Morphable Model (3DMM) to annotated landmarks [10]. The generated 3D faces overlap the face region well with pixel-level parsing accuracy. However, as a linear model built from insufficient data, 3DMM spans restricted demographic features. In previous decades, almost all the models [13], [14], [15] cover no more than $1,000$ subjects, which is far from adequate to approximate the large diversity of human faces. Besides, in 3DMM fitting, only facial features are coarsely constrained by the landmarks, and the other large regions, such as cheek, are less constrained by shape priori. Therefore, restricted by the coarse 3D ground truth, the trained neural networks [9], [10], [16] cannot capture personalized shape and suffer from model-like reconstruction results, as shown in Fig. 1.

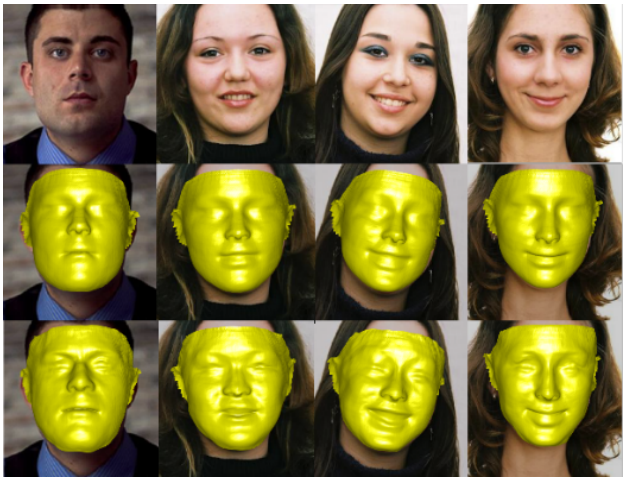


Fig. 1. The first, second and third rows show the images, the widely applied 3DMM fitting [9] results, and our results, respectively.

On the other hand, following the analysis-by-synthesis manner, the self-supervised method [11], [17], [18] learns a non-linear 3D face model from a large set of unlabeled images. Specifically, given a face image as input, a network encoder estimates the projection, lighting, shape, and albedo parameters, and two decoders serve as the non-linear

- *X. Zhu, C. Yu, Z. Lei, H. Wang and S. Li are with Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Donglu, Beijing 100190, China. Email: {xiangyu.zhu, chang.yu, zlei, szli}@nlpr.ia.ac.cn, haowang7308@gmail.com.*
- *D. Huang is with Beijing Advanced Innovation Center for BDBC, Beihang University, Beijing, China. USA. Email: dhuang@buaa.edu.cn.*

3DMM to map the parameters to a 3D face. Then, an analytically-differentiable rendering layer is employed to minimize the difference between the input face and the reconstructed face, and the learned decoder becomes the 3D face model. In this manner, the non-linear model can cover a larger shape space than the linear 3DMM. However, the improvement is still not satisfactory without straightforward supervision signals from ground-truth 3D faces. First, the successful reconstruction of the original face does not promise that the underlying 3D shape is the ground truth due to 2D-to-3D ambiguity. Second, the analysis-by-synthesis method mainly optimizes pixel values. Thus, the person-specific geometry, which does not contribute much to pixel values, is not well captured during optimization.

Beyond 2D-to-3D and self-supervised strategies, there is little research on how to enable neural networks to capture fine-trained geometry to improve visual verisimilitude. This paper explores this nontrivial problem and proposes that fully supervised learning is still a promising solution given only thousands of single-view RGB-D images. To this end, we extend the recent state-of-the-art methods and make the following contributions:

Data: We explore the construction of large-scale high-fidelity 3D face data with single-view RGB-D face images. Although it is expensive to acquire high-precision 3D faces, single-view RGB-D images can be considered an appropriate alternative because they are much easier to collect, especially considering the development of hand-held depth cameras (such as iPhone X), which strongly increases the possibility of the massive collection of medium-precision 3D data. Therefore, we register the RGB-D images and investigate two data augmentation strategies, view simulation and shape transformation, to generate a large 3D dataset **F**ine-**G**rained **3D** face (**FG3D**).

Network: A **F**ine-**G**rained reconstruction **Net**work (**FGNet**) is proposed in specific. With an initial 3DMM fitting result, we render the input image in 5 calibrated views and fuse the encoded features by a many-to-one hourglass network, where the mid-level features are aligned to a common UV space to ensure identical topology among the fused features. With lossless pose normalization and receptive field alignment, the network can concentrate on shape information and encode sophisticated features to capture the personalized shape.

Loss Function: Although commonly effective in 3D face alignment [9], [10], [19], the Mean Squared Error (MSE) loss is insensitive to fine-grained geometry, which contributes significantly to the visual effect but little to shape morphing. Considerable works [20], [21], [22], [23] also indicate that MSE cannot model how humans observe a 3D object. To address this issue, we propose a plaster sculpture descriptor to model the visual effect, where the reconstructed 3D face is rendered into several views with shading as the texture, leading to prominent improvements in visual verisimilitude.

A preliminary version of this work was published in [24]. We extend it in the following aspects: 1) For the network structure, we propose a brand-new Virtual Multiview Network (VMN) with a virtual multiview camera system and a many-to-one hourglass network, which remedies the defects of the camera view network and the model view network in [24]. 2) For the loss function, we improve the MSE loss by modeling the visual effect through a novel plaster sculpture descriptor. 3) For the data construction, we augment shape variations by transforming the underlying 3D shapes of face images and refining the face appearances accordingly, which further strengthens shape robustness. 4) Additional experiments are conducted to better analyze the motivation behind the design of the network structure and the loss function.

2 Related Work

There has been significant progress in 3D face reconstruction in recent decades. This section briefly reviews previous studies related to our proposal, including 3D morphable model fitting, non-linear 3D model construction, and fine-grained geometry reconstruction.

2.1 3D Morphable Model Fitting

In early years, some methods use CNN to directly estimate 3D Morphable Model parameters [10], [12], [25] or its variants [26], [27], [28], [29], [30], providing both dense face alignment and 3D face reconstruction results. However, the performances of these methods are restricted by the limitation of linear 3D space [27], [31], [32], [33], [34], [35]. It is also challenging to estimate the required face transformations, including perspective projection and 3D thin plate spline transformation. Recently, vertex regression networks [9], [16], which bypass the limitation of the linear model, have achieved state-of-the-art performance on their respective tasks. VRN [16] proposes to regress 3D faces stored in the volume, but the redundant volumetric representation loses the semantic meanings of 3D vertices. PRNet [9] designs a UV position map, a 2D image recording the 3D facial point cloud while maintaining the semantic meaning at each UV position. Cheng et al. [19] encode an image with a CNN and decode the 3D geometry with a lightweight Graph Convolutional Networks (GCN) for efficiency. Although breaking through the limitations of 3DMM, their reconstruction results are still model-like as their ground truth still comes from 3DMM fitting, e.g., 300W-LP [10].

2.2 Non-linear 3D Face Model by Self-supervision

Another way to bypass the limitation of 3DMM is to encode the 3D face model by a neural network, which can be learned on the unlabeled images in an analysis-by-synthesis manner. Tran et al. [11] achieve a certain breakthrough by learning a nonlinear model from unlabeled images in a self-supervised manner. They first estimate the projection, shape, and texture parameters with an encoder and then adopt two CNN decoders as the nonlinear 3DMM to map the parameters to a 3D textured face. With an analytically-differentiable rendering layer, the network can be trained in a self-supervised way by comparing the reconstructed images with the original images. Zhou et al. [18] follow a similar path but decode the shape and texture directly on the mesh with graph convolution for efficiency. Gao et al. [36] utilize a CNN encoder and a GCN decoder to learn 3D faces from unconstrained photos with photometric and

adversarial losses. In addition to learning a single model accounting for all shape variations, some works construct facial details independently. Chen et al. [37] un-warp image pixels and facial texture to the UV plane with a fitted 3DMM and predict a displacement map of the coarse shape by minimizing the difference between the reconstructed and the original images. Wang et al. [38] adopt a similar method but predict the moving rate in the direction of the vertex normal instead of the displacement map. The self-supervised nonlinear 3DMMs can cover a larger shape space than linear 3DMMs, but the fine-grained geometry, as a subtle constituting factor of face appearance, is easy to ignored due to the absence of straightforward training signals from the ground-truth shapes.

2.3 Fine-grained Geometry Reconstruction

Considering the lack of high-precision 3D scans, a common strategy for fine-grained reconstruction is **S**hape **f**rom **S**hading (**SfS**), which recovers 3D shapes from shading variations in 2D images. However, traditional SfS methods largely depend on the prior geometric knowledge and suffer from the ambiguity problem caused by albedo and lighting estimation. To address this issue, Richardson et al. [25] refine the depth map rendered by a fitted 3DMM with the SfS criterion to capture details, where a trained depth refinement net directly outputs high-quality depth maps without calculating albedo and lighting information. DF^{2}Net [39] further introduces depth supervision from RGB-D images to regularize SfS in a cascaded way. However, these methods only learn the 2.5D depth map and lack correspondence with the 3D face model. Jiang et al. [40] propose to refine the coarse shape with photometric consistency constraints to generate a more accurate initial shape. Li et al. [41] employ a masked albedo prior to improve albedo estimation and incorporate the ambient occlusion behavior caused by wrinkle crevices. Chen et al. [42] construct a PCA model for details using collected scans and adopt the analysis-by-synthesis strategy to refine the PCA model. Guo et al. [43] employ a FineNet to reconstruct the face details, whose ground truth is synthesized by inverse rendering. Although the methods capture fine-level geometric details (such as wrinkles), their global shapes still come from the 3DMM fitting. Unlike SfS, our method aims to improve the global visual effect such as facial feature topology and face contour.

In addition to single-view reconstruction, high-quality 3D faces can be better recovered with multiview inputs. Many 3D face datasets, e.g., D3DFACS [44] and FaceScape [15], use the calibrated depth camera arrays to reconstruct high-quality 3D shapes. Besides, recent deep learning based works have shown that multiview images of the same subject benefit shape reconstruction. DECA [45] proposes a detail-consistency loss between different images to disentangle expression-dependent wrinkles from person-specific details. MVF-Net uses the MultiPIE [46] samples collected by 15 cameras for multiview 3DMM parameters regression. Shang et al. [47] also adopt the MultiPIE to alleviate the pose and depth ambiguity during 3D face reconstruction. Lin et al. [48] take the RGB-D selfie videos captured by iPhone X to reconstruct high-fidelity 3D faces. Although our proposal focuses on 3D reconstruction from a single image, these multiview systems motivate our network structure to observe the input image in several calibrated views.

