Hindawi
Journal of Sensors
Volume 2021, Article ID 9921101, 13 pages
https://doi.org/10.1155/2021/9921101

Research Article

Detail 3D Face Reconstruction Based on 3DMM and Displacement Map

Tianping Li, Hongxin Xu, Hua Zhang, and Honglin Wan

Key Laboratory of Medical Physics and Image Processing in Shandong Province, Shandong Normal University, School of Physics and Electronics, Jinan, Shandong, China

Correspondence should be addressed to Hua Zhang; jnzhhua@126.com

Received 12 March 2021; Revised 19 April 2021; Accepted 8 June 2021; Published 25 June 2021

Academic Editor: Aijun Yin

How to accurately reconstruct the 3D model human face is a challenge issue in the computer vision. Due to the complexity of face reconstruction and diversity of face features, most existing methods are aimed at reconstructing a smooth face model with ignoring face details. In this paper a novel deep learning-based face reconstruction method is proposed. It contains two modules: initial face reconstruction and face details synthesis. In the initial face reconstruction module, a neural network is used to detect the facial feature points and the angle of the pose face, and 3D Morphable Model (3DMM) is used to reconstruct the rough shape of the face model. In the face detail synthesis module, Conditional Generation Adversarial Network (CGAN) is used to synthesize the displacement map. The map provides texture features to render to the face surface reconstruction, so as to reflect the face details. Our proposal is evaluated by Facescape dataset in experiments and achieved better performance than other current methods.

1. Introduction

Face is one of the most important biological characteristics of human beings, and face modeling is often used in security, animation, biometrics, and other fields [1, 2]. In recent years, due to the limitations of 2D images, the research of human face has gradually shifted from 2D plane images to 3D space models.

The steps of 3D face reconstruction are very complex if it is reconstructed step by step. Moreover, this reconstruction model will lead to more data loss and less accuracy. To stress this issue, one-step reconstruction model is presented (see Figure 1). The reconstruction system is divided into two parts: the initial face reconstruction module and the face detail synthesis module, and both are based on deep learning [3]. The initial face reconstruction module is mainly responsible for face alignment. The supervised learning method is used to train 60K face images from 300W-LP dataset to obtain the corresponding dictionary. In this process, a CNN network is used to align the negative faces and detect their feature points. The feature points are input into the principal component analysis- (PCA-) based 3DMM [4] to obtain a rough face shape. The face detail synthesis module is based on CGAN, which inputs the original image to synthesize the displacement map, and the displacement map retains the more complete details of the face [5]. The face detail synthesis module refers to DFDN to train high-quality images and get the training data, which can synthesize the displacement map from the original image.

In this paper, we propose a reconstruction system to recover the details of the face model. Our reconstruction system can better solve the problem of face pose reconstruction and facial expression reconstruction from the input image. The facial detail synthesis module of the reconstruction system can extract facial features from the input image and synthesize the displacement map containing most of the details of the target face. Compared with the initial shape model, the detail face model with displacement map has better visual effect and more accurate data.

The rest of the paper is organized as follows. Section 2 describes the researchers' related work on 3D face. Section 3 describes the initial face reconstruction module. Section 4

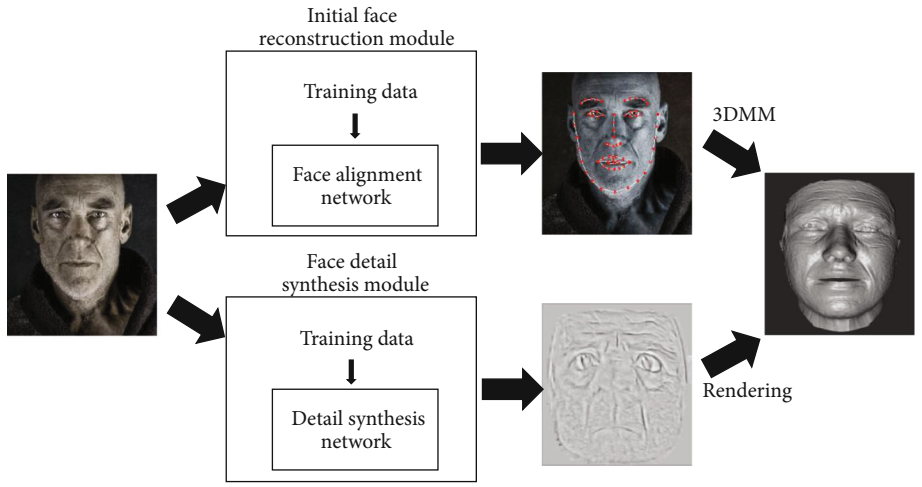

Figure 1: Reconstruction system structure.

describes the face detail synthesis module. Section 5 is the experiment and analysis. Section 6 concludes the paper.

2. Related Work

With the application of deep learning methods to graphics, the transition from 2D planar images to 3D spatial models has become one of the popular research directions. Blanz et al. proposed the concept of 3DMM and obtained a Basel Face Model (BFM) by training the objects and related data collected by the depth camera [6]. The parameterized BFM has the universal characteristics of a human face, and a deformed 3D model can be obtained by inputting shape, texture, and attribute parameters. A large number of 3DMM-based algorithms have been proposed. Tran et al. proposed a nonlinear 3D face deformation model method [7], which used a large number of unconstrained pictures as training objects to train a new architecture of 3DMM without using 3D scanning equipment. Galteri et al. used CGAN to refine 3DMM [8].

In addition to traditional 3DMM, an end-to-end method based on deep learning can also better reconstruct 3D face models. The end-to-end method can perform face alignment on the input face image. In the vector space, the detected feature points are mapped to the face model of the dense point cloud one by one. This method is simple and fast. Compared with the traditional 3DMM, its accuracy is higher in most cases. Yao et al. designed PRNet [9] based on the CNN network structure and deep residual network and used the UV vector space to complete the mapping of the 3D face model. Jackson et al. proposed a combination of 3DMM and CNN VRN [10] to reconstruct the model of nonfrontal face images. Tran et al. used the end-to-end neural network to reconstruct the details of extreme face [11].

