ORIGINAL ARTICLE

Fast 3D face reconstruction from a single image combining attention mechanism and graph convolutional network

Zhuoran Deng[1] · Yan Liang[1] · Jiahui Pan[1] · Jiacheng Liao[1] · Yan Hao[1] · Xing Wen[1]

Abstract

In recent years, researchers have made significant contributions to 3D face reconstruction with the rapid development of deep learning. However, learning-based methods often suffer from time and memory consumption. Simply removing network layers hardly solves the problem. In this study, we propose a solution that achieves fast and robust 3D face reconstruction from a single image without the need for accurate 3D data for training. In terms of increasing speed, we use a lightweight network as a facial feature extractor. As a result, our method reduces the reliance on graphics processing units, allowing fast inference on central processing units alone. To maintain robustness, we combine an attention mechanism and a graph convolutional network in parameter regression to concentrate on facial details. We experiment with different combinations of three loss functions to obtain the best results. In comparative experiments, we evaluate the performance of the proposed method and state-of-the-art methods on 3D face reconstruction and sparse face alignment, respectively. Experiments on a variety of datasets validate the effectiveness of our method.

Keywords 3D face reconstruction · Lightweight network · Attention mechanism · Graph convolutional network

1 Introduction

As a fundamental topic in computer vision and graphics, 3D face reconstruction can be used for face recognition [1–4], face alignment [5–8], emotion analysis [9], and face animation [10]. Over the years, many novel approaches to 3D face reconstruction from a single image have been proposed. The major studies focus on how to achieve robust 3D face reconstruction with high fidelity, but the considerations of speed and cost are ignored or not mentioned. In practice, the speed and cost of production use should be valued. Spending much time reconstructing a 3D face model can lead to a poor user experience, and the speed of reconstruction is largely dependent on the performance of the method. The cost of a complete 3D face reconstruction method includes, but is not limited to, data collection cost, resource consumption cost, and usage cost. It will be difficult to meet the produc-

tion needs of a high-fidelity 3D face reconstruction that takes excessive time to build or relies on special equipment. Thus, a balance needs to be struck between the various aspects.

Since a 2D image contains severely limited effective features, it is difficult to recover a detailed and highly realistic face model. Apart from the expensive device capture methods, studies have developed ways to conduct accurate 3D face reconstruction from a single 2D image based on the addition of prior knowledge. A traditional approach is to reconstruct a 3D face model by fitting a statistical model (e.g. 3D morphable model (3DMM) [11]). A mean face model with linear representations of face shape and texture is employed to fit a given image by optimization calculation. A limitation of model fitting methods is the restricted representation of a statistical model. They often fail to restore nonlinear facial details, making the 3D face model artificial.

In recent years, deep learning has become the more preferred approach for adding prior knowledge. Modelling a 3D face mesh is accomplished by learning the mapping between the 2D image and the 3D face model. With the development of neural networks, learning-based methods enable the acquisition of accurate 3D face reconstruction. Nevertheless, there is a scarcity of 2D–3D paired face datasets available, and collecting a large-scale detailed 3D face dataset is difficult

✉ Yan Liang
liangyan@m.scnu.edu.cn

Zhuoran Deng
dengzr@m.scnu.edu.cn

1 School of Software, South China Normal University, Foshan 528225, China

for ordinary users. In addition, learning 3D face reconstruction through a deep neural network involves a great deal of iterative computation, as well as processor and memory consumption. When dealing with a large amount of data, typical central processing unit (CPU) cores struggle to cope with such demands. This restricts the application of 3D face reconstruction on mobile phones even more. Moreover, the stability and robustness of the methods should be guaranteed.

In this paper, we aim to propose an efficient learning-based method of 3D face reconstruction from a single image. We strike a balance between quality, speed, and cost. In consideration of the cost of 3D data collection and preparation, we convert the task of 3D face reconstruction into a small set of 3DMM parameter regressions in the absence of accurate 3D data. To improve speed, we apply a lightweight network to extract features from images. Considering that the lightweight network is prone to loss of precision, we introduce an attention mechanism [12] and a graph convolution network (GCN) [13] in the regression. During training, three different loss functions are added to calculate the loss of 3DMM parameters, reconstructed 3D vertices, and landmarks. Various combinations of loss functions are explored to obtain the most efficient strategy. Furthermore, the performance of the proposed method is evaluated not only in a 3D face reconstruction benchmark but also in sparse face alignment.

To summarize, in this study, we concentrate on maintaining fast and robust 3D face reconstruction from a single image without accurate 3D training data. The main contributions are as follows.

– We propose a lightweight network-based framework for 3D face reconstruction to address the problems of computation speed and graphics processing unit (GPU) dependency.
– We propose a combination of an attention mechanism and a GCN for the regression of 3DMM parameters, which can improve the accuracy and robustness of the reconstructed model.
– In our experiments, we validate the effectiveness of the proposed method and obtain an optimal result by comparing different loss function strategies. Compared with state-of-the-art methods, our method achieves considerable benefits.

2 Related work

3D-from-2D face reconstruction is a long-standing topic in the field of computer vision. A large number of complex optimization calculations need to be carried out to recover a 3D face model from an image. To simplify the problems, a common way to constrain the space of solutions is by adding

prior knowledge. Relevant studies on single image-based 3D face reconstruction are mainly of two varieties: statistical model fitting methods and learning-based methods.

2.1 Statistical model fitting methods

Before the advent of deep learning, prior knowledge was embedded in a statistical face model. Specifically, an initialized mean face model is computed from a large dataset of 3D facial scans containing low-dimensional representations of the face shape and texture. The 3D face model is then fitted to an image through a series of optimization calculations so that the image generated by projecting the resulting 3D face model onto a 2D plane is as similar as possible to the input image. The most widely used models are based on 3DMM [1,11]. Blanz and Vetter [11] proposed the first morphable 3D face model in 1999, using principal component analysis (PCA) for dimensionality reduction decomposition. Subsequently, many derivatives of 3DMM have appeared, such as the basel face model (BFM) [2], a widespread model that can fit any 3D face and store its 3DMM parameters. The first BFM is unable to adjust facial expressions. A typical strategy is to join the expression basis from FaceWarehouse [14], such as the methods of [5] and [7]. The integrated BFM has 199-dimensional shape vectors, 199-dimensional texture vectors, and 29-dimensional expression vectors.

In general, the core of statistical model fitting methods is to find the optimal solution to 3DMM parameters that minimizes the difference between each input and rendered image. Piotraschke and Blanz [15] reconstructed a 3D face from a set of face images by reconstructing each face individually and then combining them into a final shape based on the accuracy of each reconstructed part. Jin et al. [16] took two images of a person's front and side view as input to develop a deformable nonnegative matrix factorization (NMF) part-based 3D face model and used an automated iterative reconstruction method to obtain a high-fidelity 3D face model. In contrast, both Jiang et al. [17] and Liu et al. [18] proposed 3D face reconstruction methods based on a single image. The former used a bilinear face model and local corrected deformation fields to reconstruct high-quality 3D face models, while the latter improved accuracy by updating contour landmarks and self-occluded landmarks. However, these methods have obvious shortcomings. Building such a 3DMM model by computing a nonlinear error function requires expensive iterative optimization. It also tends to get stuck in local minima so that its accuracy and authenticity cannot be guaranteed. Aldrian and Smith [19,20] suggested a solution that used linear methods to recover shape and texture separately. In addition, Schönborn et al. [21] proposed a different model fitting method, regarding the 3DMM fitting process as a probabilistic reasoning problem. They interpreted 3DMM as a generative Bayesian model and used

random forests as noisy detectors. The two were then combined using a data-driven Markov chain Monte Carlo method (DDMCMC) based on the Metropolis–Hastings algorithm.