3 Background of 3D Face Reconstruction

This section describes the formulation of our main task and other related background information. In general, a 3D face can be decomposed into *coarse shape*, *personalized shape* and *pose*. The coarse shape is usually formulated by the seminal work of the 3D Morphable Model (3DMM) [49], which describes the 3D face space with PCA:

$$\mathbf{S} = \overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id} + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}, \tag{1}$$

where $\overline{\mathbf{S}}$ is the mean shape, $\mathbf{A}_{id}$ is the principle axes trained on the 3D face scans with neutral expression and $\boldsymbol{\alpha}_{id}$ is the shape parameter, $\mathbf{A}_{exp}$ is the principle axes trained on the offsets between expression scans and neutral scans and $\boldsymbol{\alpha}_{exp}$ is the expression parameter. Although widely implemented in various tasks [50], [51], [52], the coarse shape covers limited shape variations, leading to model-like reconstruction results. The personalized shape contains most person-specific shape morphing, which can be formulated as the vertex displacement $\Delta\mathbf{S}$ missed by the linear subspace. Finally, the pose, which is formulated by the camera parameters $\mathbf{C} = [f, \mathbf{R}, \mathbf{t}_{3d}]$, determines the rigid transformation to the image plane:

$$\mathcal{V}(\mathbf{C}, \mathbf{S} + \Delta\mathbf{S}) = f * \mathbf{R} * (\mathbf{S} + \Delta\mathbf{S}) + \mathbf{t}_{3d}, \tag{2}$$

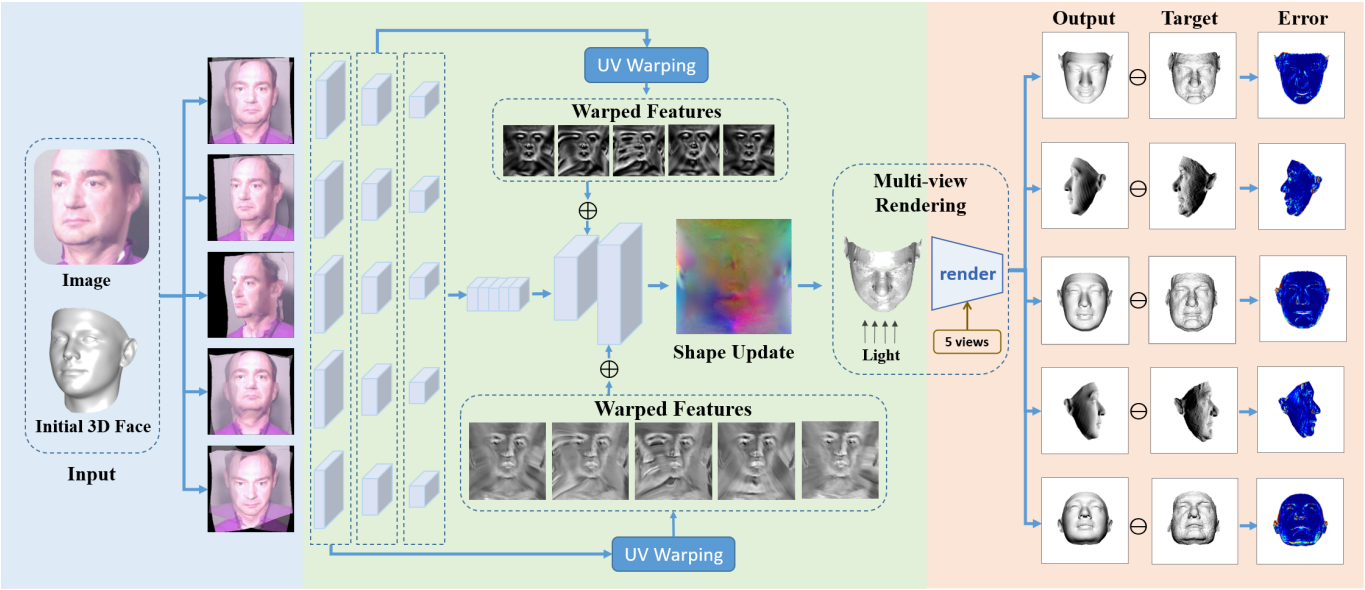
where $\mathcal{V}(\cdot,\cdot)$ is the rigid transformation, f is the scale factor, $\mathbf{R}$ is the rotation matrix, and $\mathbf{t}_{3d}$ is the 3D translation vector. Among these 3D components, camera parameter $\mathbf{C}$ and coarse shape $\mathbf{S}$ can be well estimated from a single image by recent 3DMM fitting methods [9], [10], [12]. However, there is little attention paid to the recovery of personalized shape $\Delta\mathbf{S}$ due to the shortage of training data and sophisticated recovery methods, which stimulates our motivation. It is worth noting that, different from the multiview reconstruction task [47], [48], which has a well-defined solution by stereo vision, single-view reconstruction is an ill-posed problem due to the loss of depth information. Thus, the goal of our task is *predicting* the personalized shape by the knowledge learned from large-scale training data.

To make CNN concentrate on personalized shape, we fit an *off-the-shelf* 3DMM [10] to estimate the coarse shape $\mathbf{S}$ and the camera parameter $\mathbf{C}$. Then our task can be formulated as:

$$\arg\min_{\theta} \|\mathcal{F}(\mathbf{S}^*) - \mathcal{F}(\mathit{Net}(\mathbf{I}, V(\mathbf{S}, \mathbf{C}); \theta) + \mathbf{S})\|, \tag{3}$$

where $\mathbf{S}^*$ is the ground-truth 3D shape, $\mathcal{F}(\cdot)$ is a 3D feature accounting for reconstruction quality, and the network $\mathit{Net}(\mathbf{I}, \mathcal{V}(\mathbf{S}, \mathbf{C}); \theta)$, with θ as its network weights, takes the image $\mathbf{I}$ and the 3DMM fitting result $\mathcal{V}(\mathbf{S}, \mathbf{C})$ as the input and predicts the personalized shape. It can be seen that the network structure $\mathit{Net}(\cdot)$ and the supervisory signals $\mathcal{F}(\cdot)$ are two critical topics in this formulation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Virtual Multiview (b) Many-to-one Hourglass (c) Plaster Sculpture Descriptor

Fig. 2. An overview of the High-Fidelity Reconstruction Network (HRNet). (a) The input image is rendered to 5 calibrated views to normalize pose variations while preserving the original image geometry. (b) A many-to-one hourglass network serves as the encode-decoder to fuse the multiview features and generate the vertex displacements as the personalized shape. (c) A plaster sculpture descriptor is utilized to optimize the visual effect by comparing the multiview images rendered from the shapes.

4 Virtual Multiview Network

The core problem of designing the network, i.e., $Net(\mathbf{I}, \mathbf{V})$ in Eqn. 3, is how to fully utilize the visual information of the input image $\mathbf{I}$ and the 3D prior embedded in the fitted 3DMM $\mathbf{V}$. The network should have three properties: 1) **Normalization:** The non-shape components of the input face should be normalized with the initially fitted 3DMM to make the network concentrate on capturing shape information. 2) **Lossless:** the geometry and texture in the original image should be preserved to provide complete information for the network. 3) **Concentration:** the receptive field of each outputted vertex should cover the most related image region to highlight the subtle appearance. In this section, we motivate our **V**irtual **M**ulti-view **N**etwork (**VMN**) from the multiview camera system used to generate complete 3D scans, where a volunteer sits in a fully constrained environment and several cameras circling around the head take the pictures simultaneously. Fig. 2 shows an overview of the proposed network.

4.1 Multiview Simulation

Given a face image, we aim to simulate its appearances at 5 constant poses, whose (pitch, yaw) angles are $(0°, 0°)$, $(0°, 25°)$, $(0°, 50°)$, $(15°, 0°)$ and $(-25°, 0°)$, as shown in Fig. 2(a). The view synthesis can be achieved by transferring the image $\mathbf{I}$ to a 3D object through the strong prior from the fitted 3DMM $\mathbf{V}$. Following face profiling [10], we tile anchors on the background, set their depth to the mean of the 3D face, and triangulate them to a 3D mesh $\mathbf{V}^I$, which can be rendered at any required views, as shown in Fig. 3(b). However, the texture in 3D mesh $\mathbf{V}^I$ is not complete. When the target pose is smaller than the original pose, the self-occluded region will be exposed, leading to large artifacts. To fill the occluded region, we also generate a flipped 3D mesh $\mathbf{V}^I_{flip}$ from the mirrored image and register it to $\mathbf{V}^I$ to complete the face texture, as shown in Fig. 3(c).

When rendering the 3D mesh at specified views, we find that each pixel is determined by either the original image or the flipped image, depending on the vertex visibility. The visibility score of each vertex is defined in terms of the angle between the vertex normal and the view direction:

$$vis(\mathbf{v}) = \begin{cases} \mathbf{l}^T \cdot \mathcal{N}(\mathbf{v}) + 2 & if \quad \mathbf{v} \in face \\ \mathbf{l}^T \cdot \mathcal{N}(\mathbf{v}) & if \quad \mathbf{v} \in background, \end{cases} \tag{4}$$

where $\mathbf{l} = [0, 0, 1]$ is the view direction and $\mathcal{N}(\mathbf{v})$ is the normal of vertex $\mathbf{v}$ in the original image. Note that the visibility scores of face vertices are increased by 2 to ensure that face regions overlap the background. The visibility scores, regarded as the face texture, are then rendered to the target view to obtain the visibility map, as shown in Fig. 3(e). Finally, the virtual face image at a specified view is constructed by:

$$\begin{aligned} &\mathbf{I}_v = \lambda \odot \mathcal{R}[\mathcal{V}(\mathbf{V}^I, \mathbf{C}_v)] + \lambda_{flip} \odot \mathcal{R}[\mathcal{V}(\mathbf{V}^I_{flip}, \mathbf{C}_v)], \ with \\ &\lambda(x, y) = 1, \lambda_{flip}(x, y) = 0 \quad if \ \mathbf{vis}(x, y) \geq \mathbf{vis}_{flip}(x, y) \\ &\lambda(x, y) = 0, \lambda_{flip}(x, y) = 0.5 \quad if \ \mathbf{vis}(x, y) < \mathbf{vis}_{flip}(x, y), \end{aligned} \tag{5}$$

where $\mathcal{R}(\cdot)$ renders the 3D mesh $\mathcal{V}(\mathbf{V}^I, \mathbf{C}_v)$ to a specified view, indicated by the camera parameter $\mathbf{C}_v$, λ is the weight map calculated by comparing the visibility maps **vis**, and $\odot$ is the element-wise production. As shown in Fig. 3(f), the invisible region is made transparent to indicate that the texture is not real but inpainted by face symmetry, which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

should guide the network to concentrate more on the visible texture. In the case that the input is the frontal view, we directly stretch the texture as the side-face texture, and the neural network is trained to handle the artifacts.

