Beyond 3DMM Space: Towards Fine-grained 3D Face Reconstruction

Xiangyu Zhu[1,2], Fan Yang[3], Di Huang[4], Chang Yu[1,2], Hao Wang[1], Jianzhu Guo[1,2], Zhen Lei[1,2⋆], and Stan Z. Li[5]

[1] CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] College of Software, Beihang University
[4] Beijing Advanced Innovation Center for BDBC, Beihang University
[5] School of Engineering, Westlake University
{xiangyu.zhu, jianzhu.guo, zlei, szli, chang.yu}@nlpr.ia.ac.cn,
{fanyang, dhuang}@buaa.edu.cn, haowang7308@gmail.com

Abstract. Recently, deep learning based 3D face reconstruction methods have shown promising results in both quality and efficiency. However, most of their training data is constructed by 3D Morphable Model, whose space spanned is only a small part of the shape space. As a result, the reconstruction results lose the fine-grained geometry and look different from real faces. To alleviate this issue, we first propose a solution to construct large-scale fine-grained 3D data from RGB-D images, which are expected to be massively collected as the proceeding of hand-held depth camera. A new dataset Fine-Grained 3D face (FG3D) with 200k samples is constructed to provide sufficient data for neural network training. Secondly, we propose a Fine-Grained reconstruction Network (FGNet) that can concentrate on shape modification by warping the network input and output to the UV space. Through FG3D and FGNet, we successfully generate reconstruction results with fine-grained geometry. The experiments on several benchmarks validate the effectiveness of our method compared to several baselines and other state-of-the-art methods. The proposed method and code will be available at https://to-be-released.

Keywords: 3D Face Reconstruction, Fine-grained, Deep Learning

1 Introduction

With the advent of deep learning and the development of large annotated datasets, recent works have shown results of unprecedented accuracy even on the most challenging computer vision tasks. In this work, we focus on 3D face reconstruction which recovers the 3D facial geometry from a single 2D image. Despite many years of research, it is still an open problem in vision and graphics research. Since the seminal work of Blanz and Vetter [5], 3D Morphable Model (3DMM) has been widely used to reconstruct 3D face shape. However, most of the popular

⋆ Corresponding author.

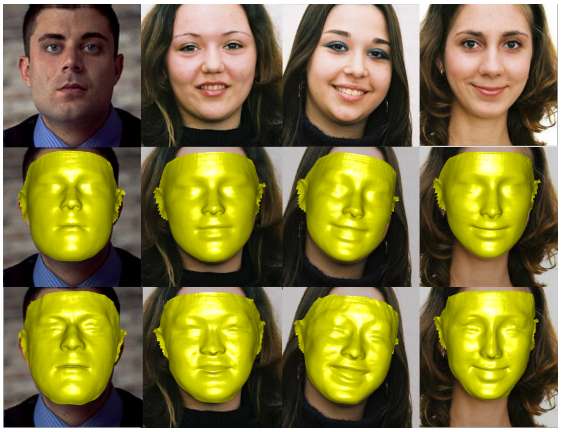

Fig. 1. The first row shows the images, the second row shows the results of the state-of-the-art method PRNet [9] and the third row shows our results.

models like BFM [21] are built from scans of only 200 subjects with a similar ethnicity/age group. They are also captured in well controlled conditions with only neutral expressions. As a result, these models are fragile to large variances in face identity. In more than a decade, almost all the models cover no more than 300 training scans. Such a small training set is far from adequate to describe the full variability of human faces. Recently, there is a surge of interest in 3D face reconstruction using deep Convolution Neural Networks (CNN) rather than the optimization based traditional methods [5,26,25,42]. However, training deep models requires large data with dense 3D annotations, which are expensive and even infeasible in some cases. In most 3D face datasets, the ground truth is constructed by fitting a 3DMM to less than 100 labelled landmarks, which loses the fine-grained geometry, especially on the cheek region. A model trained on such a dataset cannot deal well with the variations that are not present in the 3DMM space. Although recent works bypass 3DMM parameters and use the image-to-volume [11] or image-to-uvmap [9] strategy, the ground truth still comes from the space of 3DMM and the fitting results are still model-like.

In this paper, we aim to overcome the intrinsic limitation of 3D face reconstruction by improving both the training data and the reconstruction method. Firstly, we explore to construct large-scale fine-grained 3D data from RGB-D images. Although complete and high-precision face scans are expensive to acquire, the RGB-D images can be considered as a good alternative, which are much easier to collect and have been popular in face analysis [36,36,29,38,20,6,17]. As the proceeding of hand-held depth camera, we believe medium-precision RGB-D images can be massively collected in the near future. In this paper, we first employ the 3DMM texture and illumination model as a strong constraint to robustly register RGB-D images and perform high-fidelity out-of-plane augmentation, generating a large 3D dataset **F**ine-**G**rained **3D** face (FG3D) from

public RGB-D images. Secondly, to reconstruct fine-grained geometry through CNN, we propose a **F**ine-**G**rained reconstruction **Net**work (FGNet) and discuss two possible structures to capture fine-grained shapes: a camera-view structure (FGNet-CV) which directly estimates the shape update from the original image and a model-view structure (FGNet-MV) which normalizes pose variations by UV-space warping to concentrate on shape modification. The two structures are compared experimentally and the better one is adopted for reconstruction.