2.2 Learning-based methods

Thanks to the rapid development of deep learning, a 3D face model can be recovered from a single image by using a convolutional neural network (CNN) to encode prior knowledge in the weights of the trained network. Nonetheless, there are numerous challenges in reconstructing 3D face models from images.

First, the problems of training data should be fixed, mainly in terms of data volume and data diversity. Many 3D face reconstruction methods often perform poorly on images where the facial features are partially occluded or self-occluded due to large poses because of the reduction of valid features. One solution is to expand the training data [5,6,22,23]. Richardson et al. [22] randomly modified 3DMM and rendered them onto a 2D plane to generate synthetic 2D images. Zhu et al. [5] proposed synthesizing 3D faces by directly regressing the 3DMM parameters from the input images. The other solution is generally through strong regularization of shape [24]. Focusing on occlusion and pose problems, Ruan et al. [25] proposed a self-aligned dual-face regression network combined with an attention mechanism to solve them.

The second problem is improving accuracy and robustness. Deng et al. [26] proposed an accurate 3D face reconstruction by introducing a differentiable renderer and designing a hybrid loss function for weakly supervised training. Sanyal et al. [27] used multiple images of the same subject and an image of a different subject to learn shape consistency and inconsistency during training to enhance robustness. Recent studies [28–30] have demonstrated that GCNs [13,31,32] contribute to the recovery of facial details. Lin et al. [28] obtained high-fidelity 3D face models by utilizing a GCN to decode the features extracted from single face images and then produce detailed colours for the face mesh vertices. Lee and Lee [29] introduced an uncertainty-aware mesh encoder as well as a decoder that combined a GCN with a generative adversarial network (GAN), to solve the problems of occlusion and blur. Based on GCNs, Gao et al. [30] proposed decoding the identity and expression features extracted from a CNN to recover 3D face shape and albedo, respectively.

A further concern is the speed and cost of 3D face reconstruction. Most of the above methods regress 3DMM parameters based on a deep convolutional neural network (DCNN). The size of the trained network is usually large, with numerous parameters and computations, resulting in slow inference and memory consumption. The inference time would be much longer on a CPU alone, or the huge amount

of computing could not even be handled by a CPU core. To overcome these shortcomings, solutions can be found by reducing the 3DMM parameters for regression [8] or using an image-to-image CNN instead of a regression network [6,33]. Feng et al. [6] designed a novel method to record the 3D shape of a face using a UV position map, which enabled fast reconstruction. Koizumi and Smith [33] estimated the correspondence from an image to a face model based on an image-to-image CNN without ground truth or landmarks. Guo et al. [8] reduced the dimensions of 3DMM parameters and performed fast and stable 3D face reconstruction based on a lightweight CNN.

3 Proposed method

In this section, we introduce our work in detail. First, we describe the composition of 3DMM. We then detail each component of the proposed network architecture. Specifically, there are two modules: one for fast feature extraction based on a lightweight network and the other for parameter regression combined with an attention mechanism and a GCN. After that, we introduce three loss functions used in training. Our framework is shown in Fig. 1.

3.1 3DMM parameter regression

The face shape and texture of 3DMM can be defined as:

$$\mathbf{S}_{\text{model}} = \bar{\mathbf{S}} + \mathbf{B}_{\text{shp}}\boldsymbol{\alpha}_{\text{shp}},$$
$$\mathbf{T}_{\text{model}} = \bar{\mathbf{T}} + \mathbf{B}_{\text{tex}}\boldsymbol{\alpha}_{\text{tex}}, \tag{1}$$

where $\mathbf{S}_{\text{model}}$ and $\mathbf{T}_{\text{model}}$ are the face shape vector and texture vector, respectively; $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the mean face shape and texture, respectively; $\mathbf{B}_{\text{shp}}$ and $\mathbf{B}_{\text{tex}}$ are the PCA bases of face shape and texture, respectively; and $\boldsymbol{\alpha}_{\text{shp}}$ and $\boldsymbol{\alpha}_{\text{tex}}$ represent the corresponding parameters.

Typically, a full 3DMM parameter regression still needs to estimate pose parameters, illumination parameters, and camera parameters, so that the output model can be projected onto a plane and compared for similarity with the input image. For the purpose of fast 3D face reconstruction, we remove some of the parameters and reduce the dimensions of the remaining parameters, referring to previous studies [5,7,8]. Therefore, we only learn the 3DMM parameters of shape, expression, and pose in the regression task. Here, the 3D shape in Eq. 1 is described as:

$$\mathbf{S}_{\text{model}} = \bar{\mathbf{S}} + \mathbf{B}_{\text{shp}}\boldsymbol{\alpha}_{\text{shp}} + \mathbf{B}_{\text{exp}}\boldsymbol{\alpha}_{\text{exp}}, \tag{2}$$

where the 3D expression base $\mathbf{B}_{\text{exp}}$ and corresponding parameters $\boldsymbol{\alpha}_{\text{exp}}$ are added. Given a face image, the network estimates a vector with 62 dimensions $(\mathbf{T}, \boldsymbol{\alpha}_{\text{shp}}, \boldsymbol{\alpha}_{\text{exp}}) \in \mathbb{R}^{62}$,
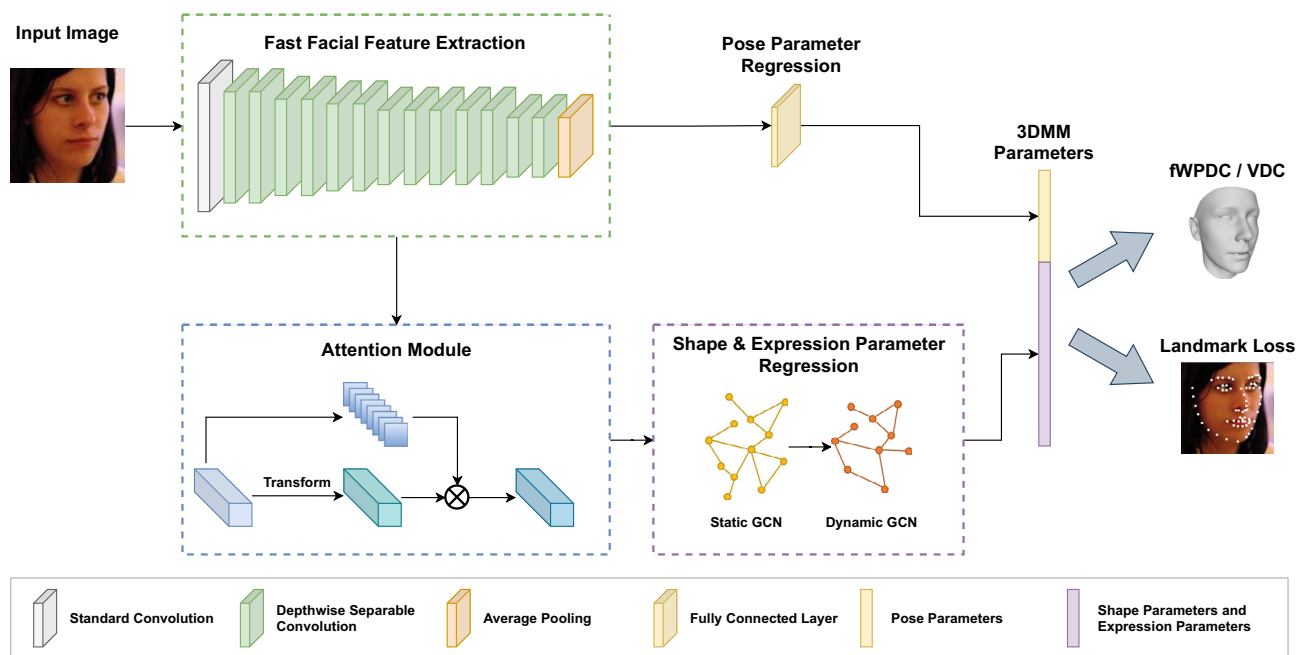
Fig. 1 Framework of the proposed method

where $\mathbf{T} \in \mathbb{R}^{3 \times 4}$ is a transformation matrix representing the face pose. $\boldsymbol{\alpha}_{\mathrm{shp}} \in \mathbb{R}^{40}$ and $\boldsymbol{\alpha}_{\mathrm{exp}} \in \mathbb{R}^{10}$ are the 3DMM parameters of shape and expression, respectively. After regression, the 3D face can be computed with Eq. 2.