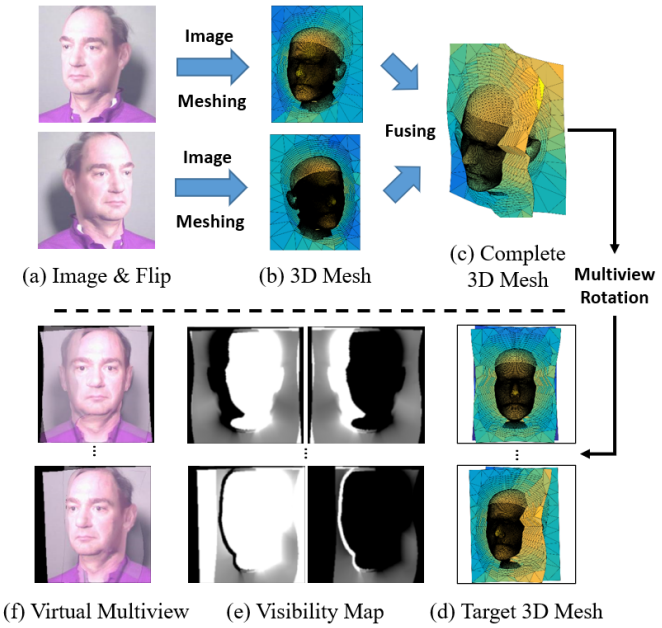


Fig. 3. An overview of multiview simulation. (a) The input image and the flipped one; (b) The 3D image mesh; (c) The complete 3D image mesh; (d) Rotating the 3D mesh to the target view; (e) The visibility map of the original image and the flipped image; (f) The generated virtual view.

4.2 Many-to-one Hourglass

To fuse the features extracted from multiview inputs, we propose a novel many-to-one hourglass network with a 5-stream encoder and a 1-stream decoder, as shown in Fig. 2(b). The encoder learns each view by a weight-shared CNN, and then concatenates the features as a multiview description. Afterwards, the decoder deconvolves the feature to the UV displacement map [24]. Besides, the intermediate features at symmetric layers are added together to merge high-level and low-level information. However, the particular structure of the many-to-one hourglass network poses a critical challenge for intermediate feature fusion. The features at one position have different semantic meanings, since the encoder has 5 different image views and the decoder is on the UV plane. Directly adding the features would degrade each other. To address the problem, each feature map in the encoders is warped to the UV plane according to the fitted 3DMM before being sent to the decoder.

Regarding the three properties desired for network design, the VMN fulfills the normalization property by transferring the input face to constant views. It also possesses the lossless property as little information is lost during the construction of virtual multiview input, i.e., the original topology and the external face regions are preserved. For the concentration property, since the intermediate features are all aligned to the UV plane, the receptive field can consistently concentrate on the most related region.

5 Loss Function

The loss function directly judges the reconstruction results according to the targets. Generally, the Mean Square Error (MSE) loss is employed to diminish the 3D coordinate error. Given the ground-truth shape $\mathbf{S}^*$, the initial 3D shape $\mathbf{S}$ and the output shape offset $\Delta\mathbf{S}$, we can optimize the MSE loss as:

$$\mathcal{L}^{mse} = \|\mathbf{S}^* - \mathbf{S} - \Delta\mathbf{S}\|^2. \tag{6}$$

In Section 1, we argue that it is the limited representation power of 3DMM and the lack of the 3D ground truth that accounts for the model-like reconstruction, as shown in Fig. 4(c). However, with the shape reconstruction network trained on sufficient data, the reconstruction results are still not visually discriminative when considering the MSE loss only, as shown in Fig. 4(d). The main reason is that the vertex coordinate error does not account for how humans observe a 3D object. Therefore, a powerful 3D feature accounting for the visual effect, i.e., the $\mathcal{F}(\cdot)$ in Eqn. 3, is crucial for loss functions. In this section, we propose a new **P**laster **S**culpture **D**escriptor (**PSD**) to model the visual effect and a **V**isual-**G**uided **D**istance (**VGD**) loss to supervise the network training.

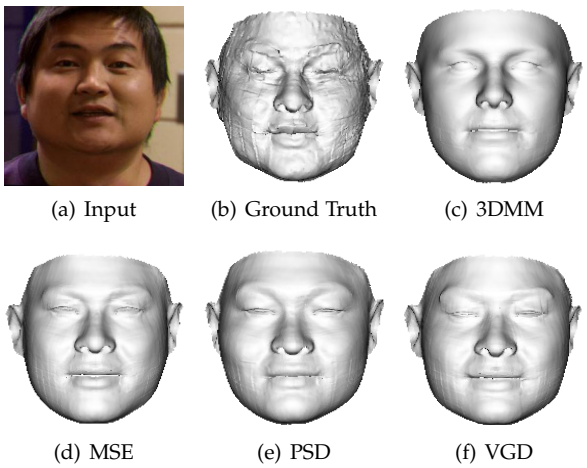


Fig. 4. The reconstructed shapes by different losses. (a) The input image; (b) The ground-truth shape; (c) The 3DMM fitting result; (d) Mean Squared Error; (e) Plaster Sculpture Descriptor; (f) Visual-guided Loss. First, the visual difference between 3DMM and MSE is significant, especially for the nose and eyes. Second, PSD slightly refines the facial feature topology but the improvement is not obvious. Finally, VGD further reconstructs a sharper jaw that resembles the ground truth.

5.1 Plaster Sculpture Descriptor

In 3D object retrieval, the light field descriptor [53] is widely computed for silhouette images of 3D shapes. Besides, recent analysis-by-synthesis 3D face reconstruction methods [17], [18], [36] optimize face images generated by a facial appearance model. Inspired by these two achievements, we consider that the images rendered by a 3D face can be utilized to model its visual effect, and explore a **P**laster **S**culpture **D**escriptor (**PSD**) to measure the reconstruction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

quality directly by how we see it. As shown in Fig. 2(c), a 3D shape is considered as a plaster sculpture with all-white vertex color and **orthogonal light** on it. During training, the 3D shape is rendered at 5 views, whose pitch and yaw angles are $(0°, 0°)$, $(0°, 90°)$, $(0°, -90°)$, $(30°, 0°)$ and $(-30°, 0°)$, and the L2 distances of the rendered images between the output and the ground-truth 3D shapes are formulated as the visual-effect distance:

$$\mathcal{D}^{psd} = \sum_{v=1}^{5} \|\mathcal{R}(\mathbf{R}_v * (\mathbf{S}+\Delta\mathbf{S}), \mathbf{T}^w) - \mathcal{R}(\mathbf{R}_v * \mathbf{S}^*, \mathbf{T}^w)\|_2, \quad (7)$$

where v is the view index, $\mathcal{R}(\cdot,\cdot)$ is the renderer whose input is 3D shape and texture, $\mathbf{R}_v$ is the rotation matrix corresponding to each view, and $\mathbf{T}^w$ is the all-white texture under orthogonal light. If we employ a differentiable renderer [54], the $\mathcal{D}^{psd}$ can be directly regarded as a loss function. As shown in Fig. 4(e), the facial features recovered by PSD are more similar to the ground truth, which is crucial in visual evaluation.

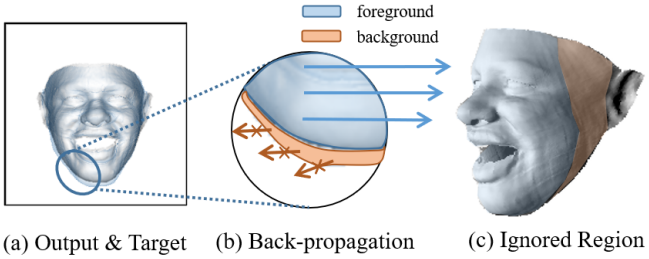


Fig. 5. The defect of Plaster Sculpture Descriptor (PSD) when regarded as a loss function. (a) The output 3D shape overlapped with its target. (b) The back-propagation of PSD. If an output pixel is located on the background, its corresponding vertex does not have back-propagated signals. (c) The trained (blue) and ignored (red) regions. The face contour is usually located on the background and loses the supervision.

5.2 Visual-guided Distance Loss

Although plaster sculpture descriptor measures how the shapes are observed by humans, there is an intrinsic problem on the face contour due to the defect of the differentiable renderer. As shown in Fig. 5, on the PSD error map, the differentiable renderer back-propagates the gradients of pixels to the corresponding vertices. Therefore, only the outputted vertices that located within the target face region have the back-propagated signals, and those on the background are not trained, leading to unsatisfactory contour reconstruction. To tackle this issue, we propose a **V**isual-**G**uided **D**istance (**VGD**) loss which considers the visual-effect distance as the vertex weights rather than the optimization target. The overview of VGD is shown in Fig. 6, whose insight is that, given the ground-truth position of each vertex, we only need to find which vertices dominate the visual effect. Specifically, we calculate the output-to-target and the target-to-output PSD errors simultaneously, on either of which the poorly fitted face contour inevitably brings large values. Then, we employ pixel-to-vertex mapping to retrace the pixel errors to vertices, and add the vertex errors from both the output and the target. Finally, the accumulated vertex errors from all the views are regarded as the vertex weights for the MSE loss:

$$\mathcal{L}^{vgd} = \mathbf{W}^{psd} \odot \|\mathbf{S}^* - \mathbf{S} - \Delta\mathbf{S}\|^2, \quad (8)$$
$$\mathbf{W}^{psd} = \sum_{v=1}^{5} [\mathcal{R}^{-1}_{\mathbf{S}+\Delta\mathbf{S}}(\mathcal{D}^{psd}_v) + \mathcal{R}^{-1}_{\mathbf{S}^*}(\mathcal{D}^{psd}_v)],$$

where $\mathbf{W}^{psd}$ represents the vertex weights, $\mathcal{R}^{-1}_{\mathbf{S}}(\cdot)$ denotes the inverse rendering function that maps pixel errors to the vertices according to the 3D face $\mathbf{S}$, and $\mathcal{D}^{psd}_v$ denotes the PSD error map in the vth view. As shown in Fig. 4(f), the visual-guided weights enable the network to focus on the visual discriminative regions, such as the face contour and the rough regions.

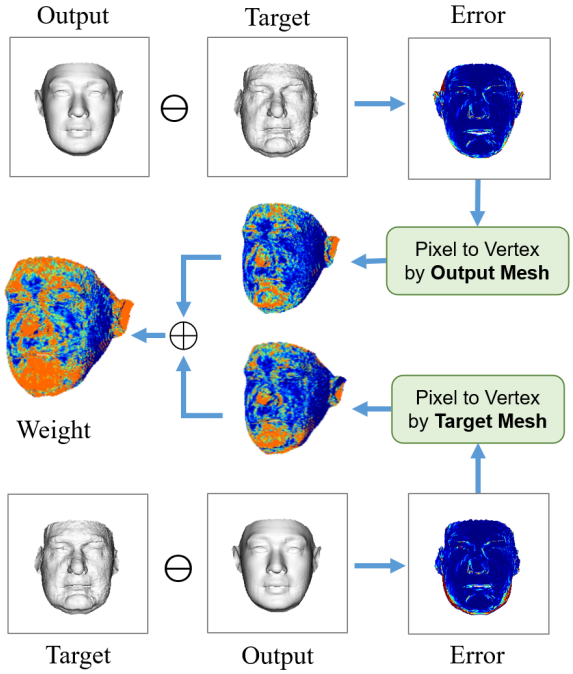


Fig. 6. Illustration of visual-guided weights calculation in one view. The output-to-target and the target-to-output PSD errors are traced back to the output and the target 3D mesh, respectively. The added errors are considered the vertex weights measuring the contribution to the visual effect.