In summary, our main contributions are: (1) In order to overcome the scarcity of 3D fine-grained training data, we develop a complete solution to generate a large number of "image to 3D face" pairs from RGB-D images. (2) We provide a new fine-grained 3D face dataset FG3D with about 200k samples for neural network training. (3) A novel network structure FGNet is proposed for fine-grained geometry reconstruction. (4) Based on FG3D and FGNet, we finally generate face-like 3D reconstruction results. Extensive experiments show that our method significantly reduces the reconstruction error and achieves the best result. Fig. 1 briefly shows some results.

2 Related Works

With the development of deep learning, 3D face reconstruction has witnessed great progress by Convolution Neural Network (CNN). In early years, some methods use CNN to estimate the 3D Morphable Model parameters [14,18,23,24] or its variants [3,8,13,31,35,4], which provide both dense face alignment and 3D face reconstruction results. However, the performance of these methods is restricted due to the limitation of the 3D space defined by the face model basis or the templates [8,10,15,19,28,30]. The required face transformations including perspective projection and 3D thin plate spline transformation are also difficult to estimate. Recently, two end-to-end works [11,9], which bypass the limitation of the PCA model, achieve state-of-the-art performance on their respective tasks. VRN [11] develops a volumetric representation of 3D face and uses a network to regress it from a 2D image. However, this representation discards the semantic meaning of points and the network needs to regress the redundant whole volume in order to restore the face shape. PRNet [9] designs a UV position map, which is a 2D image recording the 3D coordinates of a complete facial point cloud, while at the same time keeps the semantic meaning at each UV place. PRNet uses an encoder-decoder network to regress the UV position map from a single 2D facial image. Although these methods have broken through the limitations of 3DMM, their training sets are still restricted by 3DMM and the reconstruction results are still model-like. Tran et al. [33] achieve a certain breakthrough by utilizing two CNN decoders, instead of two PCA spaces, to learn a nonlinear model from unlabelled images in an weakly-supervised manner. However, the model still needs to be pre-trained on 3DMM data and the learned bilinear model does not go far beyond 3DMM space, making the results still lack fine-grained geometry information. Different from the above methods, our solution can directly obtain fine-grained 3D faces and keep the semantics of vertices.

2.1 3D Morphable Model

The seminal work of Blanz et al. [5] proposes the 3D Morphable Model (3DMM) to describe 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. The 3D face can be rigidly transformed by:

$$V(\mathbf{p}_{3d}) = f * \mathbf{R} * (\overline{\mathbf{S}} + \mathbf{A}_{id}\boldsymbol{\alpha}_{id} + \mathbf{A}_{exp}\boldsymbol{\alpha}_{exp}) + \mathbf{t}_{3d}, \tag{2}$$

where $V(\mathbf{p}_{3d})$ is the model construction and rigid transformation function, f is the scale factor, $\mathbf{R}$ is the rotation matrix constructed from Euler angles *pitch*, *yaw*, *roll* and $\mathbf{t}_{3d}$ is the translation vector. The collection of 3D geometry parameters is $\mathbf{p}_{3d} = [f, \mathbf{R}, \mathbf{t}_{3d}, \boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}]$.

3 Fine-grained 3D Data Construction

One of the main challenges of fine-grained 3D face reconstruction is the scarcity of training data. However, it is very tedious to acquire complete and high-precision 3D faces. The raw scans must be captured in well controlled conditions and registered to a face template through laborious hand labeling. Differently, RGB-D images are much easier to capture and also contain rich 3D information. In this work, we explore to construct a large 3D face dataset from public RGB-D images.

3.1 Texture Constrained Non-rigid ICP

The first task is registering all the depth images to a template face to get the topology-uniformed shape. Previous methods adopt the Iterative Closest Point (ICP) method [1] for registration. However, most of depth images are collected in semi-controlled environment, suffering from holes, spikes, occlusions and large missing regions due to self-occlusion. Hand labelling such as dense 3D landmarks is needed for robust registration. To improve the robustness of ICP on human faces, we propose to utilize the face texture, from both the RGB-D image and the face model, as a strong constraint in closest point matching. Fig. 2 shows the overview of our method.

Firstly, we fit a 3DMM with the detected 240 landmarks [40] to get the initial 3D face $\mathbf{V} = \{\mathbf{v}_i | i = 1, 2, \cdots, N\}$ (Fig. 2(c)), which is defined in Eqn. 2. Secondly, we construct the face texture as a template during ICP registration. Specifically, we utilize the PCA raw texture model from BFM [21]:

$$\mathbf{T} = \overline{\mathbf{T}} + \mathbf{B}\boldsymbol{\beta}, \tag{3}$$

where $\overline{\mathbf{T}}$ is the mean texture, $\mathbf{B}$ is the principle axes of the raw texture and $\boldsymbol{\beta}$ is the raw texture parameter. Given 3D vertices $\mathbf{V}$ and its raw texture $\mathbf{T}$, the Phong illumination model is used to produce the final face texture [5]:

$$C_i(\mathbf{p}_{tex}) = \mathbf{Amb} * \mathbf{T}_i + \mathbf{Dir} * \mathbf{T}_i * \langle \mathbf{n}_i, \mathbf{l} \rangle + k_s \cdot \mathbf{Dir} \langle \mathbf{r}_i, \mathbf{ve} \rangle^{\nu}, \tag{4}$$

where C_i is the RGB color 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 $\mathbf{p}_{tex} = [\boldsymbol{\beta}, \mathbf{Amb}, \mathbf{Dir}, \mathbf{l}, k_s, \nu]$. We fit the illumination model by optimizing Eqn. 5 through the Levernberg-Marquardt method:

$$\arg \min_{\mathbf{p}_{tex}} \|\mathbf{Img}(\mathbf{V}) - C(\mathbf{p}_{tex})\|, \tag{5}$$

where $\mathbf{Img}(\mathbf{V})$ is the image pixels at vertex positions. The optimized result $\mathbf{C} = \{\mathbf{c}_i | i = 1, 2, \cdots, N\} = C(\mathbf{p}_{tex})$ is the face texture, shown in Fig. 2(d).

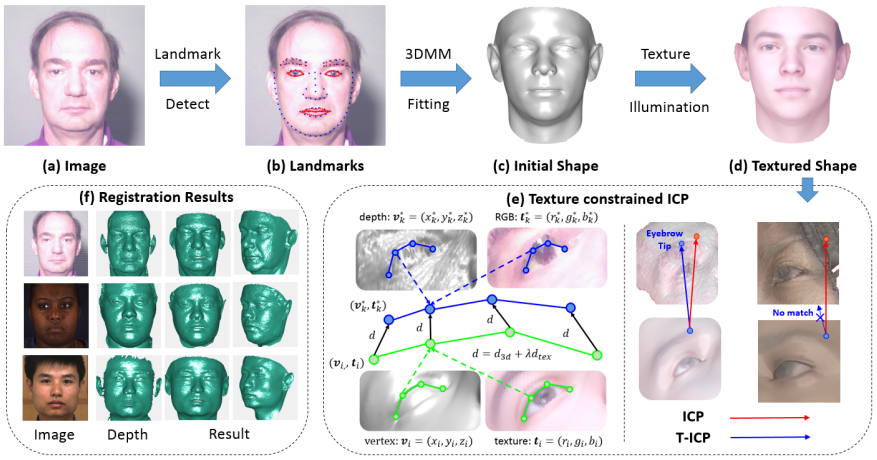

Fig. 2. The overview of 3D face registration. (a) The input image. (b) The detected 240 landmarks. (c) The fitted 3D shape by landmarks. (d) The reconstructed face texture by the texture and illumination model. (e) Left: the T-ICP searches the closest point in both 3D space (x, y, z) and color space (r, g, b). Right: the incorporation of texture improves the robustness on eye-brows and occluded regions. (f) The final registration results.

Thirdly, to register the initial shape to the depth image, we propose a **T**exture constrained Nonrigid-**ICP** (**T-ICP**) method to find the vertex correspondence

based on both geometry and texture. Suppose the target RGB-D image has the vertices $\mathbf{V}^* = \{\mathbf{v}_k^* | k = 1, 2, \cdots, K\}$ and their corresponding pixels $\mathbf{C}^* = \{\mathbf{c}_k^* | k = 1, 2, \cdots, K\}$. For each vertex $\mathbf{v}_i$ on the initial shape, its closest point $\mathbf{v}_{k_{corr}}^*$ is searched by:

$$k_{corr} = \arg\min_k (\|\mathbf{v}_i - \mathbf{v}_k^*\| + \lambda_{tex} \|\mathbf{c}_i - \mathbf{c}_k^*\|) \quad \text{if} \quad \|\mathbf{v}_i - \mathbf{v}_k^*\| < \tau_v \quad \text{and} \quad \|\mathbf{c}_i - \mathbf{c}_k^*\| < \tau_c \tag{6}$$

where τ_v and τ_c are the distance thresholds in 3D space and color space, respectively. As shown in Fig. 2(e), by incorporating the texture constraint, we improve the robustness not only on the geometry-smooth but texture-rich surfaces like eye-brows, but also on the occluded regions where the matching is filtered out by τ_c due to large texture error. With the correspondence $(\mathbf{v}_i, \mathbf{v}_{k_{corr}}^*)$, we perform Optimal Non-rigid ICP [1] to finish the registration, shown in Fig. 2(f). Different from the texture constraint used in scan-to-scan registration [27] where both scans have the texture of the same object, our task is a more challenging model-to-scan registration, where the facial template only has a texture model rather than the real texture. During registration, we must iteratively update the texture parameters and get a more reliable texture constraint.

Finally, we disentangle rigid and non-rigid transformations, getting the ground-truth shape by optimizing the following equation:

$$\mathbf{S}_{morph}^* = \arg \min_{\mathbf{S}_{morph}, \mathbf{R}, f, \mathbf{t3d}} \|\mathbf{V}_{regist} - f * \mathbf{R} * (\overline{\mathbf{S}} + \mathbf{S}_{morph}) + \mathbf{t}_{3d}\| \tag{7}$$

where $\mathbf{V}_{regist}$ is the registered 3D face, $(f, \mathbf{R}, \mathbf{t3d})$ are the rigid transformation parameters, $\overline{\mathbf{S}}$ is the mean shape and $\mathbf{S}_{morph}^*$ is the difference between the target shape and the mean shape, which will be the target of the neural network learning.