3.2 Network architecture

As shown in Fig. 1, we propose a fast parameter regression strategy based on a lightweight network, combined with an attention mechanism and a GCN. First, we employ a lightweight network, i.e. MobileNet [34], to extract features from images quickly. According to Eq. 2, constructing a 3D face model mainly depends on the estimation of shape parameters and expression parameters, which influence the final performance of the face model. As a result, we then separate the 3DMM parameters into two parts for regression. One is the parameter regression of shape and expression, where the attention mechanism and GCN are introduced to improve robustness and stability. The other is the regression of pose parameters, performed by a fully connected layer.

3.2.1 Fast feature extraction

Adopting a lightweight network enables fast and stable feature extraction. In this regard, we choose MobileNet [34] to extract features from images. Instead of standard convolutions, MobileNet introduces depthwise separable convolutions. When performing convolutions, a standard convolution kernel considers all channels in the corresponding image region simultaneously, which increases the computation mul-

tiplicatively. The depthwise separable convolution factorizes a convolution into a depthwise convolution and a pointwise convolution, decoupling channel correlation and spatial correlation. In this way, both the number of parameters and computational cost are greatly reduced, but accuracy is still guaranteed. The feature extraction network is a modification of MobileNet-V1 [34], where the last fully connected layer is replaced with two branches for the next regression step.

3.2.2 Enhanced attention to facial features

An attention mechanism is introduced with the aim of concentrating on the context-aware representation of facial features from the extracted feature map and suppressing other useless information, such as image backgrounds. With reference to [35], we generate attention masks $\mathbf{M}$ and the transformed feature map $\mathbf{X}$ from the extracted features. The attention masks $\mathbf{M}$ are treated as weights of different channels, which are then multiplied by the transformed feature map $\mathbf{X}$ to construct the final content-aware matrix $\mathbf{H_c}$. Specifically, the formula can be defined as:

$$\mathbf{H}_c^k = \sum_{i=1}^{h} \sum_{j=1}^{w} m_{ij}^k \mathbf{x}_{ij}, \tag{3}$$

where m_{ij}^k denotes the kth weight of the attention mask, $\mathbf{x}_{ij}$ denotes the feature vector of the transformed feature map $\mathbf{X}$ at (i, j), and h and w represent the height and width of the input image, respectively.

3.2.3 Graph convolutional network for robust parameter regression

Unlike CNNs, GCNs can perform convolution operations on non-Euclidean structured data. Accordingly, we introduce a static GCN and a dynamic GCN [35] to help restore unstructured details of the face model.

For the static GCN, any static graph convolutional layer can be defined as:

$$H^{l+1} = \sigma(A H^l W), \tag{4}$$

where H^l and W are the input nodes and weight matrix of the lth layer, respectively, A is the adjacency matrix, and $\sigma(\cdot)$ denotes the nonlinear activation function. Here, $H^0 = H_c$ is the input nodes of the single-layer static GCN, which comes from Eq. 3. Thus, the formula of the static GCN is defined as:

$$H_s = \sigma(A_s H_c W_s), \tag{5}$$

where H_s denotes the updated nodes, A_s and W_s are the adjacency matrix and the weight matrix of the static GCN, respectively, and the activation function $\sigma(\cdot)$ is $LeakyReLU$. The multiplication of the adjacency matrix A_s with the features H_c is equivalent to the sum of the features of the neighbouring nodes of a node. In this way, each node can use the information of neighbours to update the state.

Compared with the static GCN, a significant characteristic of the dynamic GCN is the dynamic adjacency matrix update. Since the adjacency matrix A_s of the static GCN is fixed, it is unreasonable to use the same adjacency matrix for all inputs. The dynamic GCN can overcome this weakness by adaptively constructing the adjacency matrix A_d according to the input features. Intuitively, recalculating the adjacency matrix for each input can better spread information between similar structures and speed up the learning of local semantic information. Specifically, the dynamic GCN can be defined as:

$$H_d = \sigma(A_d H_s W_d), \tag{6}$$

where H_d are the output 3DMM parameters, W_d is the state-update weight matrix of the dynamic GCN, and the adjacency matrix A_d of the dynamic GCN is defined as:

$$A_d = \delta(W_a H_s'), \tag{7}$$

where $\delta(\cdot)$ is the $Sigmoid$ activation function, W_a is the weight matrix obtained by convolution, and H_s' is constructed by concatenating H_s with its global expression.

3.3 Loss functions

During training, we adopt three loss functions to handle the optimizer: fast weighted parameter distance cost (fWPDC) [8], vertex distance cost (VDC) [5], and landmark loss.

3.3.1 Fast weighted parameter distance cost

Generally, the WPDC [5] is used to constrain the loss between the predicted parameters and ground truth. The WPDC sets different weights for each parameter. The formula can be defined as:

$$L_{\text{wpdc}} = \| w \cdot (p - p^{gt}) \|^2, \tag{8}$$

where $w = (w_1, w_2, \ldots, w_{62})$ denotes the weight of each parameter, $p = (p_1, p_2, \ldots, p_{62})$ is the predicted 3DMM parameter, and $p^{gt} = (p_1^{gt}, p_2^{gt}, \ldots, p_{62}^{gt})$ is the ground truth. To simplify the calculation, fWPDC [8] separates the parameters into two parts for calculation. That is, $p = [T, \alpha]$, where T is a transformation matrix from the predicted pose parameters and $\alpha = [\alpha_{\text{shp}}, \alpha_{\text{exp}}]$ are the predicted shape parameter and expression parameter. Equation 8 is changed to:

$$L_{\text{fwpdc}} = \| w_T \cdot (T - T^{gt}) \|^2 + \| w_\alpha \cdot (\alpha - \alpha^{gt}) \|^2, \tag{9}$$

where w_T and w_α are the weights of the corresponding parameters, and T^{gt} and α^{gt} refer to the ground truth of the corresponding parameters.

3.3.2 Vertex distance cost

The VDC [5] minimizes the vertex distances between the reconstructed 3D face shapes and the ground truth. The vertices of a 3D face model are generated by the predicted 3DMM parameters p. The formula is defined as:

$$L_{\text{vdc}} = \| V_{3d}(p) - V_{3d}(p^{gt}) \|^2, \tag{10}$$

where $V_{3d}(\cdot)$ is the reconstructed vertices calculated by:

$$V_{3d}(p) = T * \begin{bmatrix} S_{\text{model}} \\ 1 \end{bmatrix}, \tag{11}$$

where T is the same as that in Eq. 9 and S_{model} is from Eq. 2.

3.3.3 Landmark loss

To improve the robustness of face reconstruction, we adopt sparse landmark loss to constrain the 3DMM parameters to better fit the input. We additionally detect 68 facial landmarks $\{q_n^{gt}\}$ of each input image as ground truth. During training, we obtain 2D landmarks $\{q_n\}$ by projecting the 3D landmark

vertices of the reconstructed model onto images. Then, the loss is formulated as:

$$L_{\text{lmk}} = \frac{1}{N} \sum_{n=1}^{N} \| q_n - q_n^{gt} \|. \tag{12}$$

Here, N is 68. We obtain the loss values by means of Euclidean distance computation.