6 Data Construction

It is very tedious to acquire complete and high-precision 3D faces in real circumstances. The raw scans must be captured under well-controlled conditions and registered to a face template through laborious hand-labeling. To collect large-scale 3D data, we construct the 3D ground truth from single-view RGB-D images, which is more promising for massive collection, especially considering the rapid development of hand-held depth cameras such iPhone X.

6.1 RGB-D Registration

The training data requires face images and their corresponding topology-uniform 3D shapes. Thus, the first task is registering a template to the depth images, where Iterative Closest Point (ICP) method is commonly adopted.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

However, plausible registered results are not sound enough to be the supervision of neural network training. Semantic consistency, i.e., each vertex has the same semantic meaning across 3D faces is a critical criterion. As shown in Fig. 7(a), semantically consistent registration cannot be achieved by finding the closest point in (x, y, z) only. Considering that our registration target is RGB-D data, whose texture contains discriminative features for vertex matching, we introduce two terms on the RGB channels to provide more constraints in the popular optimal nonrigid-ICP loss function [55].

The first term comes from the edge. As shown in Fig. 7(b), a series of landmarks, which are dense enough to finely enclose the eyes, eyebrows and mouth, are detected to represent the edges on the facial features. The constraint is formulated as:

$$E_{edge} = \left\| \begin{bmatrix} a^{11} & a^{12} & a^{13} & a^{14} \\ a^{21} & a^{22} & a^{23} & a^{24} \end{bmatrix} \begin{bmatrix} x^{3D} \\ y^{3D} \\ z^{3D} \\ 1 \end{bmatrix} - \begin{bmatrix} x^{2D} \\ y^{2D} \end{bmatrix} \right\|, \tag{9}$$

where $\begin{bmatrix} a^{11} & a^{12} & a^{13} & a^{14} \\ a^{21} & a^{22} & a^{23} & a^{24} \end{bmatrix}$ is the first and second rows of the 3×4 affine transformation matrix of each vertex, which should be optimized, (x^{3D}, y^{3D}, z^{3D}) is an edge 3D landmark and (x^{2D}, y^{2D}) is the corresponding 2D landmark detected on the image.

The second term aims to regularize the face contour. However, it is unreliable to optimize the vertex-to-landmark distances as in Eqn. 9, since there is no strict correspondence between 3D and 2D contour landmarks [56]. To utilize contour landmarks, we perform curve fitting rather than landmark matching to reduce the semantic ambiguity. As shown in Fig. 7(c), the 2D contour landmarks are sequentially connected to a contour curve. Then, the 3D contour vertices found by landmark marching [56] are fitted to the 2D curve by:

$$E_{cont} = \left\| \begin{bmatrix} a^{11} & a^{12} & a^{13} & a^{14} \\ a^{21} & a^{22} & a^{23} & a^{24} \end{bmatrix} \begin{bmatrix} x^{3D} \\ y^{3D} \\ z^{3D} \\ 1 \end{bmatrix} - \begin{bmatrix} x_c^{2D} \\ y_c^{2D} \end{bmatrix} \right\|, \tag{10}$$

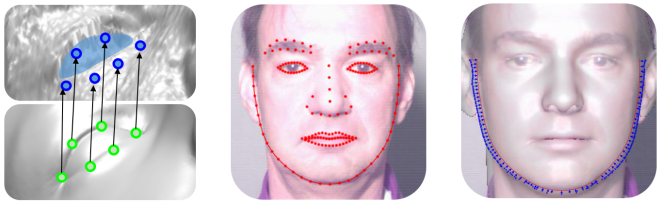
$$(x_c^{2D}, y_c^{2D}) = \arg\min_{(x,y)} \|(x, y) - (x^{3D}, y^{3D})\|, \forall (x, y) \in \mathbb{C},$$

where (x_c^{2D}, y_c^{2D}) is the 2D-closest point on the curve $\mathbb{C}$ to the 3D vertex (x^{3D}, y^{3D}, z^{3D}). During ICP registration, Eqn. 9 and Eqn. 10 are employed as additional terms constraining the first two rows of affine transformations, and the third row is optimized separately by traditional ICP constraints [55].

Given the registered face $\mathbf{V}_{regist}$, we further disentangle rigid and non-rigid transformations by:

$$\arg \min_{\mathbf{S}, \mathbf{R}, f, \mathbf{t}_{3d}} \|\mathbf{V}_{regist} - (f * \mathbf{R} * \mathbf{S} + \mathbf{t}_{3d})\|, \tag{11}$$

where $(f, \mathbf{R}, \mathbf{t}_{3d})$ are the rigid transformation parameters and the optimized shape is regarded as the ground-truth shape $\mathbf{S}^*$. The difference between the ground-truth shape and the 3DMM-fitted shape $\Delta\mathbf{S} = \mathbf{S}^* - \overline{\mathbf{S}} - \mathbf{A}_{id}\boldsymbol{\alpha}_{id} - \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}$ will be the target of the neural network training.



(a) Semantic Inconsistency (b) 2D Landmarks & Contour Curve (c) Curve Fitting

Fig. 7. RGB-D registration. (a) ICP only cannot guarantee semantic consistency. The example shows that the closest points on the target found by the eye-contour vertices are located outside the eye. (b) The landmarks on the facial-feature edge and the face contour curve. (c) The 3D contour vertices are fitted to the 2D contour curve.

6.2 Full-view Augmentation

Currently, most of the public RGB-D images are frontal faces [57], [58], [59], leading to the risk of poor pose robustness. Although RGB-D data can be rendered to other views, the results suffer from large artifacts due to the incomplete depth channel and face texture. To address this issue, a full-view augmentation method is proposed specifically for RGB-D data. Based on the face profiling method [10], we complete the depth channel for the whole image space, where the depth on the face region comes directly from the registered 3D face and the depth on the background is coarsely estimated by some anchors, as shown in Fig. 8(b). Specifically, these anchors (x_i, y_i) are triangulated to a background mesh and their depth values d_i are regularized by a depth-channel constraint and a smoothness constraint:

$$\begin{aligned} &\sum_i Mask(x_i, y_i)\|d_i - Depth(x_i, y_i)\| + \\ &\sum_i \sum_j Connect(i, j)\|d_i - d_j\|, \end{aligned} \tag{12}$$

where $Depth(x, y)$ is the depth channel of the RGB-D data, $Mask(x, y)$ indicates whether (x, y) is hollow, and $Connect(i, j)$ indicates whether two anchors are connected by the background mesh. With the complete depth channel (Fig. 8(c)), the RGB-D can be rotated (Fig. 8(d)) and rendered (Fig. 8(e)) to any views. More implemental details are provided in the supplemental materials.

Different from face profiling [10], whose purpose is enlarging medium poses to large poses, our method aims to augment frontal faces. However, the main drawback of the rotate-and-render strategy is that, when rotating from frontal faces, there are serious artifacts on the side face due to the incomplete face texture, as shown in Fig. 8(e). In this work, we employ the texture and illumination model in 3DMM as a strong prior to refine the artifacts. Specifically, the face texture is modeled by PCA [13]:

$$\mathbf{T} = \overline{\mathbf{T}} + \mathbf{B}\boldsymbol{\beta}, \tag{13}$$

where $\overline{\mathbf{T}}$ is the mean texture, $\mathbf{B}$ is the principle axis of the texture and $\boldsymbol{\beta}$ is the texture parameter. Given 3D shape $\mathbf{V}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

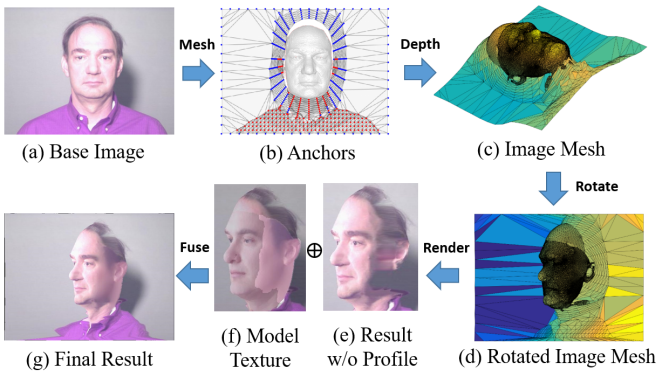


Fig. 8. An overview of full-view augmentation. (a) The base image; (b) The original depth channel and the anchors on the background. The red anchors have depth values and the blue anchors are located on the hollow, they have different constraints in Eqn. 12; (c) The completed depth channel; (d) The rotated 3D mesh; (e) Rendering with image pixels and model texture; (f) The augmentation result.

and texture $\mathbf{T}$, the Phong illumination model is used to produce the color of each vertex [49]:

$$C_i(\mathbf{p}_{tex}) = \mathbf{Amb} * \mathbf{T}_i + \mathbf{Dir} * \mathbf{T}_i * \langle \mathbf{n}_i, \mathbf{l} \rangle + k_s \cdot \mathbf{Dir} \langle \mathbf{r}_i, \mathbf{ve} \rangle^{\nu}, \tag{14}$$

where C_i is the RGB of the ith vertex, the diagonal matrix $\mathbf{Amb}$ is the ambient light, the diagonal matrix $\mathbf{Dir}$ is the parallel light from direction $\mathbf{l}$, $\mathbf{n}_i$ is the normal direction of the ith vertex, k_s is the specular reflectance, $\mathbf{ve}$ is the viewing direction, ν controls the angular distribution of the specular reflection and $\mathbf{r}_i = 2 \cdot \langle \mathbf{n}_i, \mathbf{l} \rangle \mathbf{n}_i - \mathbf{l}$ is the direction of maximum specular reflection. The collection of texture parameters is $\mathbf{p}_{tex} = [\boldsymbol{\beta}, \mathbf{Amb}, \mathbf{Dir}, \mathbf{l}, k_s, \nu]$. Given the ground-truth 3D shape $\mathbf{V}_{regist}$, the texture parameters $\mathbf{p}_{tex}$ can be estimated by the analysis-by-synthesis manner:

$$\arg \min_{\mathbf{p}_{tex}} \|\mathbf{I}(\mathbf{V}_{regist}) - C(\mathbf{p}_{tex})\|, \tag{15}$$

where $\mathbf{I}(\mathbf{V})$ is the image pixels at the vertex positions. The optimized result $C(\mathbf{p}_{tex})$ is the face texture, as shown in Fig. 8(f). With the estimated model texture, we render the 3D face with both the image pixels and the model texture, as shown in Fig. 8(e) and Fig. 8(f). Then, we inpaint the self-occluded side face with the model texture through Poisson editing, as shown in Fig. 8(g). It can be seen that the model texture is realistic enough to inpaint the smoothly textured side face.