The rendering of face model is also a key part. Ranjan et al. proposed COMA method to generate the head network and used an MPI-IS Mesh Processing Library for rendering [12]. MPI-IS Mesh Processing Library is an efficient 3D model rendering tool. Li et al. designed a Flame model to render the basic shape and expression of the face model [13]. Sanyal et al. proposed RingNet based on Flame [14]. RingNet can reconstruct the head model by inputting face image and can better simulate the facial expression. A deep 3D face reconstruction method was proposed by Deng et al. [15]. This method is based on 3DMM and coarse facial expression [16], and the rendered model is more accurate.

3. Initial Face Reconstruction Module

The initial face reconstruction module is the key module in proposed reconstruction system. This module outputs the input face image directly to the rough initial face model, which includes pose face alignment, feature point detection, and model fitting.

3.1. Construction of Rough Face Model

3.1.1. Face Alignment. In our method, the feature point coordinates are used as the input of 3DMM based on the PCA algorithm to construct a parameterized model.

Because the manual labeling is time-consuming and labor-intensive, and the traditional feature point detection is poor in robustness and accuracy, we use a CNN to deal with the face alignment of nonfrontal face images. This article uses the DLIB library to detect feature points. DLIB library uses regression tree set cascade [17] to generate feature point model through supervised learning and training image sets with feature point annotations. Input an image, the algorithm will generate the initial shape based on the target face and roughly estimate the location of the feature points. Then, a gradient boosting algorithm is used to reduce the error between the initial shape and the real landmark, and the least square method is used to minimize the error to obtain the cascade r_t of each stage.

$$\widehat{S}^{t+1} = \widehat{S}^t + r_t\left(I, \widehat{S}^t\right), \tag{1}$$

where t is the number of cascade regressions, $\widehat{S}^t$ is the shape vector of the tth secondary cascade regression, and I is the

Figure 2: Feature point detection.

input image. The key point of the cascade is that the regressor r_t predicts according to the image pixel intensity value and indexes it relative to the current shape vector $\widehat{S}^t$. The feature points of a nonfrontal face image are divided into two parts: visible and invisible. Since the latter is difficult to predict, deep learning methods can effectively deal with this problem.

We train 60K face images with face deflection angle data and feature point coordinate data in the 300W-LP dataset [18] to obtain a dictionary. Through the index dictionary, the output finds the index target that is closest to the deflection angle of the input face image. In addition, referring to the weight setting of the main components of the human face by referring to the PRNet, the feature points in the vicinity of the eyes, nose, and mouth are given greater weights to highlight the changes and recognition of the model

$$l_{fp} = M_{face} * W_{fp}, \tag{2}$$

where M_{face} detects the coordinates of face feature points and W_{fp} is the weight. Figure 2 shows an example of feature point detection.

3.1.2. 3DMM Face Reconstruction. A rough face model with a smooth surface is relatively average, without too much facial detail but contains most of the depth information of the face. Inputting the face image fitting model will change the vertex, and the topological network of the BFM will average face model. The method in this paper employ BFM2017 [19] to fit a 3D face with less detail.

Taking the original image as input, assuming that the grid vertex coordinates of the 3D model are $S = (x_i, y_i, z_i)$. The feature points according to Equation (2) are used to calculate the PCA parameters. According to [6], n shape vectors of the initial face model is

$$S_{\text{shape}} = \sum_{i}^{n} \alpha_i S_i, \tag{3}$$

where α_i is the shape weight coefficient.

According to the average face shape $\bar{S}$ obtained from the training set of 200 images, the difference between the shape of each face model and the average face shape $\Delta S_i = S_i - \bar{S}$ calculates the covariance matrix C_S of the shape vector. Through PCA, the orthogonal coordinate system formed by the eigenvector s_i of C_S is transformed into the basis:

$$S_{\text{mod el}} = \bar{S} + \sum_{i=1}^{m-1} \alpha_i s_i. \tag{4}$$

Due to the universality of main features of human face, the distribution of shape vector parameter α_i is normal

distribution (as shown in Equation (5)). Texture parameters are similar to shape parameter.

$$p(\alpha) \sim \sim \exp\left(-\sum_{i=1}^{m-1}\frac{\alpha_i^2}{2\sigma^2}\right). \tag{5}$$

For the shape parameter α, texture parameter β, and attribute parameter γ, the RGB vector of the projected image of the reconstruction model is

$$I_{\text{mod el}}(x,y) = \left(I_r(x,y), I_g(x,y), I_b(x,y)\right)^T. \tag{6}$$

The error between the projected image of the reconstructed model and the input image is

$$E_{\text{image}} = \sum_{x,y}\left\|I_{\text{input}}(x,y) - I_{\text{mod el}}(x,y)\right\|^2. \tag{7}$$

Matching the input face image with the 3D modeled face is an ill-posed problem. In the vector space of the face model, the matching quality and a priori method can be used to obtain the solution with constraints [6]. Similar to Equation (5), $p(\alpha)$ and $p(\beta)$ obey normal distribution, $p(\gamma)$ is obtained by the point-to-point method. According to Bayesian decision, the input image can be obtained through the maximum posterior probability with the parameters (α, β, γ), and the model I_{model} is reconstructed through the three parameters. But under the influence of noise, the observed image I_{input} will be disturbed.

Assuming the standard deviation σ_G of the Gaussian noise of the observed image, the parameter probability of the observed image is

$$p(\alpha,\beta,\gamma) = \exp\left(-\frac{1}{2\sigma_G^2}\cdot E_1\right). \tag{8}$$

The posterior probability of the parameter is expressed by minimizing the cost function:

$$E = \frac{1}{\sigma_G^2}E_1 + \sum_{i=1}^{m-1}\frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_{i=1}^{m-1}\frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i\frac{(\gamma_i-\bar{\gamma}_i)^2}{\sigma_{\gamma,i}^2}. \tag{9}$$

3.2. Camera Model

3.2.1. Weak Perspective Projection Function. To visualize a 3D model, the topology of the 3D model needs to be projected onto a two-dimensional plane. Compared with orthogonal projection, perspective projection can freely set the reduction and enlargement of the projected image.