4 Experiment

4.1 Datasets and evaluation metrics

In this study, we aim to produce fast and robust 3D face reconstruction from a single image without the use of accurate 3D training data. As a result, 300W-LP [5] is chosen as the training dataset. 300W-LP is composed of synthesized 3D faces with large poses from 300W [36], which includes the annotated faces in the wild (AFW) [37], labeled face parts in the wild (LFPW) [38], HELEN [39], IBUG [36], and extended Multi-Modal Verification for Teleservices and Security Applications (XM2VTS) [40] datasets. Practically, the training dataset we used consists of over 600,000 still images, extended by Guo et al. [8]. Since the extended dataset does not contain ground truth of facial landmarks, we adopt a face align network (FAN) [41] to extract 68 2D facial landmarks of each image as ground truth. During the collection of landmarks, we remove the samples that failed to be detected. In total, there are 626,088 as the training set as well as 50,807 for validation.

To evaluate the performance of our method on 3D face reconstruction, we employ the not quite in the wild (NoW) [27] dataset. The NoW dataset contains 2,054 2D images of 100 subjects, with a 3D face scan for each subject. The images are categorized into four cases: neutral, expression, occlusion, and selfie. We follow the NoW benchmark [27] to evaluate the performance of 3D face reconstruction. Specifically, the benchmark calculates the scan-to-mesh distance between the ground truth scan and the reconstructed mesh. The median distance, mean distance, and standard deviation are then recorded, as well as a cumulative error plot for all distances.

Most methods of face reconstruction support face yaw angles of less than 45° or when all facial landmarks are visible, which is not able to align faces in extreme cases such as large poses up to 90°. To test the performance of our method on images with large poses, we evaluate sparse face alignment accuracy with small, medium, and large yaw angles (i.e. yaw angle ψ corresponding to $0° \leq \psi \leq 30°$, $30° < \psi \leq 60°$, and $60° < \psi \leq 90°$, respectively) using the normalized mean error (NME) by bounding box size on AFLW2000-3D

according to [5]. AFLW2000-3D consists of fitted 3D face models of the first 2,000 samples from the annotated facial landmarks in the wild (AFLW) [42] and corresponding 68 3D facial landmarks.

Additionally, we introduce the CelebFaces Attribute (CelebA) [43] dataset in the qualitative analysis. The CelebA dataset contains 202,599 face images with 10,177 celebrities and 40 attribute annotations, including large poses, occlusion, blur, and background clutter. We group the images by certain attributes based on experimental needs. Then, we evaluate the quality of the 3D face models reconstructed from the images with different attributes.

4.2 Implementation details

We perform experiments mainly based on PyTorch. For feature extraction, we utilize MobileNet-V1 [34] with a width multiplier of 1.0 but remove the last fully connected layer and output a feature map with a channel of 1024. For 3DMM parameter regression, we adopt the activation function LeakyReLU in the GCN, with a negative slope value set to 0.2. The input image size is 120×120 and is normalized according to [8]. We train our network with a batch size of 64 on an NVIDIA GeForce RTX 3080 GPU. The initial learning rate is 0.01, but we set a warmup in the first 5 epochs with a learning rate of 0.002. During training, we use stochastic gradient descent (SGD) as the optimizer with a weight decay of 0.0005 and a momentum of 0.9. In addition, we add gradient clipping to avoid gradient explosion. First, we conduct experiments with the fWPDC computed to observe the joint effect of the attention mechanism and GCN. Second, we adjust the training loss functions to obtain the best and most stable results. As stated in Guo et al. [8], using VDC from scratch may obtain higher vertex error, but better results can be achieved by using VDC from fWPDC or combining VDC and fWPDC. Thus, we conduct separate experiments using three different combinations of loss functions. The first combination strategy is to train our network using fWPDC and landmark loss:

$$L_1 = L_{\text{fwpdc}} + w_{\text{lmk}} L_{\text{lmk}}, \tag{13}$$

where $w_{\text{lmk}} \approx 10^{m_{\text{fwpdc}} - m_{\text{lmk}}}$ is the training weight used to balance the two losses, and m_{fwpdc} and m_{lmk} indicate the magnitude of each loss. The second is to calculate three loss functions simultaneously:

$$L_2 = L_{\text{fwpdc}} + w_{\text{vdc}} L_{\text{vdc}} + w_{\text{lmk}} L_{\text{lmk}}, \tag{14}$$

where $w_{\text{vdc}} \approx 10^{m_{\text{fwpdc}} - m_{\text{vdc}}}$ has the same effect as w_{lmk} and m_{vdc} indicates the magnitude of VDC. The last strategy is to divide the training into two stages, with different loss functions calculated for each stage. The first stage is to train

the network using Eq. 13. When the training converges to fit, we adjust the loss combination in the second stage to:

$$L_3 = w_{\mathrm{vdc'}}L_{\mathrm{vdc}} + L_{\mathrm{lmk}}, \tag{15}$$

where $w_{\mathrm{vdc'}} \approx 10^{m_{\mathrm{lmk}}-m_{\mathrm{vdc}}}$.

4.3 Ablation study

We conduct comparative experiments with different schemes to verify the effectiveness of the proposed method. In our experiments, we combine the modified MobileNet, attention mechanism, and GCN for training and search for the best result with different training strategies. The original MobileNet-V1 is employed as a baseline.

4.3.1 Evaluation on the NoW validation set

We perform experiments on the NoW [27] validation set in ablation experiments, and the results are shown in Table 1. Obviously, the combined network we proposed is effective. As shown in the first and second rows of Table 1, performance improves when the attention mechanism and GCN are applied, where the median error, mean error, and standard deviation error are reduced by 0.07 mm, 0.08 mm, and 0.07 mm, respectively, on the NoW validation set. Throughout the three strategies we proposed, the effect is generally enhanced when other loss functions are added. When joined by landmark loss (i.e. using L_1), the effect is not significant. However, when experimenting with the network trained with three loss functions together (i.e. using L_2), the errors are reduced, dropping by 0.04 mm on the median error and 0.03 mm on the mean error, compared with using L_1. Finally, we achieve the best results using the third strategy to train the network, namely, using L_1 with L_3 fine-tuning. Compared with using fWPDC only, the median error is reduced by 0.06 mm, the mean error is reduced by 0.06 mm, and the standard deviation error is reduced by 0.02 mm.

4.3.2 Evaluation on AFLW2000-3D

Similar results are obtained in the evaluation of sparse face alignment on AFLW2000-3D [5], and the results can be seen in Table 2. Compared with the baseline, the combined network we proposed performs better, with a mean NME reduction of 0.20% when using the same loss function. However, unlike the above 3D face reconstruction evaluation results, there is a significant reduction in errors with the addition of landmark loss (i.e. using L_1), decreasing the mean NME by 0.41%. The trained network with L_2 performs best at yaw angles ranging from 0° to 30°, but does not perform as well as the network trained by L_1 with L_3 fine-tuning when

yaw angles exceed 30°. They differ from each other in the mean error by 0.02%.

To show the effect of the attention mechanism and GCN more clearly, we visualize the sparse alignment results of the model trained with and without the attention mechanism and GCN in Fig. 2. In the demonstrated cases, the model with the attention mechanism and GCN obtains more accurate alignment results. In the first example with the attention mechanism and GCN, the facial contour landmarks fit the input face much more precisely. In the second example, the landmarks labelled in the nose region are more correctly located. In the third example, the labelling of the mouth's openness is more appropriate. In the last example, in the case of a relatively large yaw of the face pose, the result with the added module performs better.

4.3.3 Visualization of feature maps

To demonstrate the utility of using an attention mechanism to capture facial features, we visualize images from AFLW2000-3D [5] with their corresponding learned feature maps based on class activation mapping (CAM) [44]. Some samples are shown in Fig. 3. The heatmaps of the images are generated from MobileNet-V1 and the attention module of the proposed method in the case of using the same training loss functions. We note that the attention mechanism we employ is capable of paying attention to the face region. Comparing the third, fourth, fifth, and sixth columns of Fig. 3, our method shows robustness for images with different poses. Even for images with extremely large poses, the proposed attention module still focuses on the crucial parts. Moreover, as shown in the last two columns of Fig. 3, our method improves attention to nonocclusion regions.