3.2 Out-of-plane Pose Augmentation

Large scale data is crucial for training neural networks. However, there are less than ten thousand public RGB-D samples [16,37,22] and most of them are frontal faces, leading to poor generalization across poses. To address this challenge, we improve the face profiling method [41] for RGB-D data and synthesize hundreds of thousands high-fidelity 3D data for network training. Firstly, we complete the depth channel for the whole image space, where the depth on the face region directly comes from the registered 3D face and the depth on the background is coarsely estimated by some anchors (x_i, y_i), shown in Fig. 3(b). These anchors are triangulated to a background mesh and their depth values d_i are estimated by depth constraints and smoothness constraints, as in Eqn. 8:

$$\sum_i Mask(x_i, y_i) \|d_i - Depth(x_i, y_i)\| + \sum_i \sum_j Connect(i, j) \|d_i - d_j\|, \tag{8}$$

where $Depth(x, y)$ is the depth channel of the RGB-D image, $Mask(x, y)$ indicates whether (x, y) is hollow and $Connect(i, j)$ is whether two anchors are

connected by the background mesh. By turning the whole image to a 3D mesh (Fig. 3(c)), we can out-of-rotate it (Fig. 3(d)) and render it (Fig. 3(e)) in any views.

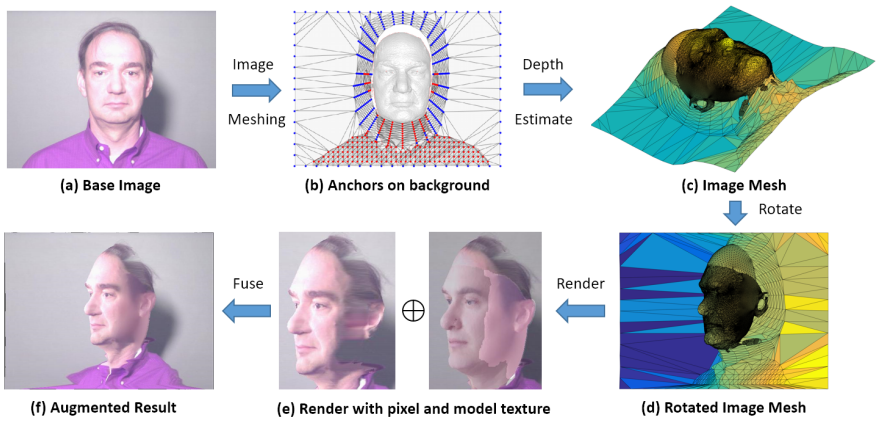

Fig. 3. The overview of out-of-plane pose augmentation. (a) The base image. (b) The original depth image and the anchors on the background. Note that the red anchors locate on the scan and the blue ones locate on the hollow, they have different constraints in Eqn. 8. (c) The complete depth of the base image. (d) The rotated 3D mesh of the image. (e) Rendering with the image pixels and the model texture. (f) The augmentation result.

Different from the original face profiling method which aims to generate large poses from medium poses, most of our base images are frontal faces. While a main drawback of face profiling is that when rotating from frontal faces, there are serious artifacts on the side face due to the lack of texture in the original image, shown in Fig. 3(e). In this work, the texture and illumination model is also used as a strong prior to refine the artifacts. With the model texture used in T-ICP (shown in Fig. 2(d)), we render the 3D image mesh with both the image pixels and the model texture, shown in Fig. 3(e). Then we detect the invisible region with the normal directions, and inpaint it with the model texture through Poisson editing, shown in Fig. 3(f). Since the side face is not texture-rich, the model texture is realistic enough to inpaint it and we finally get high-fidelity synthetic samples. In this work, we augment the images by enlarging the *yaw* angle at the step of 15° until 90°, and randomly enlarging the *pitch* angle within ±25°, generating about 200k training samples.

4 Fine-grained Reconstruction Network

With large scale training data, we are prepared to train a **F**ine-**G**rained reconstruction **Net**work (**FGNet**) to reconstruct the fine-grained geometry. In order to concentrate on face shape modification, we employ a state-of-the-art 3DMM fitting method [12] to get the rigid transformation and an initial 3D shape. Our task can be formulated as follows, given the input image **Img**, the rigid transformation $V(\cdot)$ and the initial shape $\mathbf{S}_{init}$, we aim to estimate the shape update $\Delta\mathbf{S}$ so that the final shape is closer to the ground truth $\overline{\mathbf{S}} + \mathbf{S}_{morph}$ (defined in Eqn. 7) after updating:

$$\arg\min_{\theta} \|Net(\mathbf{Img}, V(\mathbf{S}_{init}); \theta) - (\overline{\mathbf{S}} + \mathbf{S}^*_{morph} - \mathbf{S}_{init})\| \tag{9}$$

where $Net(\cdot)$ is a convolutional neural network and θ is the network parameters. To implement Eqn. 9, we should formulate the input $(\mathbf{Img}, V(\mathbf{S}_{init}))$ as a 2D map to be convolved by CNN and decide the formulation of the regression target $\overline{\mathbf{S}} + \mathbf{S}^*_{morph} - \mathbf{S}_{init}$. In this work, we discuss two structures: **camera-view (FGNet-CV)** and **model-view (FGNet-MV)** and compare them in the experiments.