6.3 Shape Transformation

In addition to pose variations, shape variations covered by the training data are also important since the personalized shape is the main goal. Unfortunately, existing data does not contain sufficient identities due to demanding data collection. In this section, we propose a shape augmentation method to transform the underlying 3D shape of a face image and refine the face appearance accordingly. In specific, we first construct the target shape to transform, whose eye, nose, mouth, and cheek come from different faces in the datasets, as shown in Fig. 9(a). Second, the base image is tuned into a 3D mesh following the same process in the full-view augmentation, as shown in Fig. 9(b). Third, we replace the 3D shape and warp the background to adjust the new face. Specifically, the background anchors are adjusted by minimizing the following function:

$$\begin{aligned} &\sum_i FaceContour(i) * (\|x_i^t - x_i^s\| + \|y_i^t - y_i^s\|) + \\ &\sum_i \sum_j Connect(i,j) * (\|(x_i^t - x_j^t) - (x_i^s - x_j^s)\| \\ &+ \|(y_i^t - y_j^t) - (y_i^s - y_j^s)\|), \end{aligned} \tag{16}$$

where (x_i^s, y_i^s) is the anchor position on the source image, (x_i^t, y_i^t) is its target position on the augmented image, $FaceContour(i)$ indicates whether anchor i is located on the face contour (the red points in Fig. 9(b)), and $Connect(i,j)$ indicates whether two anchors are connected by the background mesh. Afterwards, we render the warped 3D mesh and obtain the shape-transformed result, whose ground-truth 3D shape is the target 3D shape, see Fig. 9(d).

In addition to image pixel warping, shading adjustment is also important according to the shape-from-shading theory [32], which can be achieved by the illumination model the same as Eqn. 14:

$$\mathbf{C}_i^t = \mathbf{Amb} * \mathbf{T}_i + \mathbf{Dir} * \mathbf{T}_i * \langle \mathbf{n}_i^t, \mathbf{l} \rangle + k_s \cdot \mathbf{Dir} \langle \mathbf{r}_i^t, \mathbf{ve} \rangle^{\nu}, \tag{17}$$

where all the parameters except $\mathbf{n}^t$ and $\mathbf{r}^t$ are from the source image, and $\mathbf{n}^t$ and $\mathbf{r}^t$, which account for shading, are from the target shape. Finally, we change the facial pixels to $\mathbf{C}^t$ and obtain the final result of shape transformation, see Fig. 9(f). We provide more implemental details in the supplemental materials.

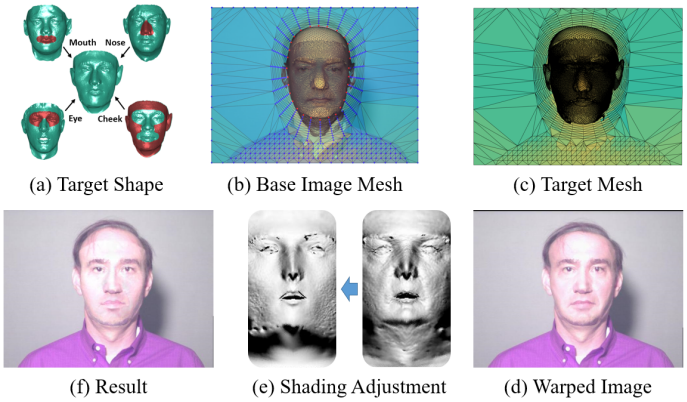


Fig. 9. An overview of shape transformation. (a) Target shape generation by fusing facial features of different identities; (b) The original image and its 3D mesh; (c) The target 3D mesh; (d) The result of shape transformation and background warping; (e) The shading adjustment; (f) The final result.

7 Benchmarks

7.1 Dataset

Fine-Grained 3D Face (FG3D) is constructed from three datasets, FRGC, BP4D, and CASIA-3D. FRGC [59] includes $4,950$ samples, each of which has a face image and a 3D

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

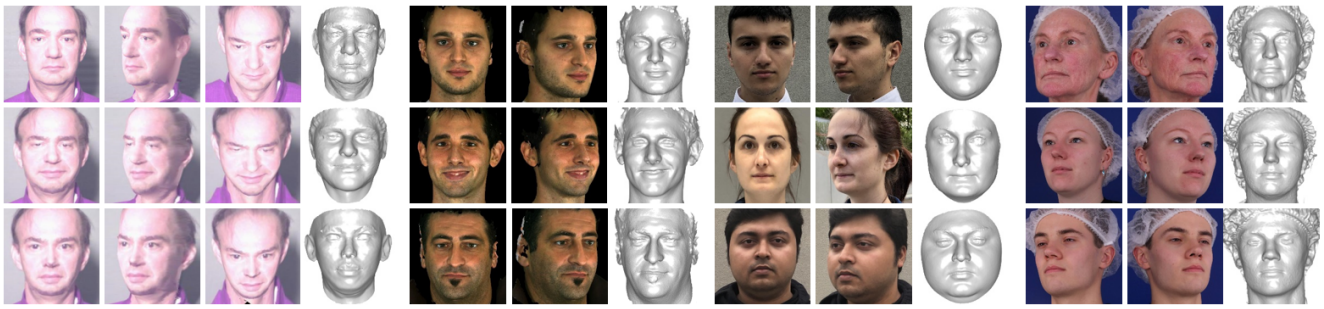


Fig. 10. Selected examples from the datasets employed in the experiments. Each face image has a 3D ground-truth shape.

scan in full correspondence. BP4D [58] contains 328 2D+3D videos from 41 subjects, where 3,376 frames are randomly selected in total. CASIA-3D [57] consists of 4,624 scans of 123 persons and the non-frontal faces are filtered out.

We divide 90% of subjects as the training set **FG3D-train** and the remaining 10% of subjects as the testing set **FG3D-test**. For the training set, we perform shape transformation as in Section 6.3 4 times to augment shape variations. Then, the out-of-plane pose variations are augmented by full-view augmentation as in Section 6.2, where the yaw angle is enlarged at steps of 15° until 50° and the $pitch$ angle is enlarged by 15° and -25°, generating $474k$ samples in total. For the testing set, we only perform full-view augmentation and generate $12k$ samples. Besides, we manually delete the bad registration results in the FG3D-test manually for better evaluation. This dataset is employed for the model training and the performance analysis of our proposal. Some examples are shown in Fig. 10(a).

Florence [60] is a 3D face dataset containing 53 subjects with 3D meshes acquired from a structured-light scanning system. Based on the protocol in [9], each subject is rendered at pitches of -20°, 0°, 20° and yaws of -45° to 45° at the step of 15°, as shown in Fig. 10(b). Besides, we register each 3D mesh with hand-labeled landmarks and carefully check the registration results. This dataset is used for cross-dataset evaluation to demonstrate the generalization.

NoW [61] contains 2054 face images of 100 subjects, which are split into an open validation set (20 subjects) and a private test set (80 subjects). Each subject has face images with different poses and expressions, and one neutral 3D face scan for reference. Under the original protocol, the reconstruction results should be disentangled from the expression to compare with the neutral 3D scan, which is not consistent with our goal. Therefore, we only select the neural-expression images of the validation set, as shown in Fig. 10(c), for cross evaluation.

Stirling/ESRC [62] is a 3D face dataset with 134 subjects (64 males and 70 females). Each subject has a 3D face scan in a neutral expression and two corresponding face images in $yaw = \pm 45^\circ$, which are captured by the Di3D camera system simultaneously. In the experiments, all subjects are employed in the cross-dataset evaluation. Some examples are shown in Fig. 10(d).

7.2 Evaluation Metric

It is still an open problem to determine the proper measurement of 3D shape accuracy. Errors calculated after projection [9], [10], [63] mostly account for pose accuracy since pose is the dominant factor of vertex positions [10]. To highlight shape accuracy, which is our main purpose, we normalize pose before error calculation. Specifically, we register a face template to the ground-truth 3D face as in Section 6.1, finding the vertex correspondence (k, k^t), where k is the vertex index on the face template and k^t is that on the ground truth. We also note $k \in \mathcal{C}$ if k belongs to the face region and the correspondence (k, k^t) is reliable (the spatial and normal distances are below thresholds). For each reconstruction result $\mathbf{V}$ and the ground truth $\mathbf{V}^*$, we estimate a rigid transformation:

$$\arg \min_{f,\mathbf{R},\mathbf{t}_{3d}} \sum_{k \in \mathcal{C}} \|(f * \mathbf{R} * \mathbf{v}_k + \mathbf{t}_{3d}) - \mathbf{v}^*_{k^t}\|, \tag{18}$$

where $(f, \mathbf{R}, \mathbf{t}_{3d})$ are the rigid transformation parameters for pose alignment. Based on this rigid alignment, we utilize the widely applied Normalized Mean Error (NME) to measure the reconstruction error:

$$NME = \frac{1}{K} \sum_{k=1}^{K} \frac{||(f * \mathbf{R} * \mathbf{v}_k + \mathbf{t}_{3d}) - \mathbf{v}^*_{k^t}||}{d}, \tag{19}$$

where K is the number of vertices, $\mathbf{v}_k$ is a vertex on the reconstructed face, $\mathbf{v}^*_{k^t}$ is the target of $\mathbf{v}_k$, and d is the 3D outer interocular distance. This error is adopted in the performance analysis on the FG3D-test, where we trust the registration results.

We also employ a Densely Aligned Chamfer Error (DACE) to measure the distances of the closest points:

$$DACE = \frac{1}{\mathcal{N}(C)} \sum_{k \in \mathcal{C}} \frac{||(f * \mathbf{R} * \mathbf{v}_k + \mathbf{t}_{3d}) - \mathbf{v}^*_{k^{nn}}||}{d}, \tag{20}$$

where $\mathbf{v}^*_{k^{nn}}$ is the nearest neighbor of $\mathbf{v}_k$ on the raw 3D scan after rigid alignment. We adopt this error in the comparison experiments on Florence, NoW and Stirling/ESRC, where the raw 3D scans are employed as the target and different methods may have different topologies. It is worth noting that, only the vertex in set $\mathcal{C}$, which indicates that its ICP matching is reliable, participates in the error calculation, which filters out the errors on the noisy, hollow and occluded regions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

8 Experiments

8.1 Implementation Details

The initial 3D face is acquired by 3DDFA [64]. The backbone follows the network defined in PRNet [65] and the shortcuts are added to generate an hourglass network. The models are trained by the SGD optimizer with a starting learning rate of 0.1, which is decayed by 0.1 at epochs 30, 35 and 40, and the model is trained for 45 epochs. The training images are cropped by the bounding boxes of the initial 3DMM fitting results [64] and resized to 256×256 without any perturbation. The UV displacement map is also 256×256. In all the following experiments, the FG3D-train is employed as the training set. The testing is conducted on the FG3D-test for performance analysis with Normalized Mean Error (NME), and several other datasets for comparison with the state-of-the-art methods, where Densely Aligned Chamfer Error (DACE) is preferred.