During the projection process, it may appear that the dense 3D coordinates are superimposed on the 2D coordinate points of the projection surface due to dimensionality reduction. Aiming at the projection of the pose face model, this paper uses a weak perspective projection function similar to the perspective projection function to deal with the problem of projecting a 3D model onto a 2D plane [20]. Figure 3 explains the difference between orthographic projection and weak perspective projection.

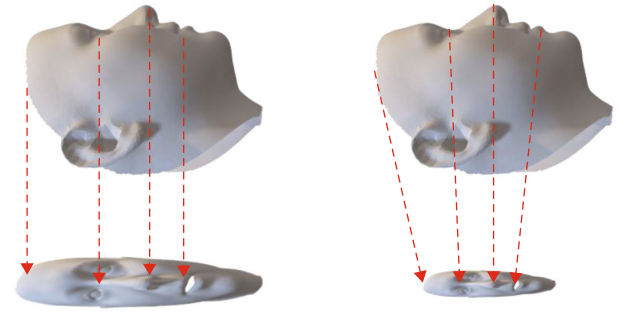

Figure 3: Orthogonal projection and weak perspective projection.

In this paper, assuming that the positive direction of the camera model is the weak perspective projection in the z direction, referring to [21], we use the orthogonal projection matrix $\Lambda \in \mathbb{R}^{3\times3}$ and the target displacement calibration x to design the weak perspective projection function:

$$P = f * \Lambda * R * S_{\text{mod el}} + x. \tag{10}$$

Optimizing Equation (10),

$$P = \sum_{i=1}^{n}\|f * \Lambda * R * \alpha_i S_i\| + x_i, \tag{11}$$

where f is the focal length ratio, R is the rotation matrix, and x_i is the displacement coefficient of the ith vertex coordinate. The weak perspective projection function projects the normalized face mesh vertices from the 3D space to the 2D plane, which is convenient for subsequent operations and processing.

Minimize the error of projecting the initial reconstruction model based on the PCA algorithm to the plane [22]:

$$E = \sum_{\text{k}}\omega_{\text{k}}\|L_{\text{k}}\text{-}P(\text{l}_{\text{k}}(\alpha,\beta))\|_2 + \lambda_k\|\alpha\|_2, \tag{12}$$

where L_k is the kth feature point of the planar face, $l_k(\alpha, \beta)$ is the coordinates of the kth vertex of the 3D model, ω_k is the weight of the kth feature point, and λ_k is the regularization coefficient of the shape parameter.

3.2.2. Hidden Surfaces Remove. In the dense 3D mesh, under nonfrontal face conditions, some vertices will always overlap, which affects the result and accuracy of feature point acquisition. In this paper, the z-buffer algorithm [23] is used to solve the ambiguity of the depth value.

The z-buffer algorithm buffers the depth value of the visible surface into the depth buffer area, and the depth value of the hidden surface is removed. So, the single view only has the depth of the visible surface. The depth z value is not the true Euclidean distance of the Cartesian space coordinate system, but a relative measure of the distance from the vertex to the viewpoint. Assuming that the model is viewed from the perspective of the z-axis as the positive direction, the projection surface is the xoy plane. (x, y) is the coordinates of each

(a) (b) (c)

Figure 4: The difference between using z-buffer depth map. (a) Original image. (b) Depth map using z-buffer. (c) Depth map without z-buffer.

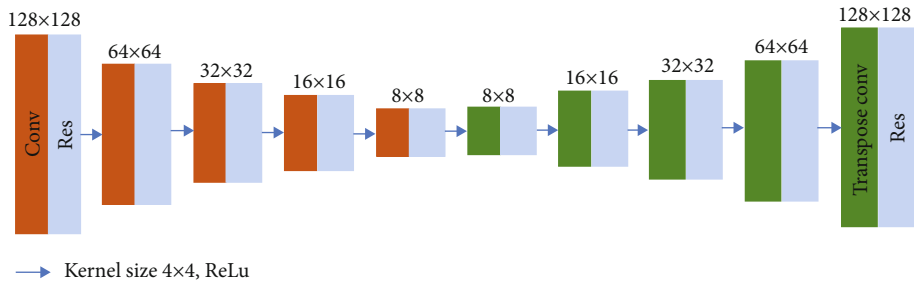

Figure 5: Face alignment network structure.

pixel in the overlapping area of the projection surface. The ray parallel to the z-axis are the depth values, which are z_1 and z_2, respectively, and the maximum of (z_1,z_2) is stored in the z-buffer.

Figure 4 demonstrates the difference between whether to use the z-buffer depth map. The depth map of the depth map using the z-buffer algorithm is distinct, and there is no ambiguity in the depth value due to the posture self-occlusion caused by a single perspective.

3.3. Face Alignment Network. The purpose of the face alignment is to obtain a dictionary through training. The input face image is indexed after face detection, etc., and then, the angle of the target face relative to the frontal perspective can be obtained, and the target face can be aligned [24]. The face alignment obtains the angle of the target face with the feature point detection of DLIB library and the improved feature point loss function. When the input is a face image with a large pose, not only the visible feature points can be accurately detected but also the feature points that are invisible due to the posture self-occlusion can be predicted more accurately.

Test an image in the test set n times, and take the average value of its location map feature points. Improved loss function L_{fp}:

$$L_{fp} = \sum \left\| \bar{S} - S_{GT} \right\| * W_{fp}, \tag{13}$$

where $\bar{S}$ is the average value of n tests of the feature point landmarks of the location map, S_{GT} is the real landmark, and W_{fp} is the weight of the feature points.

The face alignment network is a CNN architecture based on the residual network [25, 26], composed of 10 residual modules. Figure 5 is a diagram of the face alignment network structure.