4.4 Comparison with prior art

4.4.1 Qualitative analysis

For qualitative evaluation, we compare the resulting 3D shapes reconstructed from images of different properties on AFLW2000-3D [5] and CelebA [43] using different methods. Specifically, we compare our method with PRNet [6] and 3DDFA-V2 [8], which proposed fast 3D face reconstruction without accurate 3D training data, similar to our method. Some of the results are shown in Fig. 4.

First, we reconstruct the images with different yaw angles on AFLW2000-3D [5], classifying yaw angles into small, medium, and large angles. As shown in the first four rows of Fig. 4, both the proposed method and other methods obtain reasonable results. In comparison, our method shows better robustness in the case of images with large yaw angles. When the yaw angle increases to 60° or even more (the second column of the third and fourth rows of Fig. 4), the reconstructed

Table 1 Reconstruction error on the NoW validation set

Network	Loss functions	Median (mm)	Mean (mm)	Std[a] (mm)
Baseline[b]	L_{fwpdc}	1.40	1.73	1.44
MobileNet + attention & GCN	L_{fwpdc}	1.33	1.65	1.37
	L_1	1.32	1.66	1.42
	L_2	1.29	1.62	1.37
	$L_1 + L_3$ fine-tune	**1.27**	**1.59**	**1.35**

The best results are highlighted in bold

[a] Std denotes the standard deviation

[b] The baseline indicates using the original MobileNet-V1

Table 2 NME (%) of 68 landmarks with different yaw angles ψ on AFLW2000-3D

Network	Loss functions	$0° \leq \psi \leq 30°$	$30° < \psi \leq 60°$	$60° < \psi \leq 90°$	Mean
Baseline	L_{fwpdc}	3.48	4.42	5.66	4.52
MobileNet + attention & GCN	L_{fwpdc}	3.36	4.26	5.33	4.32
	L_1	3.02	3.80	4.92	3.91
	L_2	**2.73**	3.53	4.51	3.59
	$L_1 + L_3$ fine-tune	2.76	**3.46**	**4.49**	**3.57**

The best results are highlighted in bold

Fig. 2 Comparison for sparse alignment on AFLW2000-3D. Partial face regions are magnified for better visual comparison
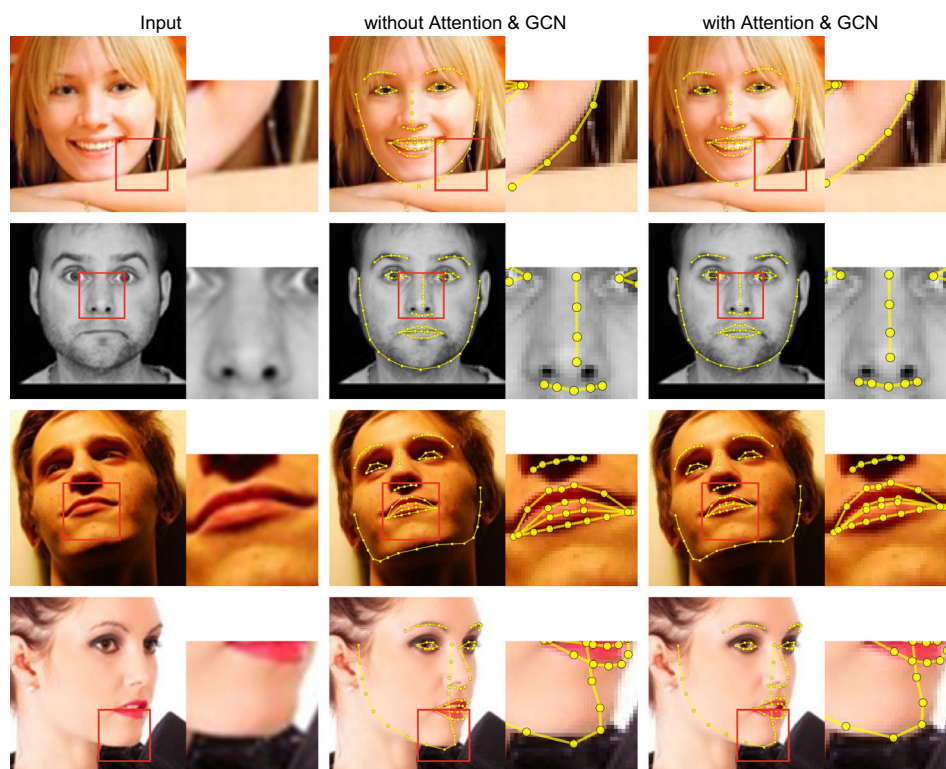

3D face shape of PRNet [6] is slightly distorted with asymmetry between the left and right of the face. In the fourth row, the mouth of the input face image is tightly closed, but the mouth cannot be closed in the reconstruction results of 3DDFA-V2 [8].

We then conduct experiments on images with different attributes from CelebA [43], and the results are shown in the last five rows of Fig. 4. Based on the annotated attributes on CelebA, we group the images by the following cases: blur, occlusion (coverage by sunglasses or bangs), insufficient light, and uneven light. For the case of blurry images, as seen in the fifth row of Fig. 4, the proposed method is not affected by the quality of images. Compared to 3DDFA-V2 [8] (the third result in the fifth row), our reconstructed 3D

Fig. 3 Visualization of the learned feature maps. The images in the first row are the input images, the images in the second row are heatmaps generated by MobileNet-V1, and the images in the third row are heatmaps generated by the attention module of our method
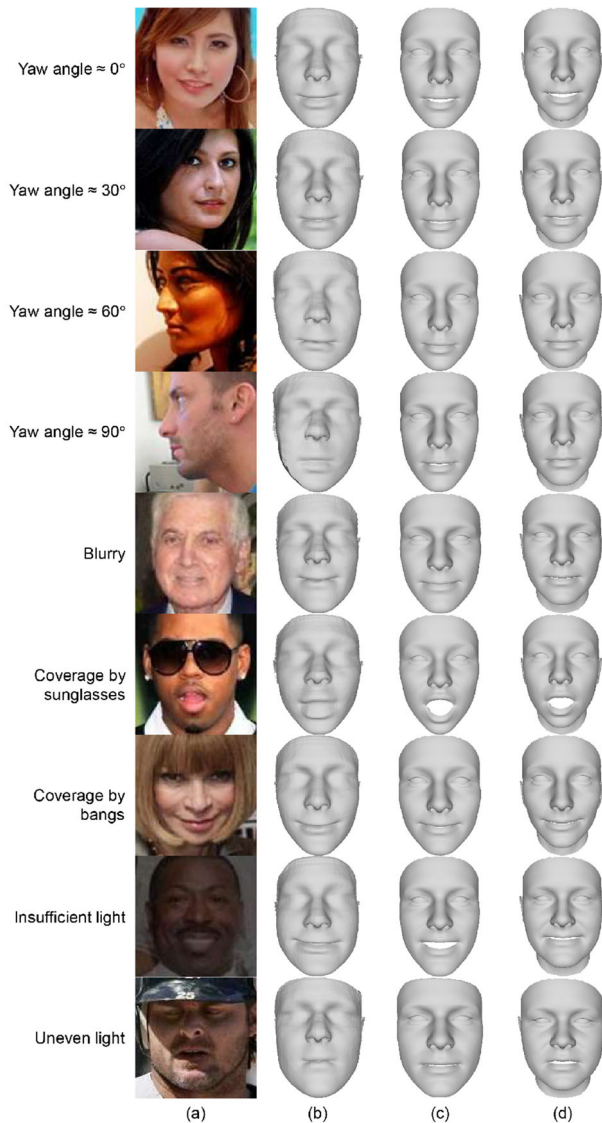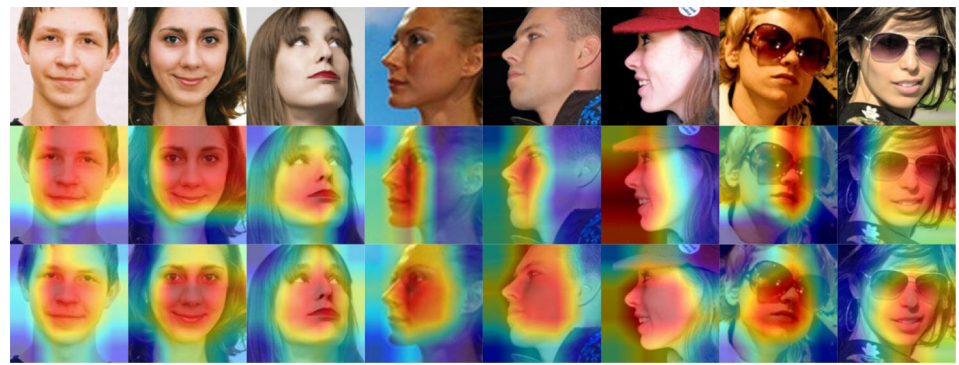
Fig. 4 Visual comparison with different methods on AFLW2000-3D and CelebA. **a** Input images. **b** PRNet [6]. **c** 3DDFA-V2 [8]. **d** Our method