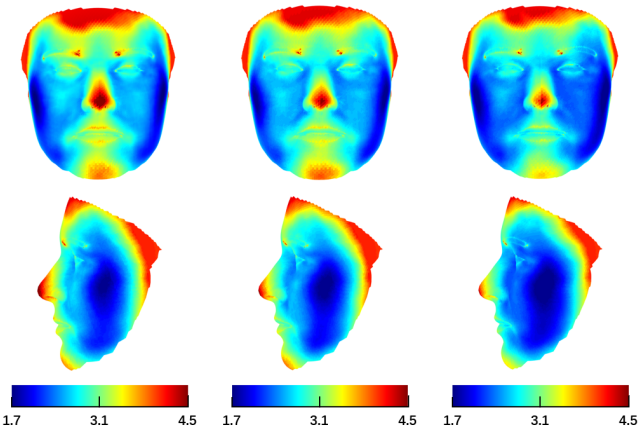
8.2 Network Structure Analysis

In this section, we thoroughly analyze how the network structure benefits the reconstruction accuracy.

8.2.1 Ablation Study on Network Properties

In Section 4, we introduce three properties desired for network design: normalization, lossless and concentration, which require the network to normalize non-shape components, preserve the image information, and align receptive fields. To evaluate the benefits of the properties, we test two alternative networks proposed in our preliminary work [24] that fulfill different levels of them. First, the Camera View Network (CVN) [24] concatenates the original image and the Projected Normalized Coordinate Code (PNCC) [10] encoded by the fitted 3DMM to regress the vertex displacements. The network fulfills the lossless property by preserving the original image. However, the pose variations are not normalized and the receptive field does not cover the most related region since the input and the output have different coordinate systems (image plane vs. UV plane). Thus, the normalization and the concentration properties are missed. Second, the Model View Network (MVN) [24] performs explicit normalization by warping the image to the UV plane, making the receptive field cover the most related region. However, the normalization loses the 2D geometry and the external face regions of the original input. Thus, the lossless property is violated. Compared with them, our Virtual Multiview Network (VMN) fulfills all the principles by constructing a virtual multiview camera system and aligning intermediate deep features to the UV plane.

We evaluate each network and the initial 3DMM fitting results in Table 1, and find that all the networks improve the initial shape. Besides, we observe that the more properties the network fulfills, the better its performance. The improvement achieved by replacing CVN with MVN reflects the effectiveness of explicitly normalizing non-shape components. Replacing MVN with VMN further promotes the performance, indicating the significance of preserving the original 2D geometry and the external face regions. Another interesting observation is that the medium poses, with yaw angles between $[30°, 45°]$, are the best for shape reconstruction, owing to the depth information for both side view and frontal view.



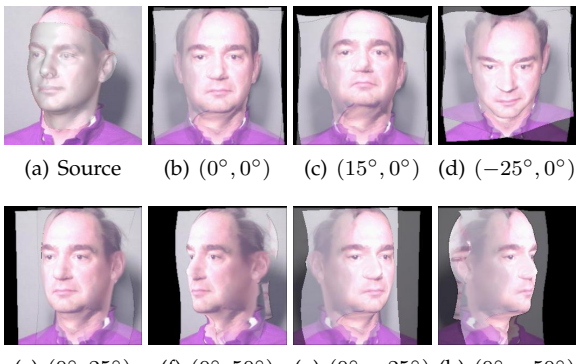
(a) CVN (b) MVN (c) VMN

Fig. 11. The vertex-level NME of our method and other baselines, calculated by the mean of FG3D-test. Lower value indicates better accuracy. (a) Camera View Network (CVN), (b) Model View Network (MVN) and (c) our method (VMN).

We further demonstrate the error at a more fine-grained vertex level, as shown in Fig. 11. First, there is a clear improvement from CVN to MVN, and further to VMN. Second, the improvements in the nose and jaw are the most obvious, followed by eye, mouth and apple cheeks. It is implied that personalized shape is mainly encoded in these regions.

8.2.2 Input Views of Virtual Multiview Network

In virtual multiview network, the input image is rendered to several constant views to simulate a multi-camera system. In this section, we further analyze how many views are appropriate to describe the face shape, ranging from the frontal view only to multiple views with rich pose variations. Totally 7 views are evaluated, whose (pitch, yaw) are $(0°, 0°)$, $(15°, 0°)$, $(-25°, 0°)$, $(0°, 25°)$, $(0°, 50°)$, $(0°, -25°)$ and $(0°, -50°)$, as shown in Fig. 12.



(a) Source (b) $(0°, 0°)$ (c) $(15°, 0°)$ (d) $(-25°, 0°)$

(e) $(0°, 25°)$ (f) $(0°, 50°)$ (g) $(0°, -25°)$ (h) $(0°, -50°)$

Fig. 12. The 7 virtual views of the input face. The angles in the bracket are (pitch, yaw).

Table 2 evaluates 4 combinations of the 7 views, where the original image is regarded as the baseline. First, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
The Normalized Mean Error (NME) of different networks, evaluated on the FG3D-test with different yaw intervals. "Lossless", "Normalize" and "Concentrate" are the three properties desired for network structures, which are discussed in Section 4, The symbol ✓ means the property is fulfilled.

Network Structure	Lossless	Normalize	Concentrate	$[0, 15]$	$[15, 30]$	$[30, 45]$	Mean
Initial Shape (3DMM)				6.13	6.13	6.09	6.11
Camera View (CVN) [24]	✓			3.68	3.56	3.46	3.57
Model View (MVN) [24]		✓	✓	3.54	3.47	3.46	3.49
Virtual Multiview (VMN)	✓	✓	✓	**3.43**	**3.26**	**3.24**	**3.32**

TABLE 2
The Normalized Mean Error (NME) with different numbers of views, evaluated on all the samples of FG3D-test. The symbol ✓ means the view in (pitch, yaw) is employed.

Pitch	$0°$	$15°$	$-25°$	$0°$	$0°$	$0°$	$0°$	NME
Yaw	$0°$	$0°$	$0°$	$25°$	$50°$	$-25°$	$-50°$	
1-View	✓							3.63
3-View	✓			✓	✓			3.40
5-View	✓	✓	✓	✓	✓			**3.32**
7-View	✓	✓	✓	✓	✓	✓	✓	3.35
Source	same as input							3.60

1-view does not perform well, even worse than the original input. It is worth noting that the 1-view setting is equivalent to the frontalization strategy [56], [66], [67], [68], [69], suggesting that, although widely applied in face recognition [69], frontalization is not appropriate in shape reconstruction due to the loss of profile. Besides, the 3-view outperforms the 1-view with the provided yaw variations, verifying the benefits of side view information. The 5-view further achieves the best result with the additional pitch variations. Finally, the 7-view indicates that the yaw angles inverse to the original pose make little difference. Thus, we exclude them.

8.2.3 *Multiview Fusion in Many-to-one Hourglass*

The virtual multiview network aims to fuse the features from 5 views and output the personalized shape on the UV plane, where a many-to-one hourglass network is employed. Since the inputs and the output lie on different planes (image plane vs. UV plane), when the features are fused, e.g., concatenated in the middle and added in the symmetric layers, the fused features have different receptive fields and may degrade each other. We first employ direct feature fusion and achieve unsatisfactory performance, as shown in Figure 13. Second, we implement a weak alignment manner that first warps the input images to the UV plane and then performs convolution. Although slight improvements are achieved, this warp-then-convolve manner violates the lossless requirement since the face contour and the external face regions are lost. Furthermore, we evaluate the proposed method that first convolves the images and then warps the intermediate features to the UV plane before feature fusion. This convolve-then-warp manner achieves the best performance, showing the importance of preserving the information in the original image. Finally, we attempt to shrink the features on the occluded region as in [10], [24], but find subtle difference, which may be attributed to the self-occlusion inpainting manner in Sec. 4.1. To simplify the network, we do not employ the shrinking module.

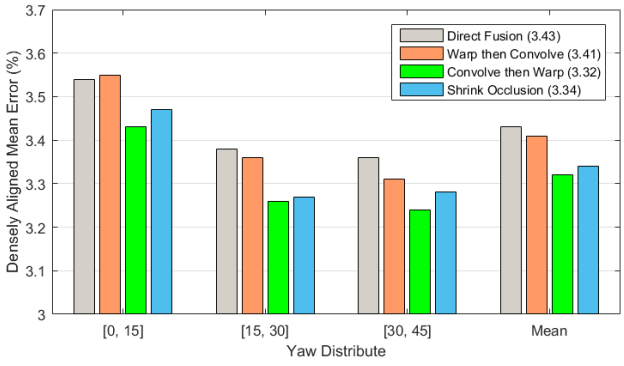


Fig. 13. The Normalized Mean Error (NME) of different multiview fusion methods, evaluated on the FG3D-test with different yaw intervals. The value in bracket is the mean NME of all the samples.

8.3 Data Augmentation Analysis

In Section 6, we augment face appearance in both pose and shape to provide adequate data for neural network training. In pose augmentation, we improve face profiling [10] by completing the depth channel and inpainting the side face with a texture model. As shown in Table 3, our full-view augmentation outperforms face profiling, especially in large poses, which is attributed to the artifacts exposed on the side face that face profiling cannot repair. Besides, by further incorporating shape transformation, the training data reaches 474k samples and the error is greatly reduced by 15.7%. It can be seen that in high-fidelity reconstruction, where shape is mostly concerned, our shape transformation is a highly effective augmentation strategy. However, adding an overwhelming number of augmented samples may not benefit the performance as the shape-augmented samples are fake images. Therefore, we further add the shape-transformed samples progressively and observe how the performance changes. As shown in Fig. 14, the error reduction converges when approximately 265k samples (2.2 times that of the original samples) are added.

8.4 Loss Function Analysis

In supervised learning, the loss function directly judges the reconstruction results according to the targets. Considering that the fine-grained geometry cannot be well captured by the traditional Mean Squared Error (MSE), we propose the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
The Normalized Mean Error (NME) of different data augmentation strategies, evaluated on the FG3D-test with different yaw intervals. "Num" is the number of training samples, "Pose Aug" indicates the full-view augmentation and "Shape Aug" refers to the shape transformation.

Augmentation	Num	[0, 15]	[15, 30]	[30, 45]	NME
Face Profiling [10]	120k	3.50	3.44	3.58	3.51
Pose Aug	120k	3.43	3.26	3.24	3.32
Pose & Shape Aug	**474k**	**2.93**	**2.71**	**2.74**	**2.80**

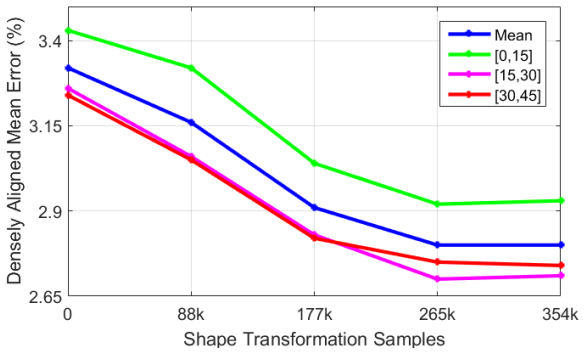


Fig. 14. The performance improvement as more shape-augmented samples are added in training, evaluated on the FG3D-test with different yaw intervals.

TABLE 4
The Normalized Mean Error (NME) and the Densely Aligned Chamfer Error (DACE) of different losses on the FG3D-test.