4.1 Camera-view Structure

In the first structure, the original image **Img** is directly sent to CNN and the Projected Normalized Coordinate Code (PNCC) [41] is employed to encode the initial fitting result $V(\mathbf{S}_{init})$ as a 2D map for CNN. It is called camera-view since the network observes the face in the same view of the camera. Besides, the shape update $\Delta\mathbf{S} = \overline{\mathbf{S}} + \mathbf{S}^*_{morph} - \mathbf{S}_{init}$ is represented as a UV map [9] and is regressed by a fully convolutional encoder-decoder network. The overview of the structure is in Fig. 4.

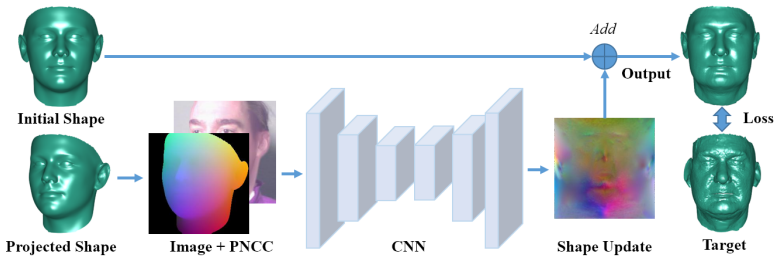

Fig. 4. The Camera-view Fine-grained Reconstruction Network.

The advantage of the camera-view structure is that it does not miss any information provided by the image. However, the structure requires the network to identify small shape variations in any poses, which is hard to learn. Besides,

the input and output have different coordinate systems (image space to UV space). For each coordinate on the output, its receptive field may not cover the most related region for reconstruction.

4.2 Model-view Structure

Different from 3DMM fitting where pose estimation is the most important [41], the purpose of fine-grained reconstruction is to modify the shape. To this end, we design a model-view structure to concentrate on shape information. The overview is shown in Fig. 5.

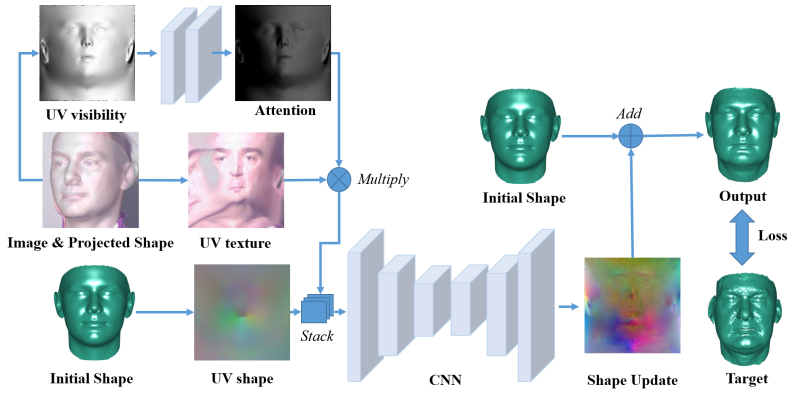

Fig. 5. The Model-view Fine-grained Reconstruction Network.

The input has three parts: we extract the **UV-texture** map according to the projected initial shape $V(\mathbf{S}_{init})$. Given that the UV-texture of a non-frontal face has invalid regions due to self-occlusion, we also construct the **UV-visibility** map which stores the z values of the vertex normals. Besides, a **UV-shape** map which stores the vertex positions of the initial shape $\mathbf{S}_{init}$ is also provided. During inference, the visibility map is first convolved by several layers to an attention map. The UV-texture is then multiplied by the attention, concatenated by the UV-shape, and sent to the backbone to regress the shape update.

The structure is called model-view since the input and the output are both in UV space and the network observes the face through the model vertices. The advantage of the structure is that each 2D position across the network has the same semantic meaning, so that the receptive field of each output coordinate always covers the most related region. Besides, by warping the image pixels to a UV map, the pose variations are implicitly normalized, making the CNN concentrate on shape updating.

5 Experiments

5.1 Datasets

To perform fine-grained reconstruction, multiple datasets listed below are used for training and evaluation in our experiments.

FG3D are constructed from three datasets. The FRGC [22] includes $4,950$ samples and each sample has a face image and a 3D scan with pixels in full correspondence. BP4D [37] contains 328 2D+3D videos from 41 subjects, where $3,376$ frames are randomly selected. CASIA-3D [16] consists of $4,624$ scans of 123 persons and we filter out the non-frontal faces. We register and out-of-plane augment the three datasets, generating a large 3D dataset **FG3D** with $212,579$ samples. Among FG3D, 90% subjects are used as the training set **FG3D-train** and the rest 10% subjects are the testing set **FG3D-test**. Besides, we manually delete the bad registration results in FG3D-test for better evaluation.