When using the face alignment network training and the angle of the training set image corresponding to the annotation, 3D point cloud and additional parameters are used as the training object, and the projection normalized coordinate code (PNCC) feature [18] that can represent the shape of the model is used to generate a dictionary.

PNCC is composed of normalized coordinate code (NCC) and z-buffer algorithm. NCC normalizes the coordinates $c = (x, y, z)$ of the vertices of the 3D average face model, and its calculation formula is

$$\mathrm{NCC} = \frac{\bar{S}_c - \min\left(\bar{S}_c\right)}{\max\left(\bar{S}_c\right) - \min\left(\bar{S}_c\right)}. \tag{14}$$

The purpose of PNCC is to use z-buffer algorithm to remove the hidden surface normalized by NCC to achieve the effect of projection. PNCC calculation formula:

$$\mathrm{PNCC} = z - \mathrm{buffer}(V_{3D}(\rho), \mathrm{NCC}), \tag{15}$$

Figure 6: Examples of texture map, normal map, and displacement map.

where $V_{3D}(\rho)$ is the 3D surface after projection and ρ is a model parameter.

4. Face Detail Synthesis Module

In the initial face reconstruction module, although the 3DMM reconstruction model based on the PCA algorithm has most of the information of the reconstruction target, it loses part of the detailed information due to dimensionality reduction. We use a face detail synthesis module to make up for face detail information.

4.1. Displacement Map Based on Texture Bump. The details of the face include gullies and wrinkles, so it is difficult to detect and extract them with a unified standard method. Undifferentiated detection and detail extraction integration can effectively solve this problem. We uses the deep learning method to build a detailed synthesis network, which detects the face in the image and extracts the texture map of the face area, and synthesizes a displacement map based on the texture map.

The displacement map is similar to the normal map. Normal map highlights the unevenness of the model. The normal map represents the normal vector corresponding to the vertices, but cannot change the vertex coordinates of the model itself. Since all the details are only reflected in the map, the displacement map can use micropolygon tessellate [27] to change details of the model surface. For a 3DMM composing of triangular meshes, first, inlay a triangular structure with the same size as the image pixel size on the effective area of the model. The bump map is grayed out, and the depth z coordinate is determined by the gray level. Then, according to the triangle mesh obtained by mosaic, the vertices are moved along the original surface normal direction. Then, determine the new normal vector for the new mesh vertex.

The lower x and y of the model's three-dimensional coordinates are represented by the uv coordinates of the texture, i.e., the image color. z coordinate is represented by the gray scale of the displacement map. The depth information of the shifted texture obtained by graying the texture is incomplete. The reason is that for face images, some face details may be treated as noise, or the depth of some details is too similar to the main area of the face, resulting in a large deviation of the model.

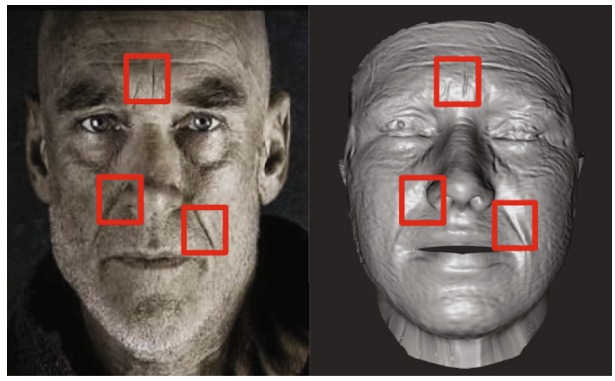

Figure 7: Rendered detailed model.

Our method proposes a detailed synthesis network based on the gray-scale displacement map, and the subtle details of the face are used as noise to extract the difficult-to-handle details of the texture map from the generator. The extracted detail noise is used as a feature map and cyclically synthesized to a displacement map. According to the gray value, the depth of the model is changed in a small amount to highlight the details. The pixels of the synthesized texture are $4096 * 4096$, and there are more pixels corresponding to details, which is more convenient for processing. The three images in Figure 6 are the RGB texture map, the normal map, and the displacement map.

In Figure 7, the red frame area of the detail model reconstructed by the method in this paper represents three details from small to large depth. Rendering the displacement map to the model can clearly see that the fit of different degrees of detail is relatively good.

4.2. Facial Expression Process. The recognition and fitting of facial expressions is a key problem that needs to be solved in the field of 3D face reconstruction. The dynamic changes and severity of the face will affect the analysis of the main components of the face. When projecting, because the 3D space dimensionality reduction will lose part of the information, the facial expression model will appear ambiguity when it is projected onto the 2D plane.

Our method mainly uses the expression fitting function of BFM2017 to realize the dynamic changes of the face. According to Equation (3), on the basis of the neutral expression face shape vector, an additional expression vector e is added to it, i.e.,

$$S_{\text{shape}} = \sum_{i}^{n} \alpha_i S_i + e_i. \tag{16}$$

However, the expression fitting function of BFM2017 mainly changes the mouth vector, and the fitting effect of other face parts is not ideal. Therefore this article uses a semantically defined emotion feature predictor and physical appearance features. The emotion feature predictor is based on deep learning training to obtain the corresponding expression parameters, and the appearance feature is the expression fitting of BFM2017.

Referring to the processing of facial dynamic expressions in DFDN, the emotion feature predictor is trained from a total of 450k images with 11 expressions in the AffectNet dataset [28]. The $e_{\text{predictor}} \in \mathbb{R}^{128}$ used to represent the feature vector of human emotion is obtained by the network training of CNN structure, and the emotion parameters are randomly generated in the standard normal distribution. The emotional feature vector with expression parameters is used to render the emotional image set, and the training set is input to the emotional feature predictor to obtain the feature vector of the face object in the image set [22]. The emotional feature vector is combined with the physical appearance feature to obtain a semantically defined feature vector.

According to the one-to-one correspondence between the feature vector of the image set and the expression parameter, a dictionary is set to represent the mapping of the feature vector to the expression parameter. Input a facial expression image, get its emotion feature vector through the emotion feature predictor, traverse the dictionary, and find the expression parameter closest to this vector.

4.3. GAN-Based Detail Synthesis Network. The Conditional Generative Adversarial Network (CGAN) [29] based on GAN is divided into two parts: generator network and discriminator network. The generator network randomly generates constrained images, and the generated images pass through the discriminator to perform feature threshold discrimination, save valid features, and cycle the generation-judgment process until the discriminator cannot determine the wrong image.

In this article, dealing with 3D face models, the loss function of CGAN is as follows:

$$V_{\text{GAN}}(D, G) = E_{x,y}[\log D(x \mid y)] + E_{x,z}[\log (1 - D(G(x \mid z)))], \tag{17}$$

where x is the input image, y is the feature point, and z is random noise. Refer to [30], optimizing Equation (17):

$$\begin{cases} L_G = V_{\text{CGAN}}(D, G) + \lambda_1 L_1(G), \\ L_D = -V_{\text{CGAN}}(D, G), \end{cases} \tag{18}$$

where L_G is the generator loss function, L_D is the discriminator loss function, $L_1(G)$ is the generator's L_1 loss function, and λ_1 is set to 100.

The U-net model based on improved FCN [31] is a structure including down-sampling and up-sampling, with the purpose of increasing the accuracy of the image. Down-sampling is used to display environmental information, and up-sampling combines the environmental information from down-sampling with the input information of up-sampling to restore detailed information, making the texture of the human face more real.

This network uses the U-net-6 structure and takes the original target image as input, to generate displacement maps from the semantically defined texture structure map. The generator network and the 4-layer fully connected layer constrain the generated data through feature points and calculate the PCA parameters. Except for the fully connected layer, every linear part is activated by the ReLU function. The LeakyReLU function is used to activate between the fully connected layers. The structure of the U-net-6 network generator is shown in Figure 8.

The network discriminator judges the validity of the output image through the threshold. In this paper, the discriminator is based on PatchGAN [32]. The input image is divided into an $N \times N$ matrix, and after convolution, an $m \times m$ matrix is output. The output matrix is averaged, the threshold is judged, and the logical result is output. The network structure of the discriminator is shown in Figure 9.

5. Experiment and Discussion

5.1. Face Alignment Evaluation. The visible and invisible feature points of nonfrontal faces obtained by face alignment will directly affect the subsequent initial face reconstruction. In our evaluation experiment, the normalized mean error (NME) calculated by comparing with real landmarks represents the accuracy of feature points.

For the face alignment experiment, this article uses the 300W-LP dataset as the training set. The dataset contains faces deflection from 0 to 90 degrees, with a total of more than 60K images. Use the DLIB library to detect human faces and crop each image into a $256 \times 256 \times 3$ face image.

Aiming at the accuracy evaluation of the feature points of face poses at different angles, this paper randomly selects 1000 images from 300W-LP dataset. Calculate the average of the normalized mean error (NME) between the 68 detected feature points of the face and the real landmarks to evaluate the accuracy in this paper. In addition, we compare our method with other two advanced face alignment methods PRNet and 3DDFA. The results obtained are shown in Figure 10.

According to Figure 10, compared with the other two methods, our method can get better results in the feature point detection experiment of 300W-LP sample set.

5.2. Reconstruction Evaluation in Constrained Scenarios. For the evaluation of face image reconstruction in constrained scenes, this experiment uses Facescape dataset [33]. Aiming at the evaluation of the 3D model [34], the evaluation

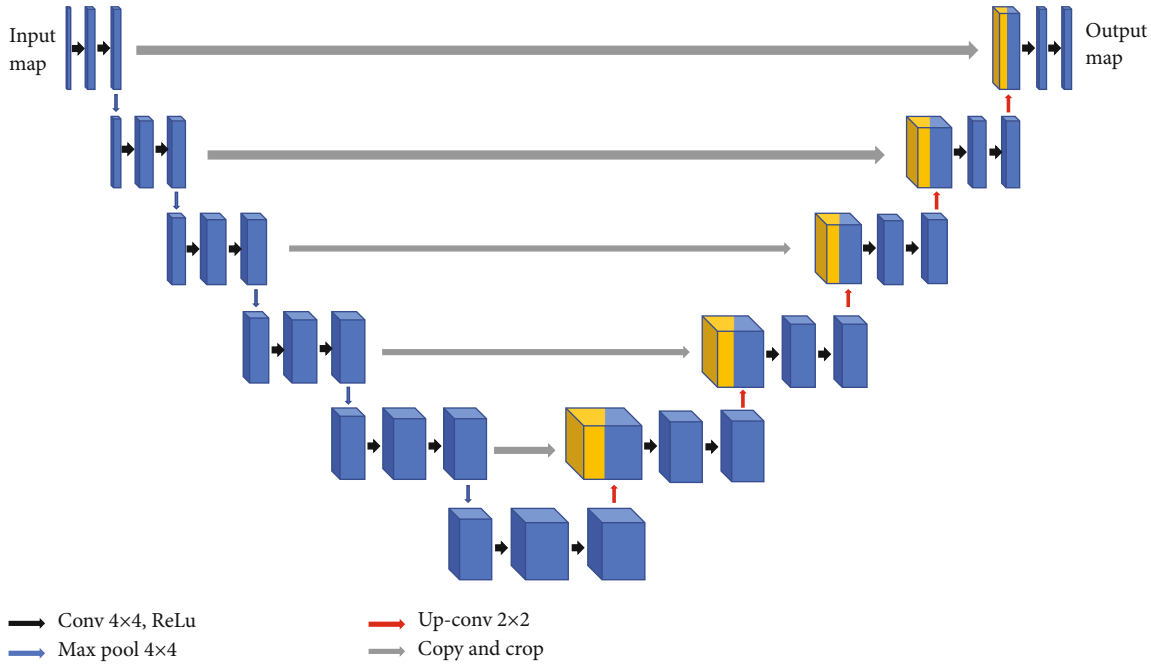

Figure 8: The structure diagram of the detailed synthesis network generator.

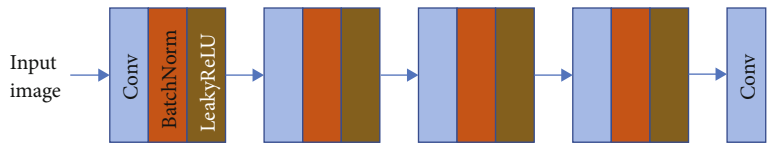

Figure 9: Network structure diagram of the discriminator.

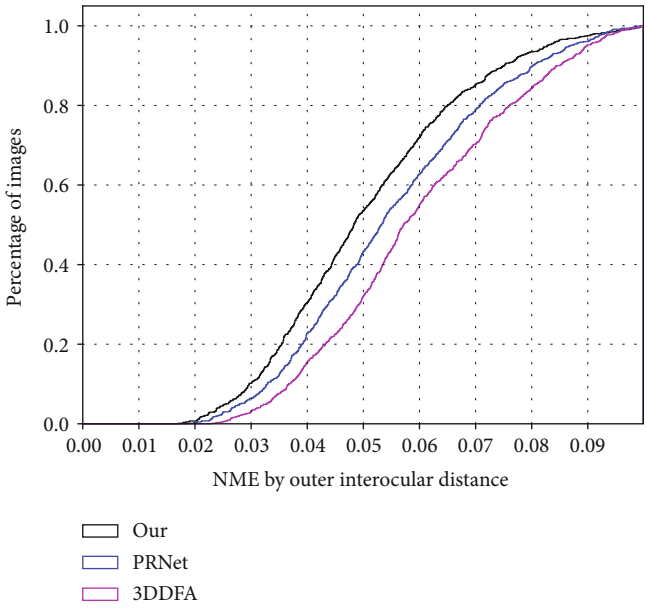

Figure 10: Comparison of cumulative error distribution (CED) curves of NME between sample landmarks and real landmarks.

Table 1: Mean RMSE 1 and mean SD of test set 1.

Method	RMSE 1	SD
Our	3.13	1.87
PRNet	3.64	2.08
3DDFA	4.70	2.96
RingNet	5.12	3.13

experiment in this paper is based on the root mean square error (RMSE) and standard deviation (SD) between the point cloud of the reconstructed model and ground truth. Among them, RMSE is used to evaluate the accuracy of the reconstructed model, and SD is used to assess the degree of dispersion of the point cloud of the reconstructed model itself. In the reconstruction evaluation, the accuracy values of neutral face evaluation, facial expression evaluation, and robustness evaluation are represented by RMSE 1, RMSE 2, and RMSE 3, respectively. The lower the RMSE and SD, the better the accuracy and dispersion of the reconstruction model.

5.2.1. Frontal Face Model Evaluation. In this experiment, the accuracy (RMSE 1) and the discrete value (SD) of the frontal face reconstruction model are used as the evaluation standard. In this evaluation process, 10 frontal face images of

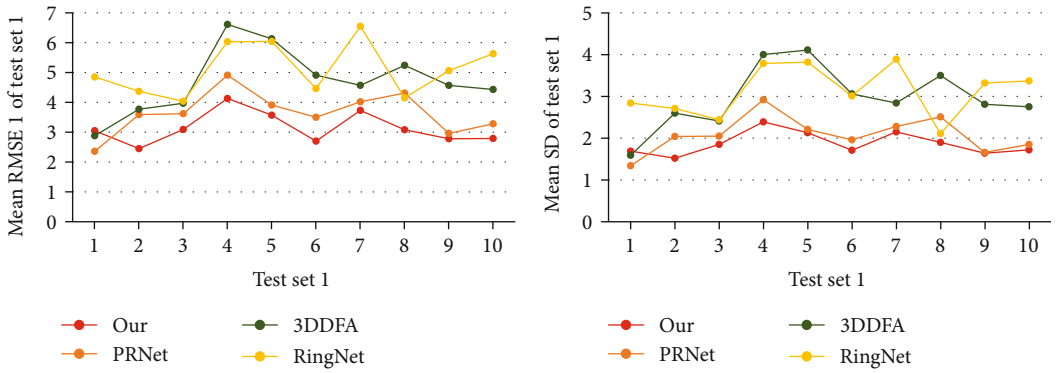

Figure 11: RMSE and SD of smooth model.

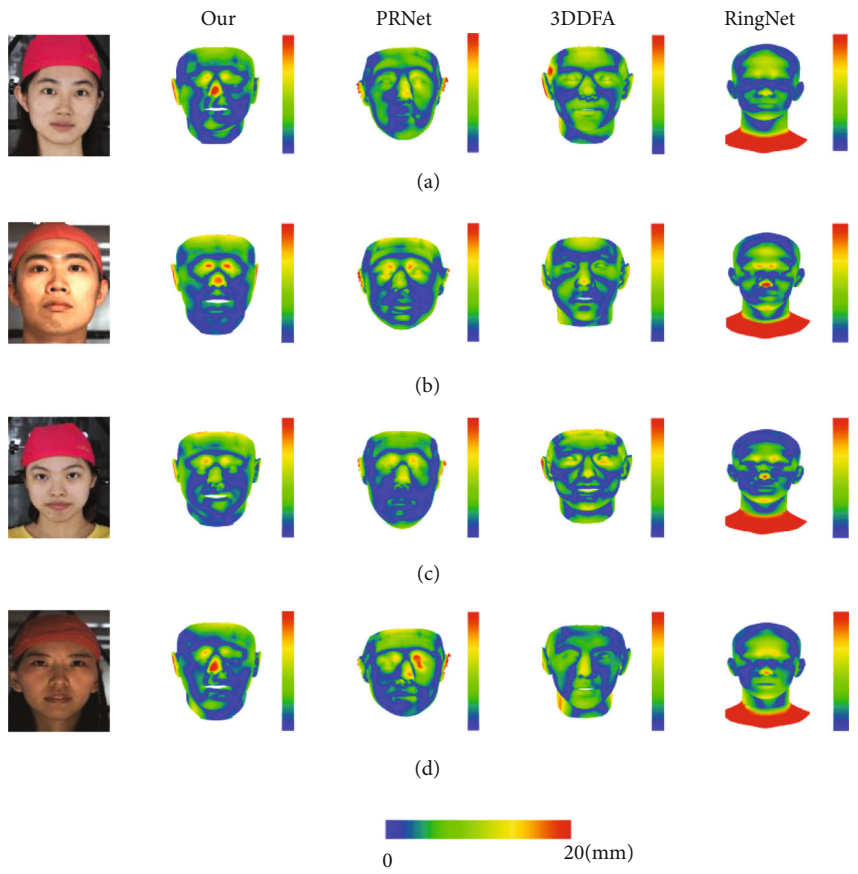

Figure 12: Error heat distribution of reconstructed model and ground truth.

the subject were randomly selected from Facescape dataset as test set 1, and the test set 1 images were reconstructed through the integrated network proposed in this paper, and 10 sets of models were obtained. In addition, this experiment compares our method with three other advanced algorithms, PRNet [9], 3DDFA [18], and RingNet [14].

According to the evaluation standard, the mean RMSE 1 and SD of the 10 groups of reconstruction models are calculated. The data of the test set 1 are shown in Table 1. The detailed data of our method and the model reconstructed by PRNet, 3DDFA, and RingNet are shown in Figure 11.

Based on the above data comparison, our method has higher accuracy and dispersion in reconstructing the frontal neutral face image compared to the other three methods. Figure 12 shows examples of the heat distribution of the sample reconstruction model error.

5.2.2. Frontal Face Model with Expression Evaluation. The difficulty of facial expression reconstruction is often greater than that of neutral expression face reconstruction. We show more reconstruction models of images in unconstrained environment in Figure 13.

Figure 13: Reconstructed models using our method. (a) Target face images. (b) Smooth model without high frequency details. (c) Fine models with high frequency details.

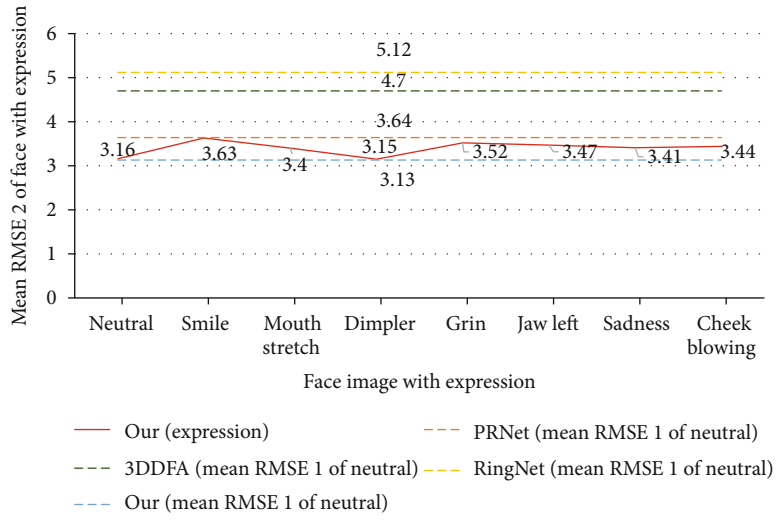

Figure 14: RMSE 2 of expression face model.

In this experiment, Facescape dataset was used to evaluate the reconstruction of facial expressions. Facescape dataset contains the depth information data of 20 facial dynamic expressions of each collected object. Eight dynamic facial expression images of the object are randomly selected from Facescape dataset for reconstruction, and the root mean square error (RMSE2) is calculated (as shown in Figure 14).

In the same method, the accuracy of facial expression reconstruction model is often slightly lower than that of neutral face reconstruction model. In Figure 14, although the RMSE 2 of face models with expression reconstructed by our method is higher than the mean RMSE 1, it is lower than the mean RMSE 1 of the face model with neutral reconstructed by other methods. The accuracy of our method for facial expression reconstruction is significantly higher than that of neutral face reconstruction model of other comparison methods, so our method also has higher advantages in facial expression fitting.

5.2.3. Robustness Evaluation under Noise Environment. In the field of 3D reconstruction, robustness is an important evaluation criterion for reconstruction model algorithms. It can

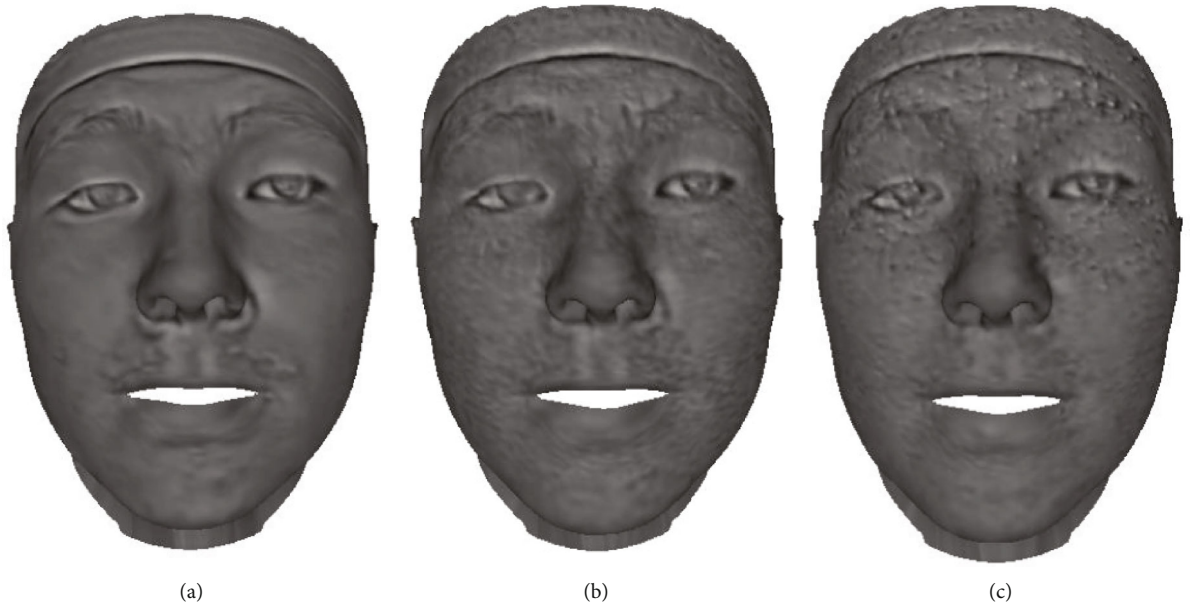

Figure 15: Robustness evaluation example model. (a) Original image reconstruction model. (b) Gaussian noise image reconstruction model. (c) Salt and pepper noise image reconstruction model.

Figure 16: Robustness evaluation.

clearly indicate the degree of adaptation of the algorithm in a complex environment and whether it can reduce the influence of interference factors on model reconstruction. The robustness evaluation in this paper is mainly about face reconstruction under noisy environment. First, randomly select 6 images from the Facescape dataset, and apply Gaussian noise and salt and pepper noise to these 6 images, respectively. As can be seen in Figure 15, an example compares the difference between detail reconstruction model of the original image and detail reconstruction model of the noise image.

The image after applying noise is the test set 2. Then, the original image and the noise image of the test set 2 were reconstructed through the integrated network, compared with the ground truth, and the root mean square error (RMSE 3) was calculated (as shown in Figure 16).

According to the test set 2 of the noise evaluation experiment and the corresponding noise image, the fluctuation interval of the RMSE3 of the noise image reconstruction and the original image reconstruction is (-0.04, 0.18). In addition, there may be a large number of noise points covering the high-frequency details, which will affect the discriminating process of the discriminator of the face detail synthesis module, resulting in the increase of iterations and the slight improvement of the accuracy of the whole model.

6. Conclusion

We propose a reconstruction system for face model. The initial face reconstruction module uses a face alignment network and 3DMM to initially reconstruct a face with a smooth surface. The face detail synthesis network generates

a displacement map, which contains most of the details of the reconstructed object. For facial expressions, we use an emotional feature predictor to fit facial expressions. The three-dimensional sense and accuracy of the detailed face model are better than the 3DMM reconstruction model based on PCA. Through the evaluation of face alignment, accuracy, and robustness in unconstrained scenes, our method obtains ideal results. Compared with other advanced methods, our method also has more advantages.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the Shandong Provincial Natural Science Foundation (ZR2020MF119).

References

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: a survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885–1906, 2007.

[2] Y. Zhang, T. Sim, C. Lim Tan, and E. Sung, "Anatomy-based face reconstruction for animation using multi-layer deformation," *Journal of Visual Languages & Computing*, vol. 17, no. 2, pp. 126–160, 2006.

[3] E. Richardson, M. Sela, and R. Kimmel, "3D face reconstruction by learning from synthetic data," in *2016 fourth international conference on 3D vision (3DV)*, pp. 460–469, Stanford, CA, USA, 2016.

[4] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[5] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, and L. Bao, "Self-Supervised Learning of Detailed 3D Face Reconstruction," *IEEE Transactions on Image Processing*, vol. 29, pp. 8696–8705, 2020.

[6] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pp. 296–301, Genova, Italy, 2009.

[7] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7346–7355, Salt Lake City, UT, USA, 2018.

[8] L. Galteri, C. Ferrari, G. Lisanti, S. Berretti, and A. del Bimbo, "Deep 3D morphable model refinement via progressive growing of conditional Generative Adversarial Networks," *Computer Vision and Image Understanding*, vol. 185, pp. 31–42, 2019.

[9] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Computer Vision – ECCV 2018*, pp. 557–574, Springer, 2018.

[10] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1031–1039, Venice, Italy, 2017.

[11] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni, "Extreme 3D face reconstruction: seeing through occlusions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3935–3944, Salt Lake City, UT, USA, 2018.

[12] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Computer Vision – ECCV 2018*, pp. 704–720, Springer, 2018.

[13] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–17, 2017.

[14] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7763–7772, Long Beach, CA, USA, 2019.

[15] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 285–295, Long Beach, CA, USA, 2019.

[16] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1294–1307, 2019.

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, Columbus, OH, USA, 2014.

[18] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: a 3D solution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 146–155, Las Vegas, NV, USA, 2016.

[19] T. Gerig, A. Morel-Forster, C. Blumer et al., "Morphable face models-an open framework," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 75–82, Xi'an, China, 2018.

[20] A. M. Bruckstein, R. J. Holt, T. S. Huang, and A. N. Netravali, "Optimum fiducials under weak perspective projection," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 223–244, 1999.

[21] F. Mokhayeri, E. Granger, and G. A. Bilodeau, "Domain-specific face synthesis for video face recognition from a single sample per person," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 757–772, 2019.

[22] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu, "Photo-realistic facial details synthesis from single image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9429–9439, Seoul, Korea, 2019.

[23] Z. S. Hakura and E. M. Kilgariff, "Calculation of plane equations after determination of Z-buffer visibility," 2014, US Patent 8,692,829.

[24] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, pp. 630–645, Springer, 2016.

[27] D. S. Rice and M. F. Deering, "Displacement mapping by using two passes through the same rasterizer," 2006, US Patent 7,148,890.

[28] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: a database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[29] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," 2014, https://arxiv.org/abs/1406.2661.

[30] H. Kuang, Y. Ding, X. Ma, and X. Liu, "3D face reconstruction with texture details from a single image based on GAN," in *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 385–388, Qiqihar, China, 2019.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Cham, 2015.

[32] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, https://arxiv.org/abs/1803.07422.

[33] H. Yang, H. Zhu, Y. Wang et al., "FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 601–610, Seattle, WA, USA, 2020.

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