Loss Function			NME	DACE
MSE	PSD	VGD		
✓			2.80	1.50
✓	✓		2.75	1.47
	✓	✓	2.72	1.47
		✓	**2.66**	**1.43**

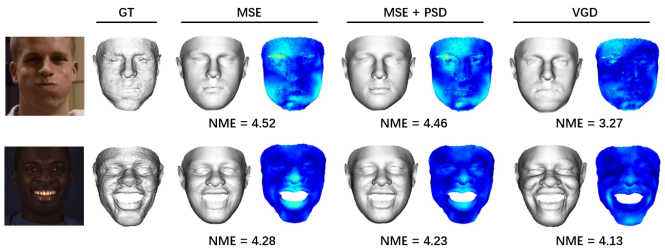


Fig. 15. Visual demonstration with different losses.

Plaster Sculpture Descriptor (PSD) to model the visual effect and the Visual-Guided Distance (VGD) to supervise the training. We evaluate the loss functions by both Normalized Mean Error (NME) and Densely Aligned Chamfer Error (DACE), where the latter concerns more about the visual effect.

The results listed in Table 4 indicate that: 1) Even with the ground-truth 3D shape as the supervision, intuitively adopting MSE cannot capture personalized shape well. 2) The introduction of PSD improves the accuracy by directly optimizing the visual effect. However, PSD cannot constrain the face contour due to its defect in back-propagation, which is discussed in Section 5.2. 3) The VGD loss further remedies the defects of PSD and captures the face contour well, achieving the best performance. 4) The unsatisfactory results achieved by the combination of PSD and VGD indicate that, it is better to regard PSD as the vertex weights than the loss function. There are also some examples in Fig. 15.

The PSD draws the 3D face as a plaster sculpture with white vertex color under frontal light, which is common in 3D structure demonstration. We also evaluate the performance of PSD by rendering the 3D face in different vertex colors and lighting conditions, finding little difference.

8.5 Comparison Experiments

8.5.1 Quantitative Comparison

Protocol: To quantitatively compare our method with prior works, we evaluate the single-view 3D reconstruction performance on Florence [60], NoW [61] and Stirling/ESRC [62], where the ground-truth 3D shape is available. For fairness, the FG3D-test is not employed due to its similar capture environment with our training data. We evaluate Florence and NoW according to different yaw intervals, and only report the mean accuracy of Stirling/ESRC due to its limited pose variations. The accuracy is measured by the Densely Aligned Chamfer Error (DACE) due to the difference in mesh topology.

Counterparts: The compared 3D face reconstruction approaches include 3DMM fitting in supervised [10], [12] and weakly-supervised [15], [63] manners, vertex regression [9], Shape from Shading [70] and Non-linear 3DMM [17].

3DDFA [10] is a representative 3DMM fitting method that regresses 3DMM parameters in a supervised fashion, which is recently improved by 3DDFA-v2 [12] in generalization through meta-learning. FaceScape [15] fits a new 3DMM with higher geometric quality than 3DDFA. Deng et al. [63] further introduce weakly-supervised learning that incorporates low-level and perception-level information. PRNet [9] bypasses 3DMM by directly regressing all the vertex coordinates in one propagation, which potentially helps the network cover a larger shape space than the linear model. Extreme3D [70] attempts to recover the geometric details by Shape from Shading (SfS). Despite the impressive recovered wrinkles and pores, SfS concentrates on the fine-level details. Their global shapes, such as face contour and facial feature topology, still originate from a fitted 3DMM. Non-linear 3DMM [17] achieves a certain breakthrough by learning a non-linear face model in an analysis-by-synthesis manner, which covers a larger shape space than linear 3DMMs.

Results: Table 5 lists the comparison results and Fig. 17 shows the corresponding Cumulative Errors Distribution (CED) curves. Taking Florence as a representative, there are several interesting observations when only shape error is evaluated: 1) Although 3DDFA-v2 is far better than 3DDFA when evaluating the projected 3D faces [12], their shapes are close. 2) As a non-linear model, PRNet performs similarly to the linear 3DDFA-v2, since its output is still limited by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

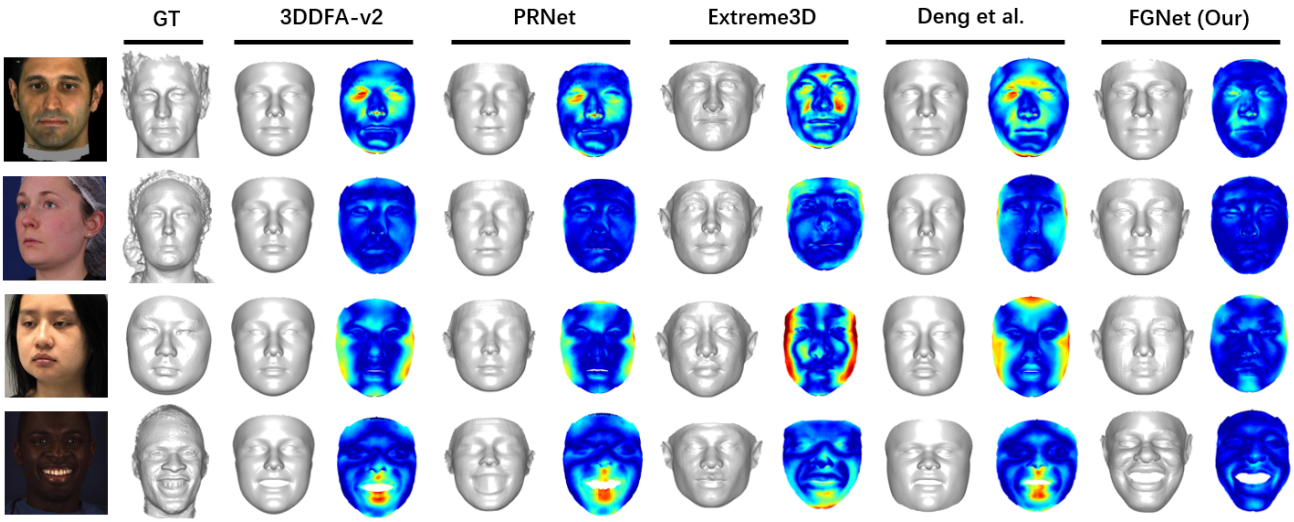


Fig. 16. Qualitative comparison of our method with the representative method of 3DMM fitting (3DDFA-v2), vertex regression (PRNet), shape from shading (Extreme3D) and analysis by synthesis (Deng et.al).

linear 3DMM due to its data-driven characteristics. 3) With the analysis-by-synthesis strategy, Deng's method ranks top among the state-of-the-art methods, demonstrating the benefits of face appearance optimization. 4) We find little difference when comparing Extreme3D with its base model (without details), indicating that the fine-level details acquired by Shape-from-Shading do not change face shapes much. 5) Our method achieves the best result, validating the feasibility of reconstructing personalized shapes in a supervised manner. In addition, we train our model 5 times and evaluate the performance variance on Florence. Compared with the variance of 0.018, the improvement of the best baseline (2.57 to 2.10) is more significant, validating the effectiveness of our method.

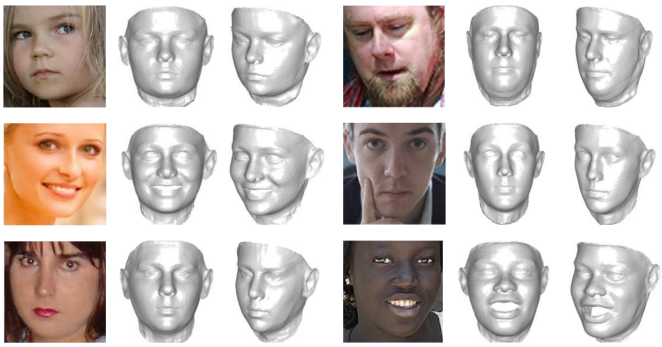


Fig. 18. Qualitative results of our method in unconstrained environment.

8.5.2 Qualitative Evaluation

We present some visual comparisons in Fig. 16 to illustrate the identifiability of the reconstructed shapes. Comparing several representative methods, including 3DMM fitting (3DDFA-v2), vertex regression (PRNet), shape from shading (Extreme3D) and analysis by synthesis (Deng's method), we can first see that the results of 3DDFA-v2 and PRNet are not discriminative enough due to the limitation of linear 3DMM. It is difficult to distinguish the 3D faces by observing the 3D geometry only. Second, Extreme3D recovers plausible geometric details by the shape-from-shading method, but their global shapes do not look identical to the corresponding person. Third, the analysis-by-synthesis method employed by Deng's method improves the visual effect by optimizing the face appearance. Finally, compared with the aforementioned methods, our proposal improves shape accuracy by three advantages: 1) straightforward training signals from the ground-truth 3D shapes, 2) a specific non-linear neural network modeling personalized shape, and 3) a visual-effect-guided loss highlighting shape errors. In the experiments, 3DDFA, 3DDFA-v2, PRNet, Extreme3D, Deng's method and FaceScape are implemented by the released codes, and the non-linear 3DMM is reproduced and trained on 300W-LP [10].

We further qualitatively evaluate the robustness to pose, illumination and occlusion in Fig. 19. The results show that our method performs well under different poses, side-light and common occlusions such as hair and eye-glasses. The robustness comes from the diverse pose and illumination variations in the training data and the sophisticated ICP registration that filters out occluded vertex matching.

We also attempt to reconstruct the high-fidelity shape in unconstrained environment, evaluated by the samples in AFLW [71]. The results in Fig. 18 indicate that, trained on the indoor-collected and 3D-augmented samples, our model generalizes well in an outdoor environment.

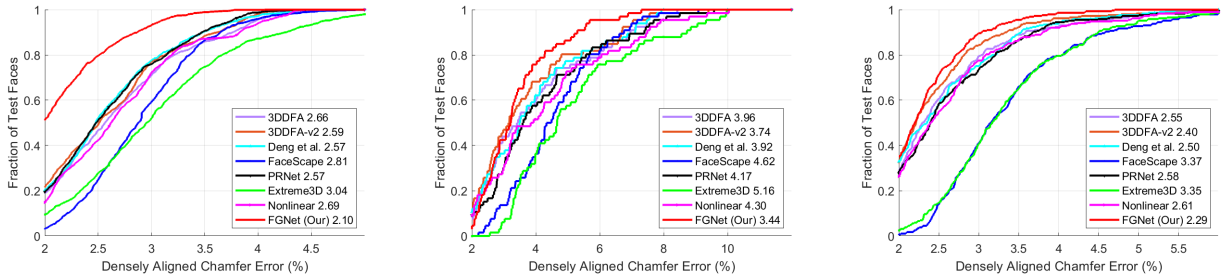
9 Conclusion

This paper proposes a complete solution for high-fidelity 3D face reconstruction, from data construction to neural network training. With the proposed Virtual Multiview Network (VMN), the input image is rendered at 5 calibrated views so that pose variations are normalized with little image information lost. Then, the features extracted from multiple views are fused and regressed to a UV displacement map by a novel many-to-one hourglass network. A

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3164131, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5
The Densely Aligned Chamfer Error (DACE) on Florence, NoW and Stirling/ESRC, evaluated by different yaw ranges.