Florence [2] is a 3D face dataset containing 53 subjects with its 3D mesh acquired from a structured-light scanning system. In the experiments, each subject is rendered at pitches of $-20°$, $0°$, $20°$ and yaws from $-90°$ to $90°$ at the step of $15°$. Besides, we register each 3D mesh with hand-labelled landmarks and carefully check the registration results. This dataset set is used for cross dataset evaluation to demonstrate the generalization. The registration results are shown in the supplemental materials.

5.2 Implementation Details

The 3DDFA [12] is used to provide the rigid transformation and an initial 3D shape for FGNet. The architecture of FGNet is a fully convolutional encoder-decoder network the same as [9]. The models are trained by the SGD optimizer and L1-Loss with a start learning rate of 0.1, which is decayed by 0.1 at epoch 20, 30 and 40, and the model is trained for 50 epochs. The training images are cropped by the bounding boxes of the initial 3DMM fitting results [12] and resized to 256×256 without any perturbation. All of the UV maps in the network are also 256×256. As for the T-ICP, the λ_{tex}, τ_c and τ_t in Eqn. 6 are set to 0.013, 0.17 and $2 * 10^{-4}$ times of the interocular distance of the 3D face, respectively.

To evaluate the reconstruction accuracy, we **rigid-align** the result to the ground truth and employ the Normalized Mean Error (NME):

$$NME = \frac{1}{N} \sum_{k=1}^{N} \frac{||\mathbf{v}_k - \mathbf{v}_k^*||}{d}, \tag{10}$$

where $k = 1, 2, \cdots, N$ are the vertices on the face region without neck and ears (the face region is given by [9]), $\mathbf{v}_k$ and $\mathbf{v}_k^*$ are vertices of the reconstructed face and the ground truth and d is the outer interocular distance of 3D coordinates. Note that this NME mainly evaluates the shape error since the pose error is normalized by the rigid alignment.

5.3 Ablation Studies

Camera View vs. Model View We first discuss the proposed two structures for fine-grained reconstruction: the Camera View (FGNet-CV) and the Model View (FGNet-MV). The FGNet-CV sends the original image to the CNN without losing any image information. While the FGNet-MV warps the image to the UV space and normalizes the pose variations implicitly. Their performance is compared in Table 1.

Table 1. The NME(%) results on FG3D-test with different structures, evaluated by different yaw ranges. The "Initial" indicates the initial 3DMM fitting result by [12].

Method	[30, 60]	[30, 60]	[60, 90]	All
Initial	5.05	5.03	5.14	5.07
FGNet-CV	3.44	3.11	3.12	3.23
FGNet-MV	3.30	3.06	2.95	3.10

Compared with the initial fitting results, both FG3D-CV and FG3D-MV greatly improve the reconstruction accuracy. Among them, FG3D-MV achieves better results by concentrating on shape modification.

Ablation Studies on Input To perform fine-grained reconstruction, we formulate several inputs to provide the face appearance and the initial fitting result for CNN. FG3D-CV has the original image and the PNCC. FG3D-MV has the UV-texture, UV-visibility and UV-shape. In this part, we analyze the effectiveness of each input, shown in Table 2.

Table 2. The NME(%) results on FG3D-test with different inputs, evaluated by different yaw ranges. UV-tex and UV-vis are short for UV-texture and UV-visibility, respectively.

Input	[30, 60]	[30, 60]	[60, 90]	All
FGNet-CV(img)	3.54	3.30	3.26	3.37
FGNet-CV(img + PNCC)	3.44	3.11	3.12	3.23
FGNet-MV(UV-tex)	3.47	3.18	3.14	3.27
FGNet-MV(UV-tex + UV-vis)	3.39	3.14	3.04	3.19
FGNet-MV(UV-tex + UV-vis + UV-shape)	3.30	3.06	2.95	3.10

In FG3D-CV, PNCC effectively improves the performance by providing the initial fitting result for the CNN and simplifying the reconstruction task. In FG3D-MV, the attention map from the UV-visibility shrinks the self-occluded

region of the UV-texture and reduces the shape error. The incorporation of UV-shape further provides the initial shape and gets better results. The combination of UV-texture, UV-visibility and UV-shape achieves the best result, which is used to represent FGNet in comparison experiments. Besides, we provide more visualizations about the attention map from the UV-visibility in the supplemental materials, illustrating the learned knowledge from this map.

Error Reduction Parallel and Orthogonal to Viewing Direction Intuitively, the shape information orthogonal to the viewing direction is easy to observe, but it is not the case for the shape parallel to the viewing direction. For example, given a frontal face, we can easily know its width and height but have to guess its thickness (such as the height of the nose bridge). We are interested in that the error in which direction is reduced by our method. Given the estimated face rotation matrix $\mathbf{R}$, the viewing direction can be set as $\mathbf{ve} = \mathbf{R} * [0, 0, 1]^T$, then the errors parallel and orthogonal to the viewing direction are:

$$E_{3d} = \| \mathbf{v} - \mathbf{v}^* \|, \quad E_{pal} = \| (\mathbf{v} - \mathbf{v}^*) \cdot \mathbf{ve} \|, \quad E_{orh} = \sqrt{E_{3d}^2 - E_{pal}^2}, \tag{11}$$

where $\mathbf{v}$ and $\mathbf{v}^*$ are the vertices of the predicted and the ground-truth shapes (shape is always a frontal face in a normalized space), respectively, E_{3d} is the original error measured by Euclidean distance, E_{pal} and E_{orh} are the errors parallel and orthogonal to the viewing direction, respectively. Based on the three types of error, we evaluate the NMEs in Table 3 and find that the error is mainly reduced parallel to the viewing direction, demonstrating that the depth is better recovered. The reason may be that the training set FG3D is constructed from RGB-D images and provides more accurate depth information than the landmark based datasets like 300W-LP [39].

Table 3. The NME(%) parallel and orthogonal to the viewing direction, evaluated on all the samples of FG3D-test.

Input	Orthogonal	Parallel	Euclidean
Initial	3.44	3.11	5.07
FGNet-CV	2.47(28.20% ↓)	1.67(46.30% ↓)	3.23(36.29% ↓)
FGNet-MV	2.38(30.81% ↓)	1.59(48.87% ↓)	3.10(38.86% ↓)

5.4 Comparison Experiments

Qualitative Comparison We present some visual comparisons to illustrate the identifiability of the reconstructed shapes. Baseline methods include the common used 3DDFA [41] and PRNet [9] which are trained on the landmark based datasets 300W-LP [39], Extreme3D [32] which reconstructs facial details

by shape-from-shading, and the released Deng's method [7] as a typical weakly-supervised method [33,34,7] which adaptively learns a nonlinear 3DMM and its fitting strategy from unlabelled images. As shown in Fig. 6, compared with other baselines, our results look more like real scans than blend models due to its better reconstructed shapes. In Extreme3D, even though plausible details are added by shape-from-shading, the face shapes are not modified. Deng's method [7] accurately reconstructs the facial features, but the cheek geometry is not well captured such as the cheekbone and the face silhouette.

Quantitative Comparison In this part, we firstly compare our method with the state-of-the-art methods including 3DDFA [41], PRNet [9], Extreme3D [32] and Deng's method [7] quantitatively on the FG3D-test. All their inputs are cropped by the ground-truth bounding boxes and only the face region is used for calculating NME. Since these methods share the topology of BFM [21], their results are comparable. As shown in Table 4, our method achieves the best result and outperforms the best of the state-of-the-art methods by 38.86%.

Table 4. The NME(%) on FG3D-test, evaluated by different yaw ranges. The FGNet employs the FGNet-MV structure.

Input	$[30, 60]$	$[30, 60]$	$[60, 90]$	All
3DDFA [41]	5.05	5.03	5.14	5.07
PRNet [9]	5.49	5.89	5.70	5.68
Extreme3D [32]	7.07	7.42	8.03	7.52
Deng et al. [7]	5.26	5.16	5.30	5.24
FGNet	3.30	3.06	2.95	3.10

Considering that FG3D-test shares the same environment with FG3D-train, we also perform cross dataset evaluation on the Florence dataset for fair comparison, shown in Table 5. First, Deng's method [7] performs better than the 300W-LP trained 3DDFA and PRNet, demonstrating that the weakly learned non-linear 3DMM [7,33,34] covers more shape variations than BFM. Second, our method achieves the best result, validating the feasibility of reconstructing fine-grained geometry in a supervised manner.

6 Conclusion

This paper proposes an solution to reconstruct 3D fine-grained face shape, from data construction to neural network training. Firstly, to prepare sufficient training data, we propose a texture constrained non-rigid ICP method to register RGB-D images robustly. Besides, an out-of-plane pose augmentation method specifically designed for RGB-D data is proposed to enrich pose variations and enlarge the scale of data. Secondly we propose a novel network structure FGNet

Table 5. The NME(%) on Florence, evaluated by different yaw ranges. The FGNet employs the FGNet-MV structure.

Input	$[30, 60]$	$[30, 60]$	$[60, 90]$	All
3DDFA [41]	6.92	6.89	6.82	6.87
PRNet [9]	6.71	6.98	8.04	7.41
Extreme3D [32]	8.03	8.38	8.53	8.37
Deng et al. [7]	6.05	6.31	6.02	6.12
FGNet	5.62	5.52	5.56	5.56

that can concentrate on shape modification by learning the image to shape mapping in UV space. Finally, our method successfully reconstructs fine-grained shape geometry and outperforms other state-of-the-art methods.

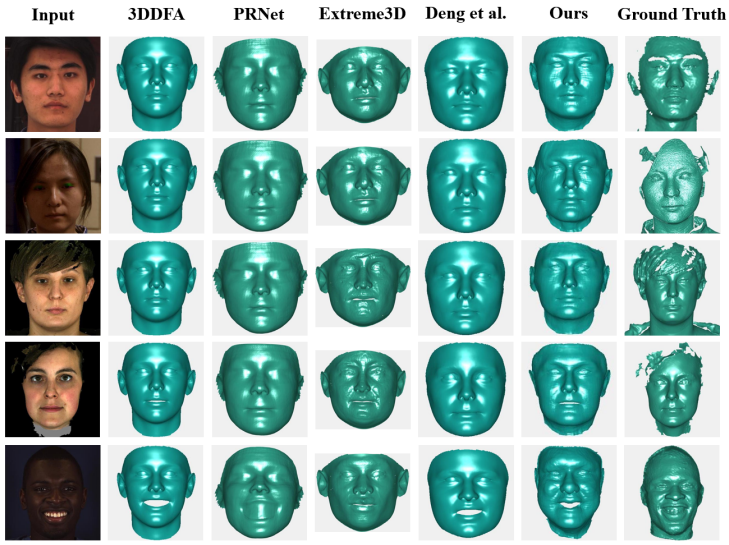

Fig. 6. Qualitative comparison. Baseline methods from left to right: 3DDFA [41] (used as our initial shape), PRNet [9], Extreme3D [32], Deng et al. [7], our FGNet and the ground-truth shape.

7 Acknowledgment

This work was supported in part by the National Key Research & Development Program (No. 2020AAA0140002), Chinese National Natural Science Foundation Projects #61806196, #61876178, #61872367, #61976229, #61673033.

References

1. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
2. Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. pp. 79–80. ACM (2011)
3. Bas, A., Huber, P., Smith, W.A., Awais, M., Kittler, J.: 3d morphable models as spatial transformer networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 904–912 (2017)
4. Bhagavatula, C., Zhu, C., Luu, K., Savvides, M.: Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses (2017)
5. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Transactions on pattern analysis and machine intelligence **25**(9), 1063–1074 (2003)
6. Cai, Y., Lei, Y., Yang, M., You, Z., Shan, S.: A fast and robust 3d face recognition approach based on deeply learned face representation. Neurocomputing **363**, 375–397 (2019)
7. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Computer Vision and Pattern Recognition Workshops (2019)
8. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5908–5917 (2017)
9. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
10. Hassner, T.: Viewing real-world faces in 3d. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3607–3614 (2013)
11. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G., Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G., Jackson, A.S., Bulat, A.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression (2017)
12. Jianzhu Guo, X.Z., Lei, Z.: 3ddfa. https://github.com/cleardusk/3DDFA (2018)
13. Jourabloo, A., Liu, X.: Pose-invariant 3d face alignment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3694–3702 (2015)
14. Jourabloo, A., Liu, X.: Large-pose face alignment via cnn-based dense 3d model fitting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4188–4196 (2016)
15. Kemelmacher-Shlizerman, I., Basri, R.: 3d face reconstruction from a single image using a single reference face shape. IEEE transactions on pattern analysis and machine intelligence **33**(2), 394–405 (2011)
16. Li, S.: Casia 3d face database - center for biometrics and security research (2004)
17. Liu, F., Tran, L., Liu, X.: 3d face modeling from diverse raw scan data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9408–9418 (2019)
18. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Joint face alignment and 3d face reconstruction. In: European Conference on Computer Vision. pp. 545–560. Springer (2016)

19. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1619–1628 (2017)
20. Mu, G., Huang, D., Hu, G., Sun, J., Wang, Y.: Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5773–5782 (2019)
21. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: 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. pp. 296–301. IEEE (2009)
22. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 947–954. IEEE (2005)
23. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 460–469. IEEE (2016)
24. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1259–1268 (2017)
25. Romdhani, S., Vetter, T.: Efficient, robust and accurate fitting of a 3D morphable model. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 59–66. IEEE (2003)
26. Romdhani, S., Vetter, T.: Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on. vol. 2, pp. 986–993. IEEE (2005)
27. Saval-Calvo, M., Azorin-Lopez, J., Fuster-Guillo, A., Villena-Martinez, V., Fisher, R.B.: 3d non-rigid registration using color: Color coherent point drift. Computer Vision and Image Understanding **169**, 119–135 (2018)
28. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1576–1585 (2017)
29. Shen, T., Huang, Y., Tong, Z.: Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
30. Snta, Z., Kato, Z.: 3D Face Alignment Without Correspondences (2016)
31. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1274–1283 (2017)
32. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.G.: Extreme 3d face reconstruction: Seeing through occlusions. In: CVPR. pp. 3935–3944 (2018)
33. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
34. Tran, L., Liu, X.: On learning 3d face morphable model from in-the-wild images. IEEE transactions on pattern analysis and machine intelligence (2019)
35. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: 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. pp. 5163–5172 (2017)

36. Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., Shi, H., Wang, Z., Li, S.Z.: A dataset and benchmark for large-scale multi-modal face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 919–928 (2019)
37. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014)
38. Zhu, K., Du, Z., Li, W., Huang, D., Wang, Y., Chen, L.: Discriminative attention-based convolutional neural network for 3d facial expression recognition. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019)
39. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)
40. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 787–796 (2015)
41. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3d total solution. IEEE transactions on pattern analysis and machine intelligence **41**(1), 78–92 (2019)
42. Zhu, X., Yi, D., Lei, Z., Li, S.Z.: Robust 3d morphable model fitting by sparse sift flow. In: 2014 22nd International Conference on Pattern Recognition. pp. 4044–4049. IEEE (2014)