face shapes fit the input images better, especially the chin part. For images with facial occlusions, such as those with

sunglasses or bangs, our method allows for reasonable reconstruction of the occluded parts, as shown in the last column of the sixth and seventh rows of Fig. 4. The performance of 3D-from-2D face reconstruction can also be easily affected when the input image is insufficiently or unevenly illuminated. However, it is clear that our method does not suffer from light problems. As shown in the last two rows of Fig. 4, the reconstructed 3D face shapes of the proposed method (the last column) achieve better results than both methods (the second and third columns).

In particular, our method can maintain a correct face shape even if the input image is in poor condition. In contrast to PRNet [6], our method reconstructs a more detailed and clearer contour of the 3D face shape, especially the area in and around the eyes and mouth. In comparison with the two state-of-the-art methods, our method shows better robustness in extreme cases.

In comparison with 3DDFA-V2 [8], we conduct a more qualitative analysis, as shown in Fig. 5. We visualize the reconstructed 3D face models under the same conditions as 3DDFA-V2 [8] on CelebA [43]. Subtle changes are difficult to visualize, so we focus on the facial region of the mouth where the comparative results are more obvious. In the first example, it is obvious that our method is much more profound in portraying the expression of the mouth. In addition, the curvature of the cheek is better. In the second and third examples with our method, the mouths are more appropriately depicted, in contrast to the open mouths in 3DDFA-V2 [8]. Overall, we can conclude that our method is able to provide more suitable facial details than 3DDFA-V2 [8].

4.4.2 Quantitative analysis

For quantitative evaluation, we perform comparative experiments of 3D face reconstruction on the NoW [27] dataset and sparse face alignment on AFLW2000-3D [5].

Evaluation of 3D face reconstruction We compare our method with the state of the art by means of the NoW [27] benchmark. The results of the comparison are shown in Table 3 and the cumulative error plot in Fig. 6. As shown in

Fig. 5 More comparison of our method with 3DDFA-V2 [8] on CelebA. The facial regions around the mouth are magnified to show the results more distinctly. The results are shown for transparency of $alpha = 0.5$ and $alpha = 0.8$
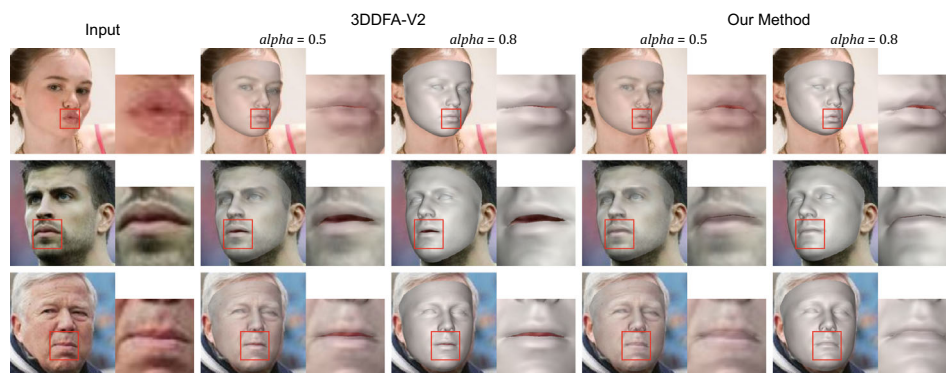

Table 3, it is obvious that our method achieves smaller reconstruction errors than 3DMM-CNN [45], UMDFA [33], and PRNet [6], where the median error is reduced by 0.55 mm, 0.23 mm, and 0.21 mm, respectively. Figure 6 also shows that the cumulative error of our method is lower than those of the three methods. A conclusion can be drawn that reconstructing 3D face models by image-to-image methods is less effective. The performance of our reconstructed 3D shapes is slightly worse than that of Dib et al. [46], 3DDFA-V2 [8], Deng et al. [26], and RingNet [27], where the median error differs by only approximately 0.03 mm to 0.08 mm, and the cumulative error is slightly higher but close to these comparative approaches. The main reason is that most of them chose to regress more 3DMM parameters to obtain more accurate results, which means more time and memory consuming processes. Dib et al. [46], Deng et al. [26], and RingNet [27] all employed a deep residual network [47] with equal to or more than 50 layers. The network would run slowly or even fail to run on typical CPUs. In contrast, our method is based on a lightweight CNN, which supports fast reconstruction on a single CPU core. Compared to 3DDFA-V2 [8], both of us adopt MobileNet to increase the speed of reconstruction. However, we target different objects of training data. We consider the case where only a single image is available as a training input, while 3DDFA-V2 [8] proposed a 3D aided short-video-synthesis strategy to allow the method to be suitable for video data. Our method is therefore more challenging in the training phase.

Evaluation of sparse face alignment We evaluate the performance of different methods in sparse face alignment by comparing the NME of 68 landmarks on AFLW2000-3D [5], and the results are shown in Table 4. This shows that our method achieves relatively good performance in face alignment for images with large poses. Overall, the mean NME of our method is smaller than that of most methods, except for 3DDFA-V2 [8]. As shown in the last three rows of Table 4, compared with 3DDFA-V2 [8] trained without the short-video-synthesis strategy, our method obtains lower errors than theirs when yaw angles are greater than 30° (3.46% vs. 3.49%/4.49% vs. 4.53%). When compared with the entire
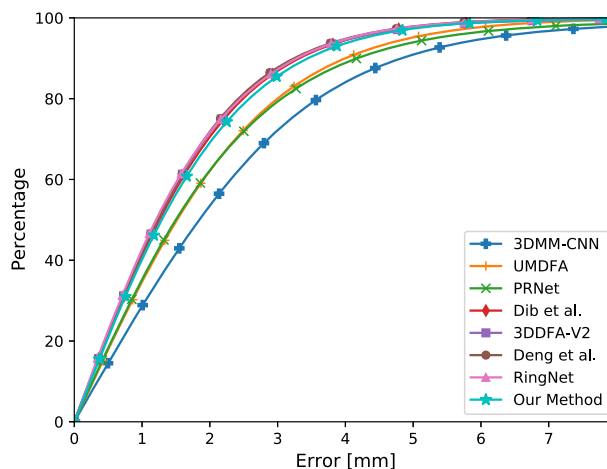
Fig. 6 Cumulative error curves of different methods on the NoW dataset

3DDFA-V2 [8], our method differs by 0.13% at yaw angles from 0° to 30° but by only 0.04% and 0.01% at yaw angles from 30° to 60° and 60° to 90°, respectively. This indicates that the proposed method performs with greater robustness when training data are more challenging, as we train the network based on only single images.

User study To measure the user preference for the reconstruction results generated by our method and 3DDFA-V2 [8], we conduct a user study. We first divide AFLW2000-3D [5] into three categories based on yaw angle ranges: small ($0° \leq \psi \leq 30°$), medium ($30° < \psi \leq 60°$), and large ($60° < \psi \leq 90°$). An equal number of images are randomly selected from each category to be reconstructed using our method and 3DDFA-V2 [8] separately. To facilitate presentation and comparison, we then re-project the reconstructed models onto the corresponding input images. In total, 120 pairs of generated images are created and evenly distributed into 10 groups. In total, we have 135 participants, each choosing a group at random and answering 12 questions. For each question, participants should decide which generated image represents the input image more closely, or it is difficult to choose one or the other. The results show that on average our method is preferred when compared with 3DDFA-V2

Table 3 Reconstruction error of different methods on the NoW dataset and the dimensions of the 3DMM parameters regressed by each method

Method	Median (mm)	Mean (mm)	Std (mm)	Dimensions of 3DMM parameters for regression						
				Shape	Texture	Expression	Pose	Camera	Light	Total
3DMM-CNN [45]	1.84	2.33	2.05	✓	✓					198
UMDFA[a] [33]	1.52	1.89	1.57	–						
PRNet[a] [6]	1.50	1.98	1.88	–						
Dib et al. [46]	1.26	1.57	1.31	✓	✓	✓		✓	✓	484
3DDFA-V2 [8]	1.23	1.57	1.39	✓		✓	✓			62
Deng et al. [26]	1.23	1.54	1.29	✓	✓	✓		✓	✓	239
RingNet [27]	1.21	1.53	1.31	✓		✓	✓	✓		159
Our method	1.29	1.63	1.41	✓		✓	✓			62

[a]UMDFA [33] and PRNet [6] use image-to-image methods instead of parameter regression

Table 4 NME (%) comparison of images with different yaw angles ψ on AFLW2000-3D

Method	$0° \leq \psi \leq 30°$	$30° < \psi \leq 60°$	$60° < \psi \leq 90°$	Mean
3DDFA [5]	3.78	4.54	7.93	5.42
3DDFA + SDM [5]	3.43	4.24	7.17	4.94
CMD [23]	–	–	–	3.90
SPDT [24]	3.56	4.06	4.11	3.88
3DDFA-TPAMI [7]	2.84	3.57	4.96	3.79
3D-FAN [41]	3.15	3.53	4.60	3.76
PRNet [6]	2.75	3.51	4.61	3.62
3DDFA-V2 without svs.[a] [8]	2.75	3.49	4.53	3.59
3DDFA-V2 [8]	2.63	3.42	4.48	3.51
Our method	2.76	3.46	4.49	3.57

[a]svs. denotes short-video-synthesis strategy

[8] (39.56% vs. 35.58%). The remaining 24.37% of participants think both methods are equally effective, while 0.49% say neither is similar. It is generally accepted that our method is better for reconstruction of the mouth region. However, in some examples, although the reconstruction of the facial features by our method fits the input image better, the reconstruction of the facial contours is not as good as by 3DDFA-V2 [8].

4.4.3 Model complexity and running time

Since our method is based on a lightweight network, the number of parameters in our network is only 7.6 M, with an input of a 120×120 size image and an output of 62-dimensional parameters. It is significantly less than other commonly used neural networks, such as ResNet-50, which has approximately 23.6 M parameters with the same input size and output size. It is worth mentioning that the multiply–accumulate operations (MACs) of our network are 298.0 M, compared with 6190 M of PRNet [6]. Table 5 shows the model sizes of different methods. We note that the model size of our method is approximately 29 MB, which is much smaller than most other methods. The model size of 3DDFA-V2 [8] is smaller

than that of ours, because they only use MobileNet-V1, while we introduce an attention mechanism and a GCN. Nevertheless, our method enables fast reconstruction even when the number of parameters increases. As shown in Table 5, we compare the inference speed of our method and 3DDFA-V2 [8] on a personal laptop with an NVIDIA GeForce GTX 1650 Ti GPU and an AMD Ryzen 7 4800H with Radeon Graphics CPU @ 2.90 GHz. Our method takes 4.1 ms on a GPU or 20.7 ms on a CPU to regress 3DMM parameters, increasing by only 0.8 ms on the GPU and 6.3 ms on the CPU compared with 3DDFA-V2 [8].

5 Conclusion

In this study, we propose a learning-based method that aims to achieve fast and robust 3D face reconstruction from a single image. We combine the lightweight network, attention mechanism, and GCN, and demonstrate the performance of the combination in experiments. The method not only demonstrates the improvement in reconstruction speed and memory consumption but also guarantees the robustness of the reconstructed 3D models. In inference, only a single image is

Table 5 Model size and running time of different methods

Method	Model size (MB)	Running time
PRNet [6]	153	–
CMD [23]	93	–
Deng et al. [26]	92	–
SADRNet [25]	60	–
3DDFA-V2 [8]	13	3.3 ms (GPU)/14.6 ms (CPU)
Our method	29	4.1 ms (GPU)/20.9 ms (CPU)

required as input, and no landmarks or other information is needed.

However, there are some limitations to our method. Since we cut a large number of 3DMM parameters, our method improves the speed of reconstruction but sacrifices the accuracy of the reconstructed model. If the parameter size is expanded, more refined 3D face models will be obtained. Furthermore, the albedo of the face and the illumination of the images have not been considered in our work. These are the steps we will take next. In the future, we will focus on reconstructing more detailed and realistic face models in a fast way, and try to put them into actual production, e.g. as an auxiliary tool for 3D animation or face recognition.

Author Contributions ZD and YL took part in conceptualization; ZD, JL, and XW were involved in methodology; ZD and YH was responsible for software; ZD, YL, JL, and YH carried out formal analysis; ZD wrote and prepared the original draft; YL and JP took part in writing—review and editing, funding acquisition, and supervision; ZD, YL, and JP contributed to resources.

Code availability Some or all of the code used during the study is available on request from the corresponding author.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003). https://doi.org/10.1109/TPAMI.2003.1227983

2. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 296–301 (2009)

3. Liu, L., Chen, S., Chen, X., Wang, T., Zhang, L.: Fuzzy weighted sparse reconstruction error-steered semi-supervised learning for face recognition. Vis. Comput. **36**(8), 1521–1534 (2020). https://doi.org/10.1007/s00371-019-01746-y

4. Bahroun, S., Abed, R., Zagrouba, E.: Deep 3D-LBP: CNN-based fusion of shape modeling and texture descriptors for accurate face recognition. Vis. Comput. (2021). https://doi.org/10.1007/s00371-021-02324-x

5. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 146–155 (2016)

6. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision–ECCV 2018, vol. 11218, pp. 557–574. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_33

7. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: a 3D total solution. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 78–92 (2019). https://doi.org/10.1109/TPAMI.2017.2778152

8. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3D dense face alignment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision–ECCV 2020, pp. 152–168. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_10

9. Jin, H., Wang, X., Lian, Y., Hua, J.: Emotion information visualization through learning of 3D morphable face model. Vis. Comput. **35**(4), 535–548 (2019). https://doi.org/10.1007/s00371-018-1482-1

10. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3D shape regression for real-time facial animation. ACM Trans. Graph. **32**(4), 1–10 (2013). https://doi.org/10.1145/2461912.2462012

11. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '99, pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999). https://doi.org/10.1145/311535.311556

12. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings, San Diego (2015)

13. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings, Toulon (2017)

14. Cao, Chen, Weng, Yanlin, Zhou, Shun, Tong, Yiying, Zhou, Kun: FaceWarehouse: a 3D facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. **20**(3), 413–425 (2014). https://doi.org/10.1109/TVCG.2013.249

15. Piotraschke, M., Blanz, V.: Automated 3D face reconstruction from multiple images using quality measures. In: 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3418–3427 (2016)

16. Jin, H., Wang, X., Zhong, Z., Hua, J.: Robust 3D face modeling and reconstruction from frontal and side images. Comput. Aided Geom. Des. **50**, 1–13 (2017). https://doi.org/10.1016/j.cagd.2016.11.001

17. Jiang, L., Zhang, J., Deng, B., Li, H., Liu, L.: 3d face reconstruction with geometry details from a single image. IEEE Trans. Image Process. **27**(10), 4756–4770 (2018). https://doi.org/10.1109/TIP.2018.2845697

18. Liu, P., Yu, Y., Zhou, Y., Du, S.: Single view 3D face reconstruction with landmark updating. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 403–408 (2019). https://doi.org/10.1109/MIPR.2019.00082

19. Aldrian, O., Smith, W.: A linear approach to face shape and texture recovery using a 3D morphable model. In: Proceedings of the British Machine Vision Conference 2010, pp. 75–17510. British Machine Vision Association, Aberystwyth (2010). https://doi.org/10.5244/C.24.75

20. Aldrian, O., Smith, W.A.P.: Inverse rendering of faces with a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **35**(5), 1080–1093 (2013). https://doi.org/10.1109/TPAMI.2012.206

21. Schönborn, S., Forster, A., Egger, B., Vetter, T.: A Monte Carlo strategy to integrate detection and model-based face analysis. Pattern Recognit. **8142**, 101–110 (2013). https://doi.org/10.1007/978-3-642-40602-7_11

22. 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, Stanford (2016). https://doi.org/10.1109/3DV.2016.56

23. Zhou, Y., Deng, J., Kotsia, I., Zafeiriou, S.: Dense 3D face decoding over 2500fps: joint texture & shape convolutional mesh decoders. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1097–1106. IEEE, Long Beach (2019). https://doi.org/10.1109/CVPR.2019.00119

24. Piao, J., Qian, C., Li, H.: Semi-supervised monocular 3D face reconstruction with end-to-end shape-preserved domain transfer. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9397–9406. IEEE, Seoul, Korea (South) (2019). https://doi.org/10.1109/ICCV.2019.00949

25. Ruan, Z., Zou, C., Wu, L., Wu, G., Wang, L.: SADRNet: self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. IEEE Trans. Image Process. **30**, 5793–5806 (2021). https://doi.org/10.1109/TIP.2021.3087397

26. 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: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 285–295. IEEE, Long Beach (2019). https://doi.org/10.1109/CVPRW.2019.00038

27. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3D face shape and expression from an image without 3D supervision. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7755–7764. IEEE, Long Beach (2019). https://doi.org/10.1109/CVPR.2019.00795

28. Lin, J., Yuan, Y., Shao, T., Zhou, K.: Towards high-fidelity 3D dace reconstruction from in-the-wild images using graph convolutional networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5890–5899. IEEE, Seattle (2020). https://doi.org/10.1109/CVPR42600.2020.00593

29. Lee, G.-H., Lee, S.-W.: Uncertainty-aware mesh decoder for high fidelity 3D face reconstruction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6099–6108. IEEE, Seattle (2020). https://doi.org/10.1109/CVPR42600.2020.00614

30. Gao, Z., Zhang, J., Guo, Y., Ma, C., Zhai, G., Yang, X.: Semi-supervised 3D face representation learning from unconstrained photo collections. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1426–1435. IEEE, Seattle (2020). https://doi.org/10.1109/CVPRW50498.2020.00182

31. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems, vol. 29 (2016)

32. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision–ECCV 2018, vol. 11207, pp. 725–741. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_43

33. Koizumi, T., Smith, W.A.P.: "Look ma, no landmarks!"—unsupervised, model-based dense face alignment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision—ECCV 2020, vol. 12347, pp. 690–706. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_41

34. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications (2017). Preprint at arXiv:1704.04861

35. Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: Computer Vision—ECCV 2020, vol. 12366, pp. 649–665. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_39

36. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 397–403. IEEE, Sydney (2013). https://doi.org/10.1109/ICCVW.2013.59

37. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: In Computer Vision and Pattern Recognition (2012)

38. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940 (2013). https://doi.org/10.1109/TPAMI.2013.23

39. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: 2013 IEEE International Conference on Computer

Vision Workshops, pp. 386–391. IEEE, Sydney (2013). https://doi.org/10.1109/ICCVW.2013.58

40. Messer, K., Matas, J., Kittler, J., Jonsson, K.: XM2VTSDB: the extended M2VTS database. In: Second International Conference on Audio and Video-based Biometric Person Authentication, pp. 72–77 (1999)

41. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks). In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1021–1030. IEEE, Venice (2017). https://doi.org/10.1109/ICCV.2017.116

42. Kostinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151. IEEE, Barcelona (2011). https://doi.org/10.1109/ICCVW.2011.6130513

43. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738. IEEE, Santiago (2015). https://doi.org/10.1109/ICCV.2015.425

44. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929. IEEE, Las Vegas (2016). https://doi.org/10.1109/CVPR.2016.319

45. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1493–1502. IEEE, Honolulu (2017). https://doi.org/10.1109/CVPR.2017.163

46. Dib, A., Thebault, C., Ahn, J., Gosselin, P.-H., Theobalt, C., Chevallier, L.: Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)

47. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision—ECCV 2016, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Zhuoran Deng is currently a graduate student in the School of Software, South China Normal University, China, from which she received her bachelor's degree in software engineering in 2020. Her current research interests focus on the fields of image processing and deep learning. One particular interest is image-based 3D face reconstruction.
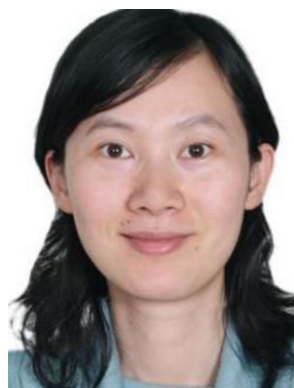
Yan Liang received the B.S. degree in automation control, the M.S. degree in signal and information processing, and the Ph.D. degree in control science and engineering from Guangdong University of Technology, China, in 2002, 2005, and 2017, respectively. In 2005, she joined South China Normal University, China, and she is now a lecturer. Her current research interests include image processing, face recognition, emotion recognition, and computer network.

Jiahui Pan received the B.S. degree in computer science in 2005 and the M.S. degree in computer application technology in 2008 from South China Normal University, China. He received his Ph.D. in pattern recognition and intelligent systems in 2014 from South China University of Technology, China. In 2008, he joined South China Normal University, China, where he is now a full Professor. His current research interests include the fields of signal processing and pattern classification.

Jiacheng Liao is currently a graduate student in the School of Software, South China Normal University, China, from which he received his bachelor's degree in Software Engineering in 2019. His current research interests include the fields of signal processing, pattern recognition, and video understanding. One particular interest is emotion recognition.

Yan Hao is currently a graduate student in the School of Software, South China Normal University, China. She is majoring in Software Engineering. Her current research interests focus on facial micro-expression recognition, deep affect prediction in the wild, and information extraction, and all studies are achieved based on the video data.

Xing Wen is pursuing his master's degree in the School of Software, South China Normal University, China, from which he received his bachelor's degree in Software Engineering in 2020. His current research interests include facial expression recognition and domain adaptation.