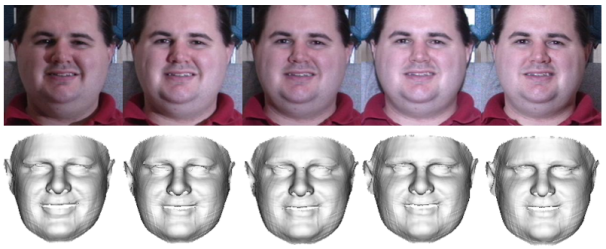
Method	Florence				NoW				Stirling/ESRC
	$[0, 15]$	$[15, 30]$	$[30, 45]$	Mean	$[0, 15]$	$[15, 30]$	$[30, 45]$	Mean	
3DDFA [10]	2.64	2.66	2.69	2.66	3.86	3.82	4.09	3.96	2.55
3DDFA-v2 [12]	2.57	2.58	2.63	2.59	3.77	3.65	3.75	3.74	2.40
Deng et al. [63]	2.48	2.59	2.67	2.57	3.66	3.57	4.22	3.92	2.50
FaceScape [15]	2.88	2.73	3.15	2.81	3.67	4.02	5.49	4.62	3.37
PRNet [9]	2.50	2.53	2.71	2.57	4.10	3.85	4.33	4.17	2.58
Extreme3D [70]	2.93	3.07	3.16	3.04	5.17	5.05	5.20	5.16	3.35
Non-linear 3DMM [17]	2.74	2.66	2.66	2.69	4.73	3.96	4.13	4.30	2.61
FGNet (Ours)	**2.12**	**2.05**	**2.15**	**2.10**	**3.45**	**3.39**	**3.45**	**3.44**	**2.29**



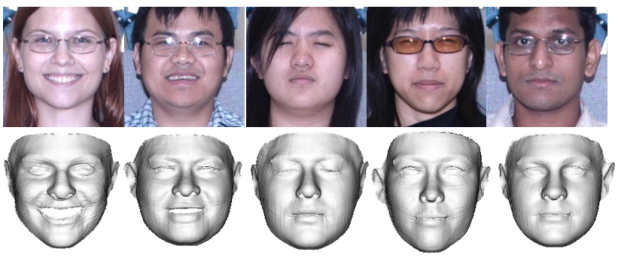
(a) Florence (b) NoW (c) Stirling/ESRC

Fig. 17. Comparisons of cumulative error distribution (CED) curves on Florence, NoW and Stirling/ESRC. The error is measured by the mean of DACE.

(a) Reconstruction across poses



(b) Reconstruction in different illumination.



(c) Reconstruction in occlusion

Fig. 19. Qualitative evaluation of the robustness to (a) pose, (b) illumination, and (c) occlusion.

novel Plaster Sculpture Descriptor (PSD) is also proposed to model the visual effect, which considers the reconstructed shape as a white plaster and measures the similarity between the multiview images rendered from the shape. Besides, to provide abundant samples for network training, we propose to register RGB-D images followed by pose and shape augmentation. Extensive experimental results substantiate the state-of-the-art performance of our proposal on several challenging datasets.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research & Development Program (No. 2020AAA0140002), Chinese National Natural Science Foundation Projects #62176256, #61876178, #61976229, #62106264, the Youth Innovation Promotion Association CAS (#Y2021131).

REFERENCES

[1] B. Echeagaray-Patron, V. Kober, V. Karnaukhov, and V. Kuznetsov, "A method of face recognition using 3d facial surfaces," *Journal of Communications Technology and Electronics*, vol. 62, no. 6, pp. 648–652, 2017.

[2] F. Liu, Q. Zhao, X. Liu, and D. Zeng, "Joint face alignment and 3d face reconstruction with application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 664–678, 2020.

[3] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5541–5550.

[4] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9821–9830.

[5] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and vision Computing*, vol. 30, no. 10, pp. 738–749, 2012.

[6] H. Bejaoui, H. Ghazouani, and W. Barhoumi, "Fully automated facial expression recognition using 3d morphable model and mesh-local binary pattern," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 39–50.

[7] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[8] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 101–10 111.

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

[10] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 78–92, 2019.

[11] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[13] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*. IEEE, 2009, pp. 296–301.

[14] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

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

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

[17] L. Tran and X. Liu, "On learning 3d face morphable model from in-the-wild images," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[18] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1097–1106.

[19] S. Cheng, G. Tzimiropoulos, J. Shen, and M. Pantic, "Faster, better and more detailed: 3d face reconstruction with graph convolutional networks," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[20] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[21] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3d reconstruction networks learn?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3405–3414.

[22] J. Jin, A. G. Patil, Z. Xiong, and H. Zhang, "Dr-kfs: A differentiable visual similarity metric for 3d shape reconstruction," in *European Conference on Computer Vision*. Springer, 2020, pp. 295–311.

[23] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, "3d human shape and pose from a single low-resolution image with self-supervised learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 284–300.

[24] D. H. C. Y. H. W. J. G. Z. L. S. Z. L. Xiangyu Zhu, Fan Yang, "Beyond 3dmm space: Towards fine-grained 3d face reconstruction,," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 343–358.

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

[26] A. Bas, P. Huber, W. A. Smith, M. Awais, and J. Kittler, "3d morphable models as spatial transformer networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 904–912.

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

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

[29] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5163–5172.

[30] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides, "Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3980–3989.

[31] T. Hassner, "Viewing real-world faces in 3d," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3607–3614.

[32] I. Kemelmacher-Shlizerman and R. Basri, "3d face reconstruction from a single image using a single reference face shape," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 394–405, 2011.

[33] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense face alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1619–1628.

[34] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1576–1585.

[35] Z. Sánta and Z. Kato, "3d face alignment without correspondences," in *European Conference on Computer Vision*. Springer, 2016, pp. 521–535.

[36] Z. Gao, J. Zhang, Y. Guo, C. Ma, G. Zhai, and X. Yang, "Semi-supervised 3d face representation learning from unconstrained photo collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 348–349.

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

[38] P. Wang, C. Lin, B. Xu, W. Che, and Q. Wang, "Low-frequency guided self-supervised learning for high-fidelity 3d face reconstruction in the wild," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[39] X. Zeng, X. Peng, and Y. Qiao, "Df2net: A dense-fine-finer network for detailed 3d face reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2315–2324.

[40] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu, "3d face reconstruction with geometry details from a single image," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4756–4770, 2018.

[41] Y. Li, L. Ma, H. Fan, and K. Mitchell, "Feature-preserving detailed 3d face reconstruction from a single image," in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018, pp. 1–9.

[42] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu, "Photo-realistic facial details synthesis from single image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9429–9439.

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

[44] D. Cosker, E. Krumhuber, and A. Hilton, "A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2296–2303.

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

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

[47] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan, "Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[48] X. Lin, Y. Chen, L. Bao, H. Zhang, S. Wang, X. Zhe, X. Jiang, J. Wang, D. Yu, and Z. Zhang, "High-fidelity 3d digital human creation from rgb-d selfies," *arXiv preprint arXiv:2010.05562*, 2020.

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

[50] S. Cheng, P. Ma, G. Tzimiropoulos, S. Petridis, A. Bulat, J. Shen, and M. Pantic, "Towards pose-invariant lip-reading," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4357–4361.

[51] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li, "Avatar digitization from a single image for real-time rendering," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 6, pp. 1–14, 2017.

[52] K. Wang and Q. Ji, "Real time eye gaze tracking with 3d deformable eye-face model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1003–1011.

[53] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3d model retrieval," in *Computer graphics forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 223–232.

[54] J. Johnson, N. Ravi, J. Reizenstein, D. Novotny, S. Tulsiani, C. Lassner, and S. Branson, "Accelerating 3d deep learning with pytorch3d," in *SIGGRAPH Asia 2020 Courses*, 2019, pp. 1–1.

[55] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

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

[57] "Casia-3d facev1," http://biometrics.idealtest.org/, 2004.

[58] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

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

[60] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2d/3d hybrid face dataset," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM, 2011, pp. 79–80.

[61] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[62] http://pics.psych.stir.ac.uk/ESRC/index.htm.

[63] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[64] X. Z. Jianzhu Guo and Z. Lei, "3ddfa," https://github.com/cleardusk/3DDFA, 2018.

[65] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Prnet," https://github.com/YadiraF/PRNet/blob/master/predictor.py, 2018.

[66] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 937–944.

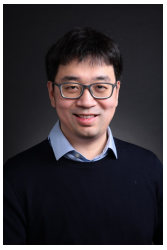
[67] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4295–4304.

[68] B. Chu, S. Romdhani, and L. Chen, "3d-aided face recognition robust to expression and pose variations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1899–1906.

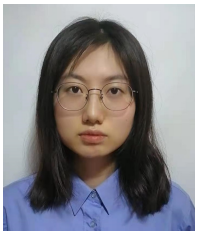
[69] I. Masi, A. T. Tran, T. Hassner, G. Sahin, and G. Medioni, "Face-specific data augmentation for unconstrained face recognition," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 642–667, 2019.

[70] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni, "Extreme 3d face reconstruction: Seeing through occlusions." in *CVPR*, 2018, pp. 3935–3944.

[71] M. Köstinger, 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 *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.



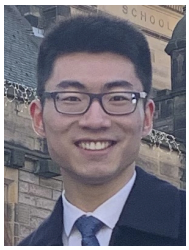
Xiangyu Zhu received the BS degree in Sichuan University (SCU) in 2012, and the PhD degree from Institute of Automation, Chinese Academy of Sciences, in 2017, where he is currently an associate professor. His research interests include pattern recognition and computer vision, in particular, image processing, 3D face model, face alignment and face recognition.



Chang Yu received the BS degree in Xi'an Jiaotong University (XJTU) in 2019. She is working toward the PhD degree in Institute of Automation, Chinese Academy of Sciences. Her research interest includes computer vision, pattern recognition and 3D face reconstruction.

Di Huang received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the École centrale de Lyon, Lyon, France, in 2005, 2008, and 2011, respectively. He joined the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, as a Faculty Member. He is currently a Professor with the research interests on biometrics, in particular, on 2D/3D face analysis, image/video processing, and pattern recognition.

Zhen Lei received the BS degree in automation from the University of Science and Technology of China, in 2005, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences,in 2010, where he is currently a professor. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular.



Hao Wang received his B.Eng from Beijing University of Posts and Telecommunications, China, in 2018, and MSc from The University of Edinburgh, the UK, in 2019. He is currently a research intern at the Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, image processing, human face analysis, biometrics, and 3D shape modeling.

Stan Z. Li received his B.Eng from Hunan University, China, M.Eng from National University of Defense Technology, China, and PhD degree from Surrey University, UK. He is currently a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He worked at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore. He was elevated to IEEE Fellow for his contributions to the fields of face recognition, pattern recognition and computer vision. His research interest includes pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance.