IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 January 2024, accepted 22 March 2024, date of publication 27 March 2024, date of current version 2 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381975

TOPICAL REVIEW

3D Face Reconstruction Based on a Single Image: A Review

HAOJIE DIAO[1], XINGGUO JIANG[1,2], YANG FAN[1], MING LI[1], AND HONGCHENG WU[1]
[1]Sichuan University of Science and Engineering, Zigong 643000, China
[2]Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644005, China

Corresponding author: Xingguo Jiang (tonny_jiang@suse.edu.cn)

This work was supported in part by the Scientific Research Foundation of Sichuan University of Science and Engineering under Grant 2019RC12, and in part by the Open Foundation of Artificial Intelligence Key Laboratory of Sichuan Province under Grant 2020RZJ03.

ABSTRACT Nowadays, along with the rise of digital human system, 3D animation, intelligent medical and other industries, 3D face reconstruction technology has become a popular research direction in computer vision and computer graphics. Traditional 3D face reconstruction techniques are affected by face expression, occlusion, and ambient light, resulting in poor accuracy and robustness of the reconstructed model, etc. With the rise of deep learning, all of the above problems have been greatly improved. Focusing on 3D face reconstruction techniques based on deep learning, this paper categorizes the existing research works into 3D face reconstruction based on hybrid learning and explicit regression. The first category of research work fits 2D faces to 3D models, which is a pathological process that requires solving the basis vector coefficients of the 3D face statistical model. The second type of research work, instead of Model Fitting, represents 3D faces with multiple data types in the display space and directly regresses 2D faces through deep networks. This review provides the latest advances in single-image-based 3D face reconstruction techniques in recent years, summarizing some commonly used face datasets, evaluation metrics, and applications. Finally, we discuss the main challenges and future trends of the single-image 3D face reconstruction task.

INDEX TERMS 3D face reconstruction, deep learning, 3DMM, model fitting, Nerf.

I. INTRODUCTION

In recent years, with the development of face-related technologies, 2D face-related technologies such as face expression classification, face detection, face recognition, face attribute editing, etc. have become more and more mature, however, due to the difficulty of 2D face images to support the application of 3D faces and the increased requirements of accuracy and precision in acquiring face-related information, compared to 2D face images, 3D faces are able to present more abundant information such as the shape, gesture, texture, and so on, and problems such as perspective conversion and angular occlusion will not affect their characterization in 3D space. The information such as the shape, posture and texture of the face can be more richly displayed in space, and the problems such as perspective conversion and angular occlusion will not affect its characterization in 3D space [1], [2]. Therefore, the technique of reconstructing high-fidelity 3D face models from 2D images has received a lot of attention from researchers. Compared with reconstructing a 3D face from multiple 2D images acquired from different viewpoints, it is more challenging to recover a 3D face using only a single unrestricted 2D image, so this paper focuses on the problem of reconstructing a 3D face based on a single 2D image.

The essence of 3D face reconstruction based on a single image is the process of recovering the 3D coordinates and RGB values of each pixel point on a 2D face image at its corresponding position in a 3D face under the condition that it describes some a priori information [3]. We retrieve the methods commonly used in 3D face reconstruction tasks in recent years, divide them into traditional methods and deep learning-based methods in time scale, and mainly elaborate some typical algorithms for 3D face reconstruction based on deep learning. The framework of the divided methods is shown in Fig. 1. From the perspective of regression process,

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

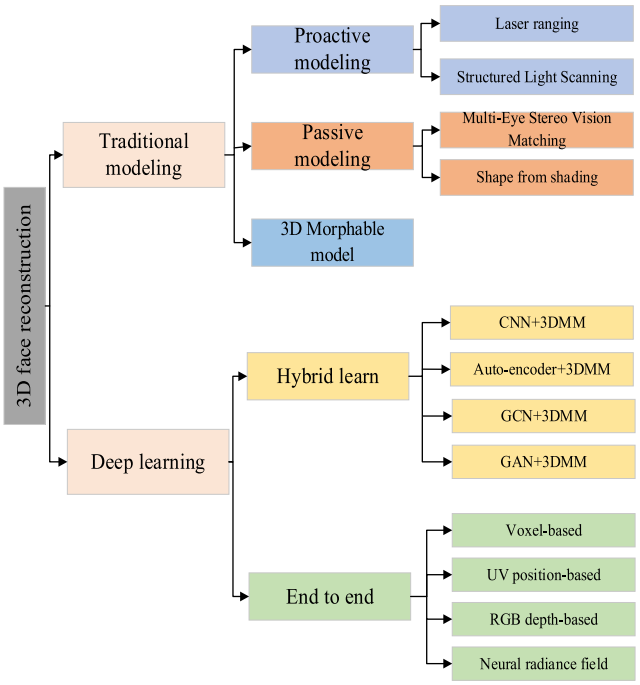

FIGURE 1. Classification of 3D face reconstruction methods.

3D face reconstruction based on deep learning can be subdivided into hybrid learning-based and end-to-end regression based on display space, in which hybrid learning-based methods first encode the 2D image into a series of vectors mapped into the hidden space through feature extraction and other operations, and decode and reconstruct the 3D face with the help of 3D deformable a priori information, while end-to-end regression based on display space can directly regress the 3D representation corresponding to each pixel position from a single 2D image.

This paper is organized as follows: section II gives a brief introduction to the relevant physical models involved in the 3D face reconstruction process. Section III introduces some 3D face reconstruction based on traditional methods and discusses the advantages and disadvantages between each method. In Section IV, the typical algorithms of 3D face reconstruction based on hybrid learning in recent years are sorted out and divided into four categories according to the different neural network architectures: based on Convolutional Neural Networks (CNN), based on Auto-Encoder, based on Graph Convolutional Networks (GCN), and based on Generative Adversarial Networks (GAN), and the ideas, steps, and features of each algorithm are elaborated in detail. Section V combs through the typical algorithms for 3D face reconstruction based on end-to-end regression in recent years and divides them into four categories including: voxel, UV position map, RGB-depth, neural radiation field, and elaborates the ideas and features of each algorithm. Section VI introduces some popular datasets and evaluation metrics used to evaluate the quality of 3D face reconstruction. Section VII shows the application scenarios of 3D face reconstruction techniques. Section VIII summarizes some current challenges in the field of 3D face reconstruction and gives some suggestions.

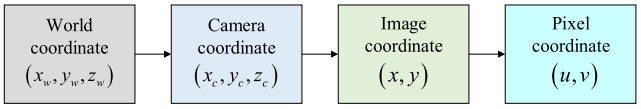

FIGURE 2. Process of imaging an object.

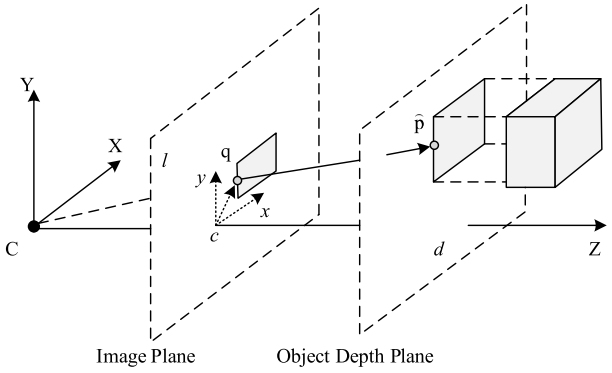

FIGURE 3. Weak perspective projection process: 3D objects are flattened to the same depth and their projections on the 2D image are scaled according to their proximity to the camera.

II. BACKGROUND KNOWLEDGE

A. CAMERA MODEL

The camera model describes the projection process of face vertices from a 3D world coordinate system to a 2D picture [4], which is shown in Fig. 2.

Given a point p in a world coordinate system, its corresponding point q on the 2D picture is calculated by the following equation:

$$q = \Pi(Rp + t) \tag{1}$$

where R and t are camera external references, which represent the rotation matrix and translation vector, respectively, and which transform the point p to $\hat{p}$ under the camera coordinate system by rigid-body transformations such as scaling, rotation translation, etc. Π denotes the projection process associated with the camera internal reference, which transforms $\hat{p}$ to the image coordinate system to obtain the point q. In 3D face reconstruction algorithms, the most commonly used projection model is the weak perspective projection. Its imaging process is shown in Fig. 3.

The weak perspective projection is widely used in 2D key tracking algorithms and dense tracking algorithms due to its computational efficiency [5]. Its projection matrix is as follows:

$$\Pi = s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{2}$$

where $s = 1/d$ denotes the similarity factor of the transformation, d is the depth of the object, and the size of the object in its plane varies with the distance from the camera. In 3D face reconstruction, the camera model can be used to project the 3D face model onto a 2D image and compare it with the

ground-truth of the input 2D image, and establish a loss to constrain the model reconstruction.

B. ILLUMINATION MODEL

In computer graphics, illumination model is to set a constant direction and intensity light source in a simulated 3D environment to interact with the surface of an object so as to generate images with real emotions. In the 3D face reconstruction task, the illumination information plays an indispensable role in rendering the face picture, but it is usually difficult to recover the illumination in the real scene, so researchers generally use a simplified illumination model such as: spherical harmonic function illumination model [6]. The internal calculation is as follows:

$$Y_k^n(\theta,\phi) = \begin{cases} \sqrt{2}H_k^n \cdot \cos(n\cdot\phi)\cdot P_k^n(\cos(\theta)) & \text{if n} < 0 \\ H_k^n \cdot P_K^n(\cos(\theta)) & \text{if n} = 0 \\ \sqrt{2}H_k^n \cdot \sin(-n\cdot\phi)\cdot P_k^{-n}(\cos(\theta)) & \text{if n} > 0 \end{cases} \tag{3}$$

where k is the index order, H_k^n denotes the normalization factor, and P_k^n denotes the accompanying Legendre polynomial of order k. Based on the spherical harmonic function, a function $T(\omega)$ defined on a sphere can be approximated as:

$$T(\omega) \approx \sum_{k=0}^{B-1}\sum_{n=-k}^{k} g_k^n Y_k^n(\omega) \tag{4}$$

where B is the order of the basis function and g_k^n is the coefficient of the basis function. For ease of understanding the basis function is expressed using a one-dimensional subscript: $Y_1 = Y_0^0, Y_2 = Y_1^{-1}, Y_3 = Y_1^0, \cdots, Y_{B^2} = Y_{B-1}^{B-1}$. Based on the above expression of the basis function, the light can be approximated by a ball-harmonic basis function:

$$C(n_i, t_i \mid \gamma) = t_i \cdot \sum_{b=1}^{B^2} \gamma_b \Phi_b(n_i) \tag{5}$$

where $\Phi_b(n_i)$ dote the ball-harmonic basis function computed from the normal vectors of the vertices of the 3D face model, while $\gamma = [\gamma_1, \ldots, \gamma_{B^2}]$ denotes the control of the illumination transformation with ball-harmonic coefficients.

C. RENDER

Rendering is the process of computationally sampling a 3D model to obtain a 2D image through different viewing conditions (e.g., orientation, distance), which is determined by the representation and storage form of the 3D model (e.g., point cloud, mesh, voxel). Among the rendering methods often used in 3D face reconstruction tasks can be categorized into traditional graphics rendering and neural network rendering. Among the rendering methods often used in 3D face reconstruction tasks can be categorized into traditional graphics rendering, neural network rendering.

In traditional graphics rendering, the surface of the face is usually regarded as a Lambertian reflection model [7], which assumes that the surface of the face is an ideal diffuse reflector and only produces diffuse reflection phenomena. As shown in Fig. 4, the luminance received by the observer does not change with the viewing angle, and the diffuse reflection follows the cosine law, i.e., the intensity of diffuse reflection depends on the angle of the normal vector of the intersection point of the incident light line and its surface, which is calculated as follows:

$$l = l_0 \cos\theta \tag{6}$$

where l is the incident light intensity, l_0 is the reflected light intensity, θ is the angle between the light vector and the normal vector. In the rendering process, the pixel values corresponding to each point of the imaging plane are calculated according to the lighting conditions of the environment where the face model is located:

$$P(x) = A(x)l^T S(x) \tag{7}$$

where x denotes the pixel coordinate position, $P(x)$ denotes the pixel color, $A(x)$ denotes the original color of the image, and $S(x)$ denotes the normal vector of the image pixels. In the actual rendering, in order to simulate the reflection effect of the smooth skin surface on the light source, an additional highlight component will be introduced as a supplement [7].

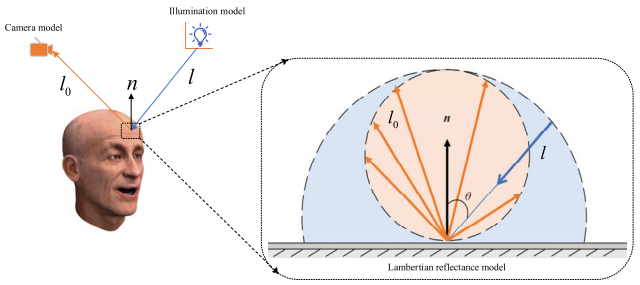

FIGURE 4. Rendering based on Lambertian reflection, the left side shows the imaging process of a given light source reflecting on a face model, and the right side is a zoomed-in illustration of the process, with light diffusely reflecting on the surface.

Neural network rendering is an emerging rendering method combined with deep neural networks in recent years, in which given parameters are input to the deep network for inference, implicit 3D information stored in the neural network is obtained and synthesized in combination with the rendering principles of graphics, and the rendered image is finally obtained. Neural Radiation Fields as the most representative neural network rendering, is widely used in the field of 3D reconstruction and high-fidelity perspective synthesis [8], its rendering flow is shown in Fig. 5.

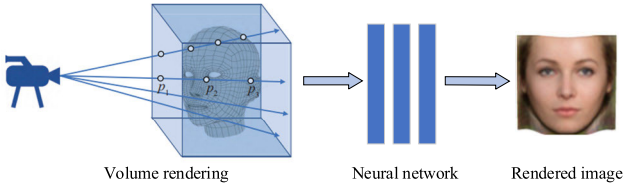

FIGURE 5. Neural radiation fields based on volume render.

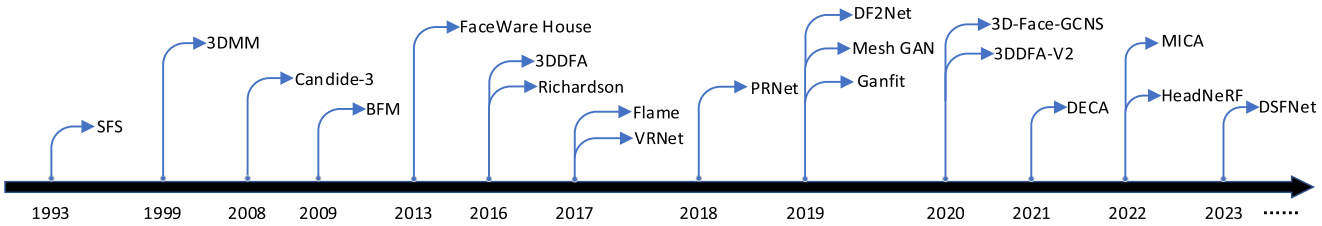

FIGURE 6. The development lineage of 3D reconstruction of faces from a single image.

Specifically, the deep network first learns the five-dimensional continuous field used to represent the 3D model through the multi-view image with its corresponding camera parameters:

$$F_\theta : (p, f) \rightarrow (\sigma, c) \tag{8}$$

where F_θ is a learnable deep neural network, $p = (x, y, z)$ is a point in 3D coordinates, $f = (\theta, \ell)$ is the direction of the light path emitted from the observation viewpoint, c is the pixel information learned by the network, and σ is the attenuation effect produced by the light during propagation, and the attenuation process is satisfied:

$$dl = \sigma(t)dt \tag{9}$$

where dt is the integral differential of the propagation path, and $\sigma(t)$ is the attenuation coefficient at time. Centered at the camera coordinates o, for a point p in the pixel coordinate system of the target image, its corresponding imaged ray is:

$$r(t) = o + f \cdot t \tag{10}$$

In summary, $w(t)$ is the transmission ratio of light propagating in the target volume, and the imaging color $C(p)$ of pixel p point is satisfied:

$$\begin{cases} C(p) = \int_0^\infty w(t) \cdot c(f, r(t))dt \\ w(t) = \exp\left\{ -\int_0^t \sigma((r(s))ds \right\} \cdot \sigma(r(t)) \end{cases} \tag{11}$$

Traditional graphics rendering is mainly applied to display represented 3D face models, and the models and techniques involved in its rendering process have a long history of development, while neural network rendering is applied to implicit 3D face models, where integral computation is carried out in the target volume through the light path and inference is performed in the neural network, and the rendering accuracy and realism are greater than the former, due to the fact that the implicit information stored inside is obtained from the images within the continuous scene and is the real result of the action of objects and light, its rendering accuracy and realism are greater than the former. Table 1 shows the advantages and disadvantages of the two rendering methods.

TABLE 1. Comparison of advantages and disadvantages.

Methods	Traditional graphics rendering	Neural network rendering
Advantage	(1) Faster rendering (2) Support for dynamic scene rendering (3) There is more mature research on improving rendering speed and optimizing rendering quality.	(1) High generalizability to complex scenarios (2) High rendering accuracy and realism (3) Deep learning supporting technologies are fast iterative and promising
Disadvantage	(1) Rendering requires additional materials and textures to display the 3D model. (2) High demands on computer storage space	(1) Inability to render face models in real time (2) Rendering is time-consuming and has a high demand on GPU performance

D. DEVELOPMENT OF 3D FACE RECONSTRUCTION BASED ON A SINGLE IMAGE

With Parke et al [9]. in 1974 using a computer to generate the first three-dimensional model of the face to this point, based on a single image of the 3D face reconstruction has been 50 years of development, the rise of deep learning in the field of image reconstruction for the 3D face reconstruction provides a new way of thinking, Fig. 6 shows the emergence of representative 3D face reconstruction algorithms at various points in time.

III. TRADITIONAL METHODS

A. PROACTIVE MODELING

Proactive modeling, also known as active ranging 3D face modeling, generally immobilizes a person in a specific scene and uses 3D data acquisition equipment to capture geometric information about the face. Common proactive modeling methods are mainly laser ranging technology [10] and structured light technology [11]. The working principle of laser-based scanning ranging is to use a scanning system to emit laser light into the target face, and its light wave will be reflected back to the sensor by the surface of the face, the longer the time required for this process represents the deeper the depth of the surface of the face, and this process is known as time-of-flight, Another method is a structured light-based modeling system, whose system is composed of one or more cameras and projectors. The working principle

of structured light scanning is to use light waves in a finely calibrated pattern to be projected onto the scanned object, and the light rays generally fall on the surface of the face in the form of regular and isometric stripes, and the depth value and texture information of the face is calculated from the light information of the surface of the face, so as to obtain a high-precision 3D face Point cloud data. The form factor of the 3D laser scanner and 3D structured light scanner is shown in Fig. 7.

FIGURE 7. Laser scanner (left) and structured light scanner (right) used to capture 3D face data.

B. PASSIVE MODELING

Passive modeling does not require specialized physical scanning equipment, but only multiple 2D face images or a series of 2D image sequences containing visual information for modeling. The passive modeling approaches include: multimetric stereo vision matching, photometric based method. Among them, the method based on multimetric stereo visual matching is to recover the shape from multiple face images taken from different viewpoints, taking binocular visual matching as an example to illustrate the principle: it utilizes the polar geometry to transform the problem into Euclidean geometric conditions, and then utilizes the triangulation to compute the depth information of stereo-matched face images, so as to obtain the dense three-dimensional spatial point cloud. The method can be divided into five steps, which are image acquisition, camera calibration [12], image correction [13], stereo matching [14], and 3D reconstruction. The binocular vision 3D face reconstruction system is shown in Fig. 8.

FIGURE 8. Binocular vision 3D face reconstruction system.

The most typical photometric-based face reconstruction is shape from shading [15], which uses the light and dark information in the two-dimensional grayscale image and the luminance generation principle to simulate the light changes of the environment, so as to obtain the normal vector of each pixel point in the three-dimensional space, and then the depth value is obtained by calculating according to the normal vector to obtain the reconstruction model. Smith and Hancock [16] proposed a statistical method to recover the surface normal field from a single-intensity image by embedding surface normals into the SFS framework by embedding surface normals in place of surface depth, the correlation between surface orientation and image intensity is utilized to recover the surface normal field from single image. This method assumes a Lambertian model and a single point light source with known orientation, and therefore cannot handle the lighting conditions in complex cases. Hu et al. [17] used a sufficiently large number of images to generate a reliable source template face, and input a single face image to estimate the target face's light by a sparse transmission model to estimate information such as illumination and albedo of the target face to generate a 3D face model. SFS is suitable for objects with uniformly concave and convex surfaces, however, the reconstructed model tends to be less robust for more complex face regions due to the fact that small changes in the surface normals lead to significant changes in the corresponding composite surfaces. Castelán and Hancock [18] integrated a localized shape-based method integrated into the SFS framework to enhance the concavity of the integrated surfaces, where they utilize local descriptors of shape indices and curvature to constrain regions and make the necessary corrections to the surface normal vectors to enhance the convexity of the surfaces and ensure that the integrated surfaces have a global height maximum. However, SFS-based 3D face reconstruction has three main limitations: (1) it cannot handle 3D reconstruction under complex illumination, and most of the methods make simple assumptions about the illumination conditions; (2) it is only applicable to a single frontal photo of the face, and is unable to recover large-pose facial images. (3) Cannot reconstruct a realistic feeling 3D face model.

C. 3D MORPHABLE MODEL

The 3D Morphable Model (3DMM) was proposed by Blanz et al. [19] in 1999. The principle is that 3DMM treats faces as distributions in a linear subspace of a high-dimensional space, and any face can be represented by a linear combination of the average face and other faces constructed from a database. Specifically: they constructed a 3D face database using laser scanning, reduced the collected face data to a linear basis using Principal Component Analysis (PCA), computed the face geometry and texture parameters by matching with the target image, and combined the linear basis of the principal components and the average face parameters to obtain the target 3D face model, the formula for its calculation is as follows:

$$\begin{cases} S_{\text{mod}} = \sum_{i=1}^{N} a_i S_i \\ T_{\text{mod}} = \sum_{i=1}^{N} b_i T_i \\ \sum_{i=1}^{N} a_i = \sum_{i=1}^{N} b_i = 1 \end{cases} \tag{12}$$

where S_i and T_i are the shape vector and texture vector of the ith face in the database. Since they are not orthogonally correlated, S_i and T_i cannot be used directly as basis vectors in the actual model construction, and we need to use PCA to perform dimensionality reduction decomposition: firstly, we compute the average face shape vector $\overline{S}$ and the average texture vector $\overline{T}$, and secondly, we center the face data to obtain $\Delta S = S_i - S$ and $\Delta T = T_i - T$. Then, we compute their covariance matrices C_s and C_t, respectively, and finally, we obtain the eigenvalues α_i, β_i, and eigenvectors and the eigenvalues and eigenvectors for the shape and texture covariance matrices. and eigenvectors s_i and t_i. In summary, the shape and texture of a 3D face can be represented by equation (13):

$$\begin{cases} S_{\text{model}} = \overline{S} + \sum_{i=1}^{N-1} \alpha_i s_i \\ T_{\text{model}} = \bar{T} + \sum_{i=1}^{N-1} \beta_i t_i \\ \sum_{i=1}^{N-1} \alpha_i = \sum_{i=1}^{N-1} \beta_i = 1 \end{cases} \tag{13}$$

The $N-1$ eigenvectors are chosen based on the descending order of the eigenvalues, and taking the first few components of s_i and t_i at the same time gives a good approximation of the original samples, and therefore better reduces the number of parameters that need to be estimated without loss of accuracy.

The flow of 3D face reconstruction algorithm based on 3DMM is shown in Fig. 9. The essence of how to fit a 2D image to a 3D model is to solve the base vector coefficients of shape and texture. The solution idea is as follows: input the 2D face image, and then combine it with the existing 3D face database to solve the base vector coefficients $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{N-1})$, $\beta = (\beta_1, \beta_2, \cdots, \beta_{N-1})$ of shape and texture, as well as the external rendering parameters including the camera position, the rotation angle of the image plane, the ambient light component, the image contrast and so on, a total of more than 20 dimensional parameters. Under the control of these initial parameters, the reconstructed 3D face model can be rendered onto a 2D image, the error is calculated with the input initial image, and then the relevant parameters and 3D model are adjusted by back propagation of the error, and iterated until reaching the optimization.

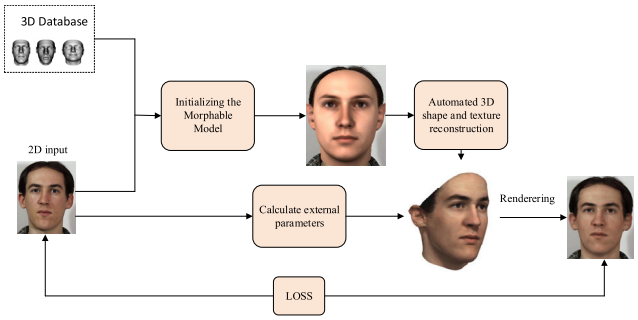

FIGURE 9. 3DMM-based 3D face reconstruction process.

TABLE 2. Comparison of advantages and disadvantages of traditional based 3D face reconstruction methods.

Methods	Advantage	Disadvantage
Laser ranging	(1) Capable of obtaining high-precision face data (2) Reconstructed 3D face models are more realistic	(1) The obtained 3D point cloud data is large and needs to be matched (2) Capture equipment is expensive
Structured Light Scanning	(1) Only one image of a face is needed to Obtain 3D shape; (2) Reconstruction structure with high precision and strong anti-interference ability	(1) Slow reconstruction (2) Face capture for indoors only
Multi-Eye Stereo Vision Matching	(1) 3D point cloud capable of providing dense parallax information for recovery (2) Reconstructed 3D face model is more detailed	(1) High cost, high computing power, susceptibility to light conditions (2) Inability to solve the problem of self-obscuring face images
SFS	(1) Reconstruction results are more accurate and have a wide range of applications (2) Reconstruction with just a single image	(1) Susceptible to light sources, Dependent Mathematical operations, Robustness Poor (2) Inability to solve the problem of large-pose face pictures
3DMM	(1) The 3D topology is known and the reconstruction results are complete (2) Its face is composed of linear superposition, and a new face model can be generated by changing any parameter.	(1) The creation of a 3D face database requires high levels of acquisition equipment and processing. (2) Constrained by linear space, the reconstructed face model is more average

3DMM-based face reconstruction algorithms were widely used at that time, but they usually suffer from the following drawbacks: (1) the reconstructed face model is poorly personalized, and the accuracy is highly dependent on the diversity of the data in the 3DMM database; (2) the model fitting is a pathological problem, which does not have a globally optimal solution per se and is prone to fall into a localized solution; and (3) due to discontinuities in the error function itself, the face image background noise and self-occlusion can seriously affect the reconstruction accuracy. (4) The linear space formed by principal component analysis is a low-dimensional space, which contains coarse shapes and textures. Table 2 shows the advantages and disadvantages of 3D face reconstruction based on traditional methods.

IV. HYBRID LEARNING

A. CONVOLUTIONAL NEURAL NETWORK BASED RECONSTRUCTION

While the classical 3D morphable model is able to recover the global geometric information well for 3D face reconstruction, the model is difficult to capture the deep detailed features of the face, resulting in a lack of detailed representation in the reconstructed 3D model. In recent years, due to the excellent performance of deep learning in image feature extraction, researchers have combined 3DMM with convolutional neural networks in 3D face reconstruction tasks and designed a more effective loss function to constrain the learning process, which enables the model to better learn and characterize the 3D details of the face. This hybrid learning-based approach has attracted academic attention and become a research hotspot in the field of 3D face reconstruction. Its algorithm flowchart is shown in Fig. 10.

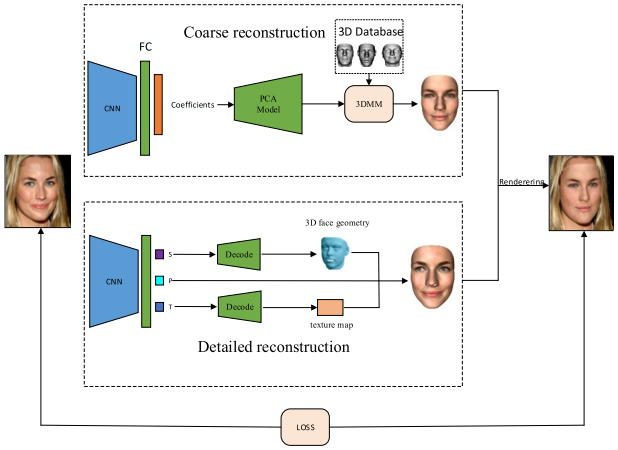

FIGURE 10. CNN-based 3D face reconstruction algorithm process.

In Fig. 10, $\{S, P, T\}$ denote the predicted three-part feature encoding for shape, pose, and texture, respectively. The first convolutional neural network extracts the global features of the face image, and predicts the 3DMM basis vector coefficients through the fully connected layer (FC), after which the basis coefficients are combined with the 3D morphable model in the 3D dataset to obtain a rough 3D face model. The second fully connected layer predicts the shape and texture parameters, and decodes the 3D face geometry and texture mapping through a decoder, and then combines the pose parameters to get the face model with detailed features, and finally calculates the loss function between the rendered 2D image and the input 2D image, and then optimizes the network.

Compared with the traditional 3D morphable model, convolutional neural network has the shortcomings of high algorithm complexity and lack of a priori information. How to combine convolutional neural networks and 3DMM to reconstruct 3D face models with detailed features has become a greater concern for researchers. In 2016, Richardson et al. [20] applied convolutional neural networks to 3D face reconstruction for the first time, and the authors proposed a coarse-to-fine training method. Specifically, the algorithm consists of two networks, respectively CoarseNet and FineNet. Among them, CoarseNet computes the base vector coefficients of 3D face by 3DMM, which serves to initially reconstruct the coarse 3D face geometry, while FineNet further refines the face features by means of SFS, which can reconstruct the 3D face model with details very well, but for the reconstruction of self-obscuring images that is less robust. Similarly, Dou et al. [21] used identity and expression fusion to express face shapes, Luo et al. [22] proposed a SCNN algorithm to compute more robust and personalized 3DMM coefficients, and Fan et al. [23] proposed a dual neural network-based face reconstruction algorithm based on both, which maximally combines the advantages of convolutional neural networks and 3DMM, and further improves the accuracy of face reconstruction. In 2020, Zhu et al. [24] proposed an image fitting algorithm (ReDA), which, with the help of the idea of soft rasterization, splits the mesh into a number of blocks along one direction, each of which is rasterized in a traditional way, using the operations of a number of different convolutional kernels to aggregate along the space and across layers. Experiments show that the ReDA algorithm is able to establish dense correspondences and reconstruct 3D faces with high quality.

In order to get higher accuracy in the reconstructed model, most researchers use face landmarks alignment to calculate the error, and face alignment has become the key in the reconstruction task. In 2016, Zhu et al. [25] proposed 3D Dense Face Alignment algorithm (3DDFA). The algorithm encodes the orientation information between the key points of the image into the pose adaptive convolution (PAF) to facilitate the fusion of the pose information in the training phase and improve the robustness and accuracy of the model to pose changes. The authors also innovatively propose the normalized projected coordinate coding PNCC, which normalizes the position information of each pixel point on the face and encodes it into a vector for representing the features of the face shape. The fitting process can be expressed as follows:

$$P^{k+1} = P^k + \mathrm{Net}^k\left(\mathrm{Fea}\left(I, P^k\right)\right) \tag{14}$$

where Net^k denotes the parameters used for the regression network, and $Fea(I, P^k)$ denotes the coefficients of the 3DMM basis vectors iterated in the 2D image. 3D reconstruction is performed based on $P^k = \left[f^k, R^k, t_{2d}^k, \alpha_{id}^k, \alpha_{exp}^k\right]$ and 3DMM models to obtain the face shape S_{k+1} in the K+1st generation, which is normalized to obtain NCC.

$$\begin{cases} NCC_{k+1}^d = \dfrac{\overline{S_{k+1}} - \min\left(\overline{S_{k+1}}\right)}{\max\left(\overline{S_{k+1}}\right) - \min\left(\overline{S_{k+1}}\right)} \\ d = (x, y, z), NCC \in \mathbb{R}^3 \end{cases} \tag{15}$$

Then a Z-Buffer renderer is invoked to get the 2D image PNCC based on NCC:

$$\begin{cases} PNCC = Z - Buzzfer(V(p), NCC) \\ V(P) = R^* S + [t_{2d}, 0]^T \end{cases} \tag{16}$$

The transformation process of NCC and PNCC of a face is shown in Fig. 11.

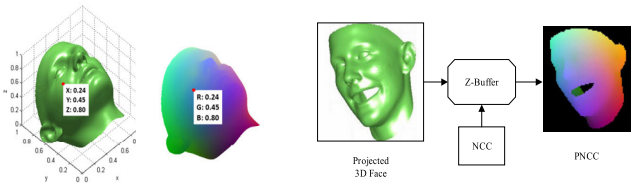

FIGURE 11. Face normalized coordinate code (left) and face normalized projected coordinate code (right) [25].

The advantage of 3DDFA is that it skillfully transforms the problem of annotating the key landmarks into the problem of fitting the image to the model, and is able to accurately capture the shape and structure of the face even in the presence of large variations in facial expressions or head postures. Accurate 3D face reconstruction often requires an iterative algorithm for continuous optimization, which leads to a long model training time and is not conducive to real-time and fast applications. MobileFace is an algorithm that performs 3D face reconstruction in real-time on mobile devices [26], which is based on a lightweight network, such as MobileNet CNN, to regress the shape of the 3D face, and introduces the face texture into the energy function to improve the detailed representation of the model. Similarly, Wang et al. [27] used MobileNet to regress the 3DMM basis vector coefficients.3DDFA_V2 is based on the improvement of the 3DDFA algorithm [28], and its improvement strategy is to replace the traditional dilb with a fast face detector Face-Boxes and introduce a lightweight network MobileNet_V2. Compared to the original version, 3DDFA_V2 can achieve higher detection accuracy and faster operation speed.

In 3D face reconstruction based on hybrid methods, the ideal situation is to have a large number of accurate 3D scanned faces as label data, but in most cases, a large number of 3D scanned faces are not available due to cost, privacy and other reasons, and usually only a small number of 3D scanned faces can be utilized to train the model, and the reconstructed model is often poorly robust in this way. To compensate for the lack of 3D face data, Sela et al. [29] proposed an image transformation network that can be used to synthesize 3D face datasets, Liu et al [30]. designed a method to fit a 3D Morphable model to multiple 2D images and then reconstructed a 3D face, and used this kind of generated data as labeled data to train the model, but this method of training the network by synthesizing 3D data It is usually difficult to reconstruct a face model with high accuracy. Jackson et al. [31] proposed a method to reconstruct a 3D model by direct volume of convolutional neural network. This method uses a multi-feature fitting approach to generate a training model by fitting a 3DMM combining the BFM model [32] and the FaceWarehouse model [33] to a 300W dataset [34]. This method does not need to use the 3DMM directly to generate faces that fit arbitrary poses and expressions. In order not to rely too much on 3D face data, researchers have focused on unsupervised or self-supervised methods. When the dataset is only 2D images, the reconstructed 3D face model is rendered back to the 2D image and loss calculation is established with the input 2D face image to supervise the training, this method is called unsupervised training. Similarly, on the basis of unsupervised, after the reconstructed 3D face model is rendered back to the 2D image, the image is reconstructed to a 3D face model and loss calculation is established with the previous 3D face model to supervise, this method is called self-supervised. Weakly supervised learning method is to add some weakly supervised information such as facial landmark, skin masks, etc. during the training process. The 4 methods of supervision are shown in Fig. 12.

In 2017, Tewari et al. [35] designed a novel deep convolutional self-encoder that achieves face reconstruction in a self-supervised manner without any 3D face data, the encoder first extracts semantic information such as pose, shape, expression, texture, illumination, etc. from the input image and encodes them, followed by decoding the semantic features into 3DMM basis vector coefficients and external parameters for reconstructing the face via a model-based decoder. Finally self-supervised training is achieved by reconstructing the rendered 2D image again. It has been pointed out that using only pixel loss and perceptual loss to constrain the model training is insufficient, and more loss functions need to be combined to achieve better results, so Deng et al. [36] proposed a method that allows unsupervised monocular reconstruction on a single image or dataset, which designs a set of complex hybrid loss functions that can take into account the supervision of both low-level and perceptual levels, and utilizes the complementary information between different images for aggregation, and uses an attention mask for some regions of the face to reduce the effect of occlusions on the reconstruction quality. In 2020, Tu et al. [37] proposed the 2DASL (2D-assisted self-supervised learning) algorithm, they based on the self-supervised training approach by regressing the convolutional neural networks on the base vector coefficients of 3DMM to solve the shortage of 3D face data. The algorithm is trained using two sets of images. The first set of images contains the ground-truth of 3DMM coefficients, which is supervised by minimizing the error of 3DMM coefficients. The second set of images has the ground-truth of facial landmarks, which is supervised by minimizing the facial landmarks errors to supervise, the method allows reconstruction in ground-truth conditions without 3D faces, but the model is poorer in terms of detail representation. Wu et al. [38] proposed an unsupervised training approach based on multilevel loss function. Combining the traditional 3D Morphable model, a convolutional neural network is utilized to learn 3D face features directly from a large number of 2D face images. In addition, in order to solve the self-obscuration problem, they use a facial parsing segmentation algorithm based on the CelebAMask-HQ [39] dataset to preprocess the image to remove the occluded region. Experiments show that the reconstructed face model has a large

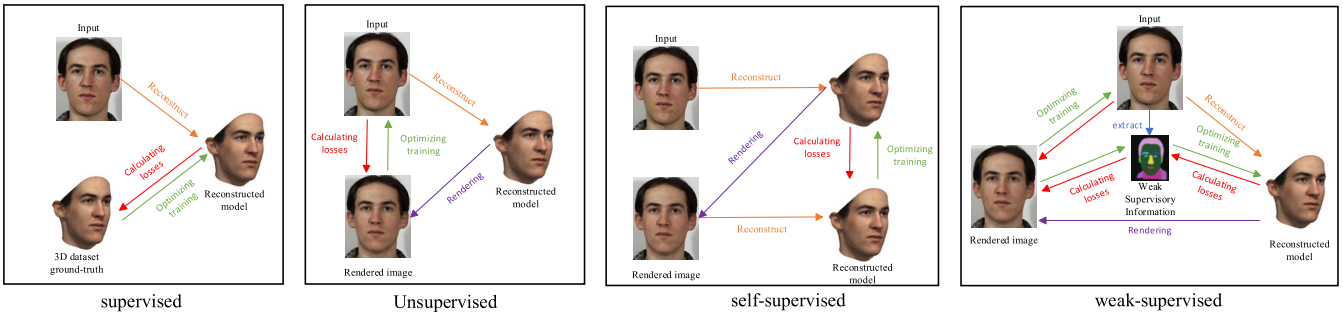

FIGURE 12. Schematic diagram of supervision methods.

quality and accuracy improvement. In 2021, Zhang et al. [40] designed a self-supervised framework (WM3DR), that uses a single convolutional neural network to predict multiple face parameters simultaneously, the framework not only greatly reduces redundant computations in feature extraction, but the architecture of a single network makes the model easier to deploy, and more importantly uses a global camera model for the reconstructed face in each image, allowing the method to be applied without detection or pre-processing such as cropping. Zielonka et al. [41] proposed MICA algorithm, in order to address the effect of perspective projection on the reconstructed face dimensions in self-supervised training, the authors proposed a training scheme based on hybrid supervision. Due to the lack of large-scale 3D databases, the authors synthetically labeled small and medium-sized databases to generate a model that contains 2300 identities and corresponding image labels of the 3D database, a pre-trained face recognition network arcface is introduced, which provides different features for different faces and is robust to scenes with different expressions and lighting variations, and these features are utilized to train a face shape estimator, which inherits the robustness and generalization of the face recognition network. Most existing 3D face reconstruction algorithms are sensitive to severe facial occlusion and large perspective transformations. For this reason, Li et al. [42] proposed a dual-space fusion network (DSFNet) in 2023 to solve the facial occlusion problem, which firstly performs image feature regression in the visible region of the face, predicts the model coefficients based on the regression features of the visible region, and utilizes the 3DMM's a priori information to reconstruct the occluded region, the network combines the advantages of image space and model space prediction, which is robust and accurate for reconstruction of input images in the presence of facial occlusion. Lei et al. [43] proposed a novel hierarchical representation network (HRN), specifically, the method decouples facial geometry into low-frequency geometric structures, mid-frequency details, and high-frequency details. In the low-frequency part, a rough geometric appearance reconstruction is performed, while in the mid-frequency part, a deformation map in UV space is used to characterize the facial details, and finally a displacement map is used in the high-frequency part for texture modeling on a pixel scale. The algorithm can reconstruct high frequency details such as wrinkles and dimples very well.

B. AUTOENCODER BASED RECONSTRUCTION

The traditional 3D Morphable model constitutes a linear space by means of principal constituents, which reconstructs a face model with poor diversity and fails to portray individualized details, in order to enhance the expressive ability of 3DMM, Tran et al. [44] proposed a nonlinear 3DMM model, whose algorithmic flow is shown in Fig. 13. The principle is to use a mesh encoder to estimate the projection, shape and texture parameters, and then use two decoders to decode the generated 3DMM parameters into shape and texture maps respectively. In addition, the authors fit the 3DMM through a 300W dataset to obtain the new projection parameter m_1, face shape S_1, and face texture T_1, and compute their losses L_m with respect to the projection parameters estimated by the encoder. L_l, L_T, L_S, and L_{adv} are the losses between the generated model and the real 3D facial landmarks, shapes, and textures, as well as the loss between rendered image and the real image, respectively.

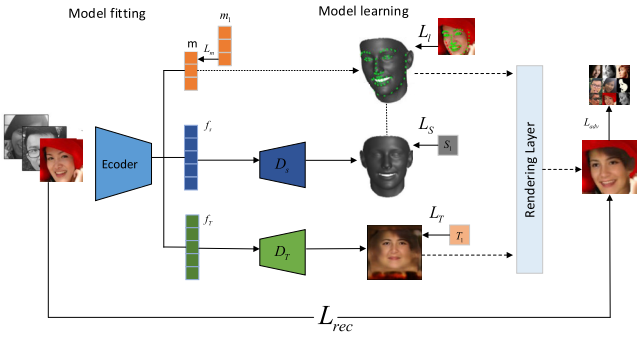

FIGURE 13. Nonlinear 3DMM algorithm flow [44].

In 3D face reconstruction based on autoencoder, most of the algorithms use a single encoder to estimate the face attributes, but the feature extraction processes of different attributes of the face are independent, and the use of a single encoder affects the discriminative power of each attribute. To solve this problem, Li et al. [45] proposed a dual-pathway encoder-decoder network structure, which encodes the features by two different encoders to obtain the identity and expression parameters respectively, and then applies the local

decoder and global decoder to regress the expression and identity of the 3D face model respectively. The algorithm utilizes two encoders to distinguish the expression and identity features of the face, and the reconstructed face model has high accuracy. In order to be able to learn the parameter information of 3DMM from the image, Tran et al. [46] used encoder network to regress the 3DMM parameters on the basis of nonlinear 3DMM, and then reconstructed the 3D model by shape and albedo decoder, the difference is that the output of this algorithm consists of two branches, in addition to the shape and albedo parameters, there are also the results of the fitted shape and albedo, and the four sets of parameters are two by two The combination of the four sets of parameters is input to the rendering layer to reconstruct a face model with high fidelity. Feng et al. [47] designed a novel face reconstruction model (DECA), which takes advantage of the consistency of some of the details of the same face under different environments and expressions, and designed a detail consistent loss to separate the face-specific details from the expression-related details, and realized the face with personalized wrinkles by controlling the expression parameters. animation, specifically the authors also designed two autoencoders to extract the low-dimensional face parameters and detail parameters respectively, after two decoders to get the albedo mapping, 3D face geometry, expression displacement maps, and finally fitting to get the 3D face model. Experiments show that the reconstructed 3D face can be fitted to other human face expressions to generate face animation.

In addition, when the convolutional neural network is regressing the 3DMM parameters, most cases use fully connected layers or 2D convolutional layers to model the parameterized space, which can lead to the generation of large networks with many parameters, which take up a large amount of computer storage space and reduce the efficiency of the algorithm. For this reason, Zhou et al. [48] proposed a method that combines DCNNS and Auto-Encode, which can learn the texture and shape of the face directly through the mesh convolution, which can be more intuitive to the definition of the nonlinear 3DMM, and also has a high computational efficiency because the mesh convolution is defined by a network with fewer parameters, which makes the algorithm have a faster computational speed.

C. GRAPH CONVOLUTION BASED RECONSTRUCTION

Convolutional neural network shows its powerful learning ability in face 3D reconstruction task, which can effectively extract the features in the image, but in convolutional neural network, a picture can be regarded as a Euclidean structure that consists of every pixel neatly arranged, and the number of neighboring nodes around each pixel point is equal and has a certain degree of correlation, so the use of convolutional computation on a picture is reasonable, and the face feature point Sorting is presenting a non-regular network structure, for this reason some researchers have proposed to introduce graph convolution networks that deal with the topology of the network to reconstruct the 3D face. In graph convolution, the number of neighboring nodes of each node is likely to be different, as shown in Fig. 14.

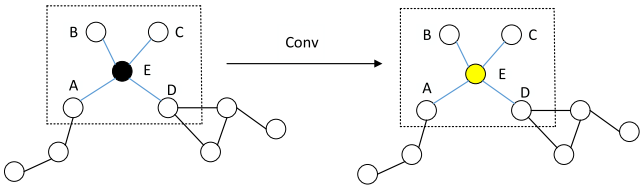

FIGURE 14. Graph convolution process.

In Fig.14, the black E node connected to ABCD has undergone convolutional computation to extract the features of the surrounding nodes, and the features of the new yellow node E have been generated. The formula for its convolution operator is as follows:

$$H_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \frac{1}{c_{ij}} H_j^l W_j^l\right) \tag{17}$$

where H_i^{l+1} is the feature of node i in layer $l+1$, H_j^l is the feature of node j in layer l, $\frac{1}{c_{ij}}$ is the normalization factor, W is the weight at node j in layer l, and N is the number of all the surrounding neighboring nodes including its own node. $\sigma(\cdot)$ is the activation function. The Chebyshev interpolation of order K is usually chosen as the graph convolution kernel g_θ. The equation is as follows:

$$\begin{cases} g_\theta = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) \\ \tilde{L} = 2L/\lambda_{\max} - I \end{cases} \tag{18}$$

where L is the Laplace matrix, $\lambda_{\max}$ is the maximum eigenvalue of the matrix, $T_k \in R^{n \times n}$ is the Chebyshev polynomial, and θ_K is the Chebyshev polynomial factor of order K. In summary, each graph convolution layer after extracting features can be represented as $y = \sigma(g_\theta H)$, H are the input features.

Due to the fact that existing linear 3DMM methods are unable to reconstruct high-quality 3D faces with details, while the spectral decomposition of nonlinear 3DMM representations is unstable in different graphs, Yuan et al. [49] designed a multi-scale graph convolutional self-encoder network in 2019, where the authors processed the face as a graph structure, and utilized graph convolution for feature extraction on the face mesh, and in face reconstruction accuracy is improved. Based on the former, Lin et al. [50] proposed an algorithm for graph convolutional networks in 2020. The framework of the algorithm is shown in Fig.15. which consists of three modules: a pre-trained face recognition module and a feature extraction module, as well as a texture-finishing graph convolution module. First of all, a ''wild'' face image is input, and the feature encoding is generated by the face recognition network, and the regressor regresses the base vector coefficients of 3dmm, including a total of 257 dimensional vectors of pose, illumination, etc.

The graph convolution module consists of three parts, the GCN decoder is responsible for decoding the features of face recognition and generating the detailed mesh vertex colors, and the graph convolution module is composed of three parts. GCN decoder is responsible for decoding the features of face recognition and generating detailed mesh vertex colors, GCN Refiner refines the mesh vertex colors generated by Regressor module, Combine Net is responsible for integrating the mesh vertex colors of GCN Decoder and GCN Refiner to output the final mesh vertex colors, and the discriminator combines the reconstructed face model in 3dmm and the reconstructed face model after graph convolution module refining. The discriminator compares the reconstructed face model of 3dmm with the face model refined by the convolution module to realize supervised training.

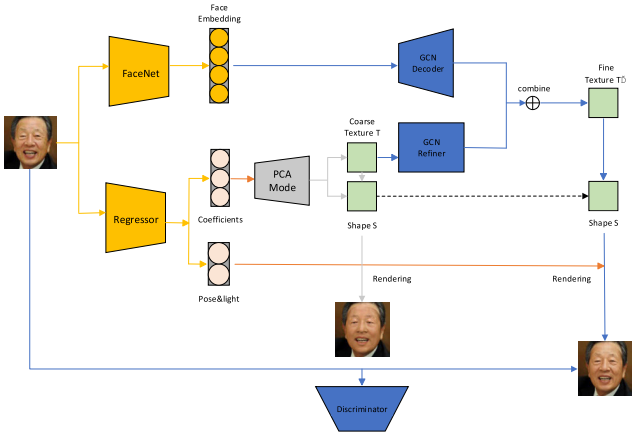

FIGURE 15. GCN based 3D face reconstruction algorithm [50].

In 2021, Qiu et al [51] reconstructed 3D cartoon faces with exaggerated expressions using graph convolutional networks for the first time in the field of cartooning. In order to solve the problem of insufficient datasets, the authors constructed a 3D cartoon dataset containing 2000 high-quality 3D cartoons handmade by professional artists, 3DCaricShop, which also provides rich annotations including 2D cartoon images, camera parameters, and 3D facial features. For the accuracy of the facial shape in the reconstruction task, the authors propose a method to transfer the high-fidelity geometric face reconstructed by the implicit function generator into a mesh with topological structure. To achieve mesh alignment, a new view cooperative graph convolutional network VCGCN is designed for extracting keypoints from the implicit mesh for accurate alignment. The method generates high-fidelity 3D exaggerated facial models in a predefined mesh topology, which can be directly used in animation creation.

D. GENERATIVE ADVERSARIAL NETWORK BASED RECONSTRUCTION

Generative Adversarial Networks [52] is a generative structural model based on deep learning, the idea of which is derived from the two-player zero-sum game in game theory. It is generally composed of a generator and a discriminator, which generates similar data with the characteristics of the training set by learning the data characteristics of the training set, and fitting the distribution of random noise to the real distribution of the training data as much as possible under the guidance of the discriminator. The discriminator, on the other hand, is responsible for distinguishing whether the input data is real or not and feeding back to the generator. The two networks are trained alternately until the generated data of the generative network can be faked as real. The 3D face reconstruction based on generative adversarial network aims to generate high-precision texture maps and then fit them into the 3D face geometry, and to achieve the purpose of reconstructing a high-quality texture by rendering the newly generated 3D face back to the 2D image through the discriminator to distinguish the authenticity. The algorithm flow is shown in Fig. 16.

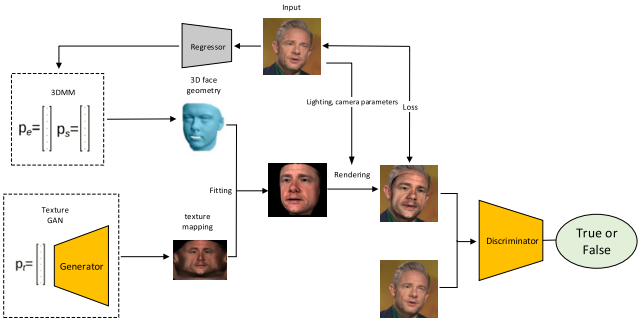

FIGURE 16. GAN-based 3D face reconstruction algorithm.

In hybrid learn based face reconstruction work, the texture features are composed of linear basis vectors of 3DMM, which can lead to low accuracy of the reconstructed face model. Whereas generative adversarial networks have shown excellent performance in generating image data, So Gecer et al. [53] proposed the GANFIT algorithm, in which the authors used GAN to train a face texture generator in UV space, which generates different and fine texture maps by inputting arbitrary noises, which are used as a priori information to be fitted to the face geometry generated by the 3DMM to reconstruct a 3D face model with texture details. Moreover, the authors design four loss functions to constrain the model: (1) Pixel loss of the reconstructed 3D model after rendering it back to a 2D face with an input 2D image through a differential renderer (2) Comparison of the 68 landmarks loss of the rendered image with the input image through a Landmark Detector (3) Inputting the reconstructed 3D face model together with the rendered 2D image as well as the input 2D image into a face recognition network, the identity loss and content loss are established separately. respectively. The advantage of this algorithm is that it can improve the realism of the texture, but the disadvantage is that it requires a large amount of face UV data for training and the texture reconstructed from the ''wild'' image lacks realism. Lattas et al. [54] proposed a method to reconstruct a 3D face model with high fidelity texture from a single ''in the wild'' image using a generative adversarial network (AvatarMe). Firstly, the UV texture of the input image is up-sampled to

obtain plausible face details, and the diffuse reflections of the high-frequency details are obtained by removing the texture information on the UV map, which is combined with the normals of the 3DMM to obtain pixel diffuse reflections and specular diffuse reflections required for the mapping. The two reflections are mapped back onto the UV image, and then the UV map is fitted to the 3D face geometry in order to reconstruct a 3D face model with high-frequency texture details.

In generative adversarial networks, 3D face shapes are often composed of discrete voxels, which can lead to the generation of rough, low-quality face shapes. CoMA [55] is a method for optimizing and improving the 3DMM coefficients on a non-Euclidean space, and the authors innovate the use of meshes as a nonlinear representation of 3D faces. And the proposed MeshGAN [56] algorithm effectively improves the above situation, the algorithm can directly perform the convolution operation on the 3D mesh to generate 3D face meshes with different expressions and different identities, and this method of using mesh representations can simulate the distribution of faces well. However, the algorithm is more complicated in network training. Lee et al. [57] proposed a method that combines graph convolution with generative adversarial networks. Specifically, firstly, the face distribution is encoded using an unqualified perceptual encoder, and the average face mesh is used as the graph structure in the decoder part, the mesh vertices are convolved with graph convolution, and then the high resolution of the generative adversarial network is utilized to generate a high-quality texture mapping, which not only guarantees the authenticity of the texture of the reconstructed model but also guarantees the accuracy of the face shape. In 2023, Sun et al. [58] devised a new 3D GAN framework (Next3D) for unsupervised learning to generate high-quality and 3D-consistent facial images from unstructured 2D images. In order to improve the accuracy and topological flexibility of the 3DMM model, the authors proposed a 3D representation called generating texture rasterized triangular faces. The proposed representation learns to generate neural textures on top of parameterized network templates, and then by rasterization to project them onto three orthogonal viewpoint feature planes to form a triangulated volume rendering. This approach combines the fine-grained representation control of mesh-guided explicit deformation with the flexibility of implicit volume representation. Experiments show that the reconstructed 3D face model not only has high accuracy but also can be directly used for animation.

Nowadays, reconstruction algorithms that introduce 3DMM prior information are based on rendering techniques in computer graphics to render the reconstructed 3D model back to a 2D image, but most of these algorithms use older differentiable renderers, and using newer renderers complicates the model solution. Piao et al. [59] proposed a generative adversarial network-based neural renderer(GAR), specifically, this renderer is no longer limited to the rules of graphics, and ensures the accuracy of the reconstructed model by taking the face normal vector map and other feature encodings as inputs, and by performing a minimization operation between the real rendered face image and the input image.

V. RECONSTRUCTION BASED ON EXPLICIT REGRESSION

In the field of computer vision, 3D faces are mainly characterized by four types of explicit data: point cloud, mesh, voxel and depth, and the various types of data can be transformed into each other as shown in Fig. 17.

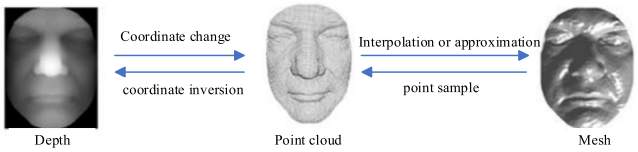

FIGURE 17. Different types of 3d face data.

In terms of detail performance, 3D face reconstruction based on 3DMM a priori information leads to a reconstructed face that is too average and lacks personality details such as wrinkles due to its linear space limitation. Some researchers have proposed to bypass the 3DMM using an end-to-end approach to reconstruct 3D face models from 2D face images. In this chapter, explicit regression-based reconstruction is categorized into four types: voxel, UV position map, depth, and neural radiation field based.

A. VOXEL REGRESSION BASED RECONSTRUCTION

In 2D space, each 2D image is composed of a single pixel, and if the smallest unit in 3D space is abstractly represented according to the 2D method, each 3D model can be composed of a single volume element, referred to as voxel. If the resolution of the voxels composing the 3D model is higher, the more concrete the 3D model is. Jackson et al. [31] proposed an end-to-end 3D face reconstruction network (VRNet) in 2017, which converts each accurately scanned 3D face into a voxel binary, where the portion of pixels containing the face takes the value of 1, and the value of 0 otherwise, as shown in Fig. 18. The algorithm innovatively transforms 3D face reconstruction into a 2D to 3D segmentation task. The network framework consists of two hourglass networks, the first one reconstructs the rough 3D face and the second one gets the fine 3D face by correcting the error.

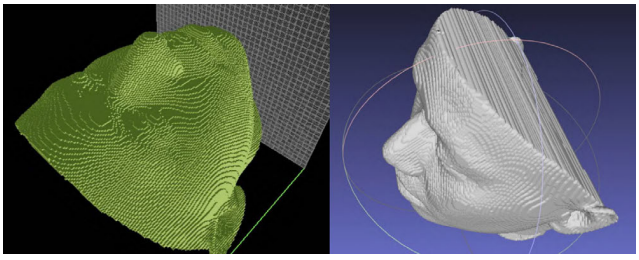

FIGURE 18. 3D face structure based on voxel representation.

The voxel-based regression method has not been widely used for the following reasons: (1) 3D face data acquisition is costly and difficult. (2) In terms of resolution, 3D faces

represented by voxels will be lower than those represented by meshes. (3) The voxels are too sparse and discrete in 3D space, and some details of the face cannot be well expressed.

B. UV POSITION REGRESSION BASED RECONSTRUCTION

High-quality 3D point cloud data contains more positional information and details, which can be of great help when performing 3D face regression. However, the acquisition process of 3D point cloud is very tedious and complicated, which requires the use of laser scanning equipment to acquire face information and save it in the form of points. Each point contains coordinate information (x, y, z) in 3D space. If both shape and texture information are saved while scanning the face, a point cloud of information (x, y, z, p, q) containing sparse coordinates is formed. The disadvantages of using point clouds to represent 3D face information are obvious; the point cloud data needs to be traversed when performing the fitting and the point cloud data is saved in an unorganized manner, which not only increases the time and difficulty of the fitting process. Inspired by Aaron et al. [31], Feng et al. [60] proposed a method to represent 3D face point cloud coordinates using a 2D UV position map in 2018, which is based on the principle of mapping the 3D point cloud coordinates according to the three channels of RGB to the UV position map as shown in Fig. 19. The UV position map can record the geometric information of the 3D face in its entirety, which cleverly transforms the reconstruction problem into a UV position image prime value prediction problem.

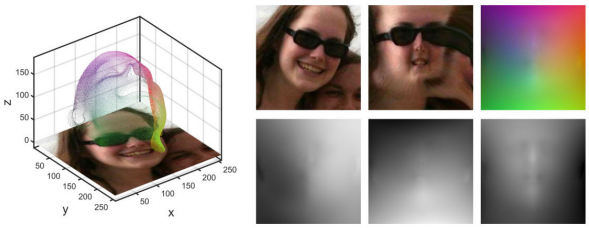

FIGURE 19. The illustration of UV position map. Left: is the input face image and its corresponding 3D point cloud, right: the first row represents the 2D face image, UV texture mapping map, and UV position map, respectively. The second row represents the x, y, and z channels of the UV position map, respectively [60].

PRNet [60] is a network capable of reconstructing a face model without the aid of 3DMM while still achieving dense alignment, which uses a simple residual convolutional neural as an encoder to encode the features of the input image, regresses the UV position map of the face through back-convolution, integrates a weight mask into the loss function during training, and also improves the performance of the network by adding the Attention Mechanism [61] to the UV position map. Chen et al. [62] encoded a 2D face into implicit vectors by a 3DMM encoder and input it into the UV space along with the reconstructed rough model to estimate the displacement depth map of the face. This method fused 3DMM parameters with the UV position map, which not only preserves the face detail information but also obtains the depth information of the face. Kao et al. [63] designed a deep neural network (PerspNet) to solve the problem of perspective projection distortion caused by a person's close proximity to the camera or moving perpendicularly to the camera axis. The face UV position map is extracted by an encoder-decoder network, and the 3D face shape is reconstructed in the canonical space, and the perspective projection is accurately represented by learning the correspondence between the 2D pixels and the 3D point cloud, and then estimating the face pose. Lin et al. [64] proposed an affine convolutional network, which not only preserves the face details but also obtains the depth information. an affine convolutional network that aims to solve the problem that the spatial misalignment between the input face image and the output UV position map causes the network to be ineffective in the feature encoding and decoding process. The authors considered that the affine transformation matrix of each spatial location in the feature map is learned by the affine convolutional layer, which regresses part of the parameter vectors of the diffuse mapping, UV position map, illumination map, and head pose, respectively, through four sub-networks, and fuses them to reconstruct a 3D face. Bai et al. [65] constructed a large-scale facial UV-texture dataset (FFHQ-UV), and designed a UV generator and texture decoder based on StyleGAN, which is capable of generating multi-view normalized UV position maps and texture parameters for fitting from a single face image. Experiments show that the method is able to generate high-quality texture maps to enhance the details of the model.

The advantage of 3D face reconstruction based on UV position map regression is that it gets rid of the linear space of 3DMM, and the high-dimensional UV space has better operability, and the reconstructed 3D face model is more realistic, and at the same time, the structure of the encoder-decoder makes the network lighter, and the operation efficiency has been greatly improved, but the mapping between the input 2D face image and the UV position map is more complicated to solve. However, the mapping between the input 2D face image and the UV position map is more complicated, which is easy to cause problems such as accuracy degradation.

C. DEPTH REGRESSION BASED RECONSTRUCTION

Depth map is different from mesh and point cloud in that they do not preserve all the 3D information, but use the form of a 2D image to represent the information in 3D space. The deeper the location in 3D space, the higher the color contrast in the depth map, and vice versa, the lower the color contrast, as a way of representing the depth information in the 3D information. Sela et al. [29] concluded that although the method based on 3DMM prior information simplifies the face reconstruction process, its low-dimensional linear space leads to the limitations of the model in the expression of the details, and they utilized an image translation network to map the input face image pixels into the depth values for regression, and the method has better real-time performance and robustness. Their reconstruction results are shown in Fig. 20.

FIGURE 20. Depth based 3D face reconstruction results [29].

Unsup3D [66] is an unsupervised 3D face reconstruction method, specifically, the method designs multiple encoder decoders to decompose 2D face images into Depth map, light map, albedo mapping, camera parameters, and confidence, and reconstructs a 3D face model by synthesizing these factors, and experiments show that the algorithm is not only capable of reconstructing high-precision faces, but also reconstructing faces with approximately symmetric nature. Zeng et al. [67] proposed a reconstruction method based on unsupervised training (DF2Net), whose network framework is shown in Fig. 21. The network consists of three different sub-networks, the first network D-Net maps the input image into a dense depth map, the second network F-Net improves the output of D-Net by integrating the features in the depth map and RGB domains, and the third network Fr-Net serves to enhance the resolution of the RGB-Depth, and the individual networks establish jump connections between multi-layer features. Experiments show that the reconstructed 3D face model has realistic shape and texture details, but the reconstruction speed is slow.

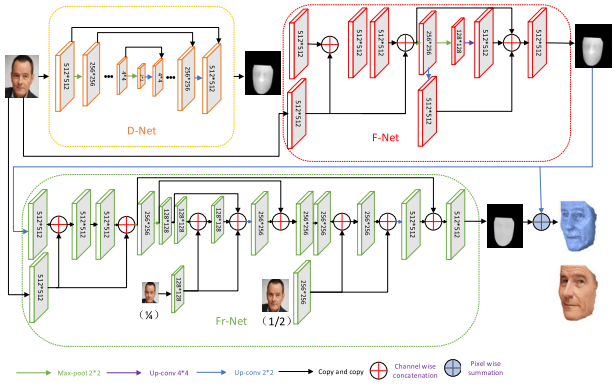

FIGURE 21. DF2Net network architecture [67].

The advantage of 3D face reconstruction based on RGB-Depth regression is that it can facilitate the data processing for generating 3D faces, and the geometry of its reconstructed model is more realistic, but it suffers from the problems of slow reconstruction rate and difficulty in obtaining the depth information used for supervised training.

D. NEURAL RADIATION FIELD REGRESSION BASED RECONSTRUCTION

In recent years, Mildenhall et al. [8] applied neural radiation fields (Nerf) to the task of new viewpoint synthesis, and achieved very stunning rendering synthesis results. Nerf is actually based on neural network rendering that implicitly encodes spatial geometric information about the target through the predicted Density attribute, which in turn allows for the rendering results from different viewpoints to maintain quite excellent rendering consistency. Gafni et al. [68] introduced a dynamic neural radiation field in order to deal with the material properties and complex geometries of 4D face avatars using stereoscopic rendering, which can capture arbitrary geometries and topologies such as hair, eyepatch, etc., of the input image, and which, in contrast to other volumetric methods that require expensively calibrated multiview devices, requires only a stationary camera to obtain a single view to render a reconstructed 3D face model. reconstructed 3D face model rendered onto the image. In order to solve the problem of view inconsistency in neural radiation fields, Chan et al. [69] proposed a generative model based on the combination of Nerf and GAN called Periodic Implicit Generative Adversarial Network (Pi-GAN), which utilizes a periodic activation function and volumetric rendering to represent the scene as a view-consistent radiation field, improving the quality of the 3D model of the face. Hong et al. [70] proposed a rendering method based on Nerf fusion face parametric model, where the authors used 3DMM to initialize the implicit coding of each image, and the low-dimensional base vectors of identity, expression, and reflectance were rendered together with the camera parameters to obtain the generated face image. The advantage of this method is the real time and the ability to control the camera, pose and other parameters separately. FENerf [71] is a neural radial field based face reconstruction method, the structure of its network model is similar to that of Pi-GAN, a single mapping network is used to generate the corresponding shapes and textures in a spatially aligned 3D voxel with a shared geometry to serve as the input to the network, the network's The output of the network is rendered by the voxel to get a 2D face image and a semantic mask, and the two discriminators of the GAN are used to determine whether the image is true or false, and a loss is established to constrain the network, which is conducive to the generation of more detailed face geometries. The results based on the Nerf reconstruction are shown in Fig. 22.

Nerf is able to reconstruct higher definition and more realistic 3D faces due to its neural volume based render, and its texture accuracy is higher than other face reconstruction methods, but due to its own limitations, it has obvious drawbacks: (1) it is not possible to reconstruct faces in real time it needs to be optimized scene-by-scene (2) it is time-consuming (3) it requires a high-performance GPU.

FIGURE 22. Results based on Nerf reconstruction [70].

VI. FACE DATASETS, EVALUATION METRICS AND COMPARISON OF METHODS

A. FACE DATASETS

Face databases are the basis for 3D face reconstruction and evaluation. Currently, there are a large number of 2D face datasets containing a variety of difficult samples (occlusion, large poses, etc.) that can be used to improve the reconstruction performance. Table 3 shows the current popular 2D face datasets.

To reconstruct a 3D face using 3DMM theory, a 3D face dataset is first needed. Blanz and Vetter [19] proposed an acquisition method to scan out 3D face data through a 3D scanner in their 1999 paper, but did not open source their database. Paysan et al [32]. in 2009 acquired 3D face data from 200 young people by an advanced ABW-3D structured light scanning system, and used an improved alignment algorithm to reduce shape artifacts, as well as an optical flow algorithm to densely align the data, keeping the vertices of the faces in one-to-one correspondence, to constitute the Basel Face Model (BFM), which made it possible to utilize the library of models The quality of 3D faces reconstructed using this model library as a priori information has been greatly improved. In order to solve the shortage of 3D face datasets, Booth et al [72]. established the Large Scale Morphable Model (LSFM), which collects 9663 faces with different identities and uses a 3D face with a fixed topology as a template for non-rigid matching of different 3D face data. face model, it can also build personalized face models according to different groups. However, none of the above models contains expression coefficients, so some researchers proposed to add a pair of expression bases in order to better portray the expression details. Chen et al [33]. added expression principal components to the BFM by capturing the expressions of 150 faces aged between 7 and 80 years old, which can be used to fit 3D faces with different expressions by controlling the change of the expression parameter. The authors of the BFM also provided the expression coefficients in the subsequent The author of BFM also provided expression coefficients in the subsequent release of BFM2017 [73], which contributed to the development of 3D face reconstruction. In 2017, MAPI open-sourced the FLAME dataset [74], which contains 5,023 vertices and four joints for each 3D face model. Specifically, FLAME splits the head into four parts: the left eyeball, the right eyeball, the chin, and the neck, which can be wrapped around customized ''joints''. These four parts can be rotated around the customized ''joints'' to form a new 3D representation. Three heterogeneous datasets are used to train the model, which not only contain rich expressions and postures, but also include the 3D geometry of the head, and represent the texture details of the face by aligning with the mesh of the BFM. The face models of BFM and FLAME are shown in Fig. 23. Table 4 shows the currently available 3D face datasets.

B. EVALUATION METRIC

The main metrics typically used to measure the effectiveness of 3D face reconstruction are:

1) The normalized mean error (NME) is commonly used in 3D face reconstruction work to measure the accuracy of facial landmarks detection. Its formula is as follows:

$$NME(P, \hat{P}) = \frac{1}{N} \sum_{i=1}^{M} \frac{\|p_i - \hat{p}_i\|_2}{d} \tag{19}$$

where p_i and $\hat{p}_i$ denote the facial landmarks of the reconstructed face and the real face, respectively, d is the normalization factor, usually the distance between the pupils of both eyes or the distance between the corners of the eyes of the outer corners of the two eyes, and N is the number of landmarks. This metric evaluates the facial landmarks error between the reconstructed 3D face rendered back to the 2D image and the input 2D image.

2) Root Mean Square Error (RMSE): when comparing the similarity between the reconstructed face and the original face image, the RMSE between the reconstructed face image and the original image identity feature vectors is usually extracted by a face recognition network and the RMSE between them is calculated.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=0}^{N-1} \| l_i - l_i' \|_2^2}{N}} \tag{20}$$

where N is the number of output images, l_i and l_i' denote the face feature vectors extracted by the face recognition network from the original face image and the reconstructed face image.

3) peak signal-to-noise ratio (PSNR): Typically used to evaluate the reconstruction quality of the textured part of a 3D face, For an $m \times n$ face image I, which is reconstructed and rendered, the mean square error (MSE) of the resulting synthesized image I_R can be expressed as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - I_R(i,j)]^2 \tag{21}$$

Then its PSNR can be expressed as:

$$PSNR = 10 \cdot \lg\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \lg\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{22}$$

where MAX_I represent the maximum pixel value of an image, usually expressed in binary form.

4) Scale Invariant Depth Error (SIDE) and Mean Angular Error (MAD) are commonly used to evaluate 3D face reconstruction based on RGB-Depth regression, given a 2D face

TABLE 3. Commonly used 2D face datasets.

Dataset	Subjects	Images	Landmarks	Male%	Attributes	Year	Access
Multi-PIE[75]	337	750K	68	—	Pose, Expression, lighting	2010	Required to apply
Helen[76]	—	20330	194	—	Pose, Lighting, Expression, Occlusion, and Individual Differences	2012	downloadable
300-W[34]	300	600	68	—	Expression, occlusion, head pose	2013	downloadable
MTFL[77]	—	12995	68	—	Expression, occlusion, head pose	2014	downloadable
CelebA[78]	10177	202K	5	50%	Occlusion, pose, expression, lighting	2015	downloadable
VGG-Face[79]	2,622	2.6M	—	47%	Occlusion, pose, expression, makeup	2015	Required to apply
LFW[80]	5749	13K	—	77%	Occlusion, pose, expression	2018	downloadable
WFLF[81]	—	10K	98	—	Occlusion, pose, makeup, lighting, blur, expression	2018	downloadable
FFHQ[82]	—	—	70K	—	Occlusion, pose, expression	2020	Required to apply
FFHQ-UV[65]	—	—	50K	—	High-quality, evenly illuminated facial texture UV	2023	Required to apply

TABLE 4. Commonly used 3D datasets and 3DMMs.

Dataset	Subjects	Expression	Source	Male%	Age	Year	Access
BU-3DFE[83]	100	25	Structured Light System	44%	All ages	2006	Mentors or institutes only
BU-4DFE[84]	101	6	Structured Light System	43%	—	2008	Mentors or institutes only
BJUT-3D[85]	1200	3	Laser Scanners	50%	—	2009	Required to apply
FRGC Ver2.0[86]	465	—	Laser Scanners	—	All ages	2014	Mentors or institutes only
BP4D+[87]	140	7	Multi-view stereo matching	41%	All ages	2016	Required to apply
300W-3D[25]	—	—	Fitting with 3DMM	—	—	2016	downloadable
AFLW-3D[25]	—	—	Fitting with 3DMM	—	—	2016	downloadable
UHDB31[88]	77	1	Multi-view stereo matching	69%	—	2017	Required to apply
Facescape[89]	938	20	Multi-view stereo matching	—	All ages	2020	Mentors or institutes only
BFM2009[32]	200	No	Structured Light System	50%	All ages	2009	Required to apply
FacewareHouse[33]	150	Yes	Kinect RGBD	—	All ages	2014	Mentors or institutes only
Multilinear Wavelet [90]	99	Yes	—	—	—	2014	downloadable
LSFM[72]	9663	—	—	48%	All ages	2016	Mentors or institutes only
BFM2017[73]	200	Yes	Structured Light System	50%	All ages	2017	Required to apply
LYHM[91]	1212	Yes	—	50%	All ages	2017	Required to apply
FLAME[74]	3800	Yes	—	48%	All ages	2017	Required to apply
COMA[55]	12	Yes	—	—	—	2018	downloadable
MICA[41]	2315	Yes	Trained from 8 sub-datasets	—	—	2022	Required to apply

image, predict the depth map d in the canonical view, and then combine with the viewpoints to obtain the actual depth map $\bar{d}$. The depth map in the actual view is compared with the real depth map d^* to calculate the corresponding scale invariant depth error:

$$E_{SIDE}\left(\bar{d}, d^*\right) = \left(\frac{1}{WH}\sum_{uv}\Delta_{uv}^2 - \left(\frac{1}{WH}\sum_{uv}\Delta_{uv}\right)^2\right)^{\frac{1}{2}} \quad (23)$$

where $\Delta_{uv} = \log \bar{d}_{uv} - \log d^*_{uv}$, Rendering artifacts at object boundaries are usually ignored in the calculation process, comparing only valid depth pixels. MAD is used as an evaluation metric to measure the quality of the obtained surfaces, and the correlation value is obtained by calculating the normal angular deflection error corresponding to the predicted depth map and the true depth map. The correlation value is obtained by calculating the normal angle deflection error corresponding to the predicted depth map and the true depth map.

5) densely aligned chamfer error (DACE): It is usually used to calculate the distance between the aligned 3D face model and the neighboring vertices between the ground-truth in the

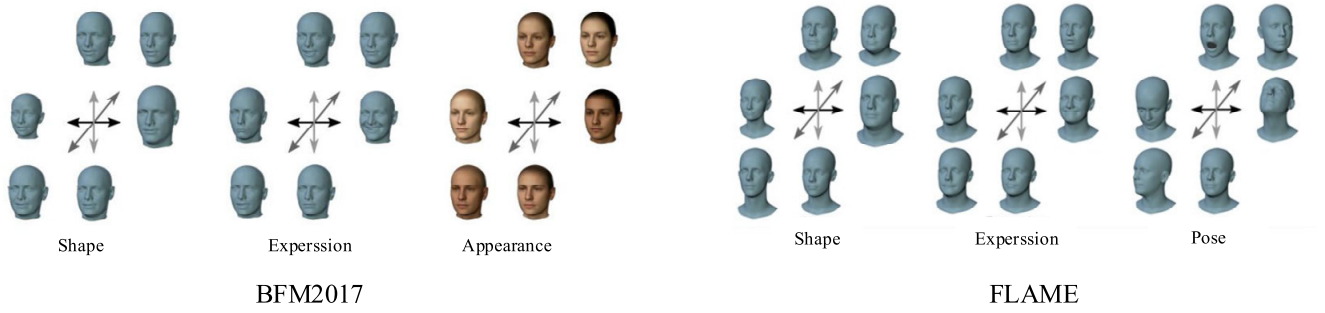

FIGURE 23. Model bases for BFM and FLAME.

dataset with the following formula:

$$DACE = \frac{1}{N(C)} \sum_{k \in C} \frac{\left\| v_k - v_{k^n}^* \right\|}{d} \tag{24}$$

where $N(C)$ denote the number of vertices, v_k denotes a point on the reconstructed face, $v_{k^n}^*$ denotes the nearest neighboring point of the point on the ground-truth provided by the 3D dataset, and d is the outer eye distance of the 3D face.

6) Frechet inception distance (FID): Usually used in 3D face reconstruction based on generative adversarial networks, it can be used to calculate the statistical similarity between the reconstructed face image and the input image in terms of visual features, with a lower score representing a higher similarity. the formula is as follows:

$$FID = \|u_1 - u_2\|_2^2 + T_r \left[C_1 + C_2 - 2(C_1 C_2)^{1/2} \right] \tag{25}$$

where u_1 and u_2 are the characteristic means of the input image and the synthesized image, $T_r[\cdot]$ is the trace of the matrix and C is the covariance matrix of the image features.

C. COMPARISON OF TYPICAL METHODS

With the advancement of 3D face reconstruction techniques, it is seen that the reconstruction accuracy and speed are improving, and the algorithms are more robust and generalizable. Whether it is 3D face reconstruction based on hybrid learning or display-based regression, both types of methods have their own advantages and disadvantages. Table 5 shows the advantages and disadvantages of representative methods in recent years.

VII. APPLICATION OF 3D FACE RECONSTRUCTION

A. 3D FACE RECOGNITION

Face recognition is a technology that uses the face as a biometric trait for identification, and compared with fingerprints, iris and other biometrics, face recognition has a broader application prospect. Although 2D face recognition has made significant progress in recent years, its accuracy is limited by factors such as posture, illumination and occlusion, and 3D face recognition can make full use of the reliable facial geometry of the face to overcome the shortcomings, and based on the reconstruction of the results for recognition, the principle of this class of methods is simple, the solution speed is fast, and the topology is known [92]. The flow of 3D face recognition is shown in Fig. 24.

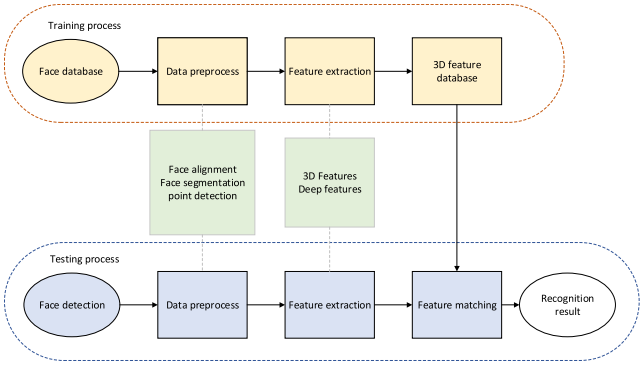

FIGURE 24. 3D face recognition process.

B. DIGITAL ENTERTAINMENT

The field of computer graphics and vision has been developing various tools and digital images of real human faces reproduced in computers for a long time, and these digitized face models are often used in movies, 3D games [93], digital stand-ins, Augmented Reality (AR), Virtual Reality (VR), and so on. An example of generating a game character based on a single image is shown in Fig. 25.

FIGURE 25. Generate game characters from a single image [93].

C. FACE VIDEO EDITING

Synthesizing high-definition face video sequences matching speech has a wide range of applications in chatbots, virtual video conferencing, etc. The problem can be regarded as a cross-modal mapping problem from speech to face.

TABLE 5. Comparison of typical algorithms for 3D face reconstruction based on a single image.

Year	Methods	Network type	Pros	Cons	Dataset	Performance	Principle
1993	SFS [15]	——	Wide range of application scenarios	Susceptible to light sources	——	——	The change in brightness of the imaged surface is used to reso- lve vector information about the surface of the object, which is converted to surface depth information.
1999	3DMM [19]	——	Representation of 3D faces using linear basis vector coefficients (shape, texture, etc.)	Linear space without detailed features	3DMM	——	A linear subspace is formed by PCA and combined with basis vector coefficients to form a specific 3D face
2016	Richardson et al. [20]	CNN	Training on coarse and detail respectively, the reconstructed model is able to represent the detail better	Poor generalizability	FRGC V2.0	SIDE:3.22m m	Combining CNN with 3D face reconstruction for the first time based on model optimization
2016	3DDFA[25]	CNN	First proposed PNCC, robust to occlusion and large attitudes	Requires large amounts of training data	AFLW 2000-3D	NME:5.42%	Converting the common 2D face feature point labeling problem to a 3D fitting task
2017	Nonlinear 3DMM [44]	Encoder Decoder	Only a single "in the wild 2D image" is needed to get the nonlinear 3DMM coefficients.	Poor detail	300W-LP	NME:4.70%	Regression of 3DMM coefficients by auto-encoder
2017	Jackson et al. [31]	Encoder Decoder	An End-to-End Approach to Regressing Voxel Representations of 3D Faces	Accurate voxel information is difficult to obtain	AFLW 2000-3D, BU-4DFE, MICC	NME:6.37%, 5.55%, 5.09%	Representation of 3D faces by CNN prediction of 3D binomials
2018	PRNet [60]	Encoder Decoder	2D representation of 3D faces using UV positional maps, end-to-end	Topology of 3D faces is unknown	Now, AFLW2000-3D	NME: 1.98%, 3.62%	Mapping 3D faces into UV space converts the reconstruction problem into a 2D UV position prediction problem
2019	GANFIT [52]	GAN DCNN	Combining GAN with 3D face reconstruction	Poor realism	MICC	Mean error using point-to- plane distance: 0.95	Using GAN to generate texture details
2020	DF2Net [67]	Encoder Decoder	Depth map based regression of 3D faces with high detail and realism, end-to-end	Slower reconstruction	BU-3DFE	SIDE:3.37m m	Regressing 3D faces by predicting depth maps
2020	Lin et al. [50]	Encoder Decoder GCN	High-fidelity textures, no need for large-scale face texture databases	Poor reconstruction of occluded faces	CelebA	PSNR: 29.69	Refining the color of mesh vertices with GCN

TABLE 5. Comparison of typical algorithms for 3D face reconstruction based on a single image.

2021	DECA [47]	Encoder Decoder	Supervised training using consistency loss, robust to occlusion and noise	Slower reconstruction	NoW	NME:1.38%	Predicting 3DMM coefficients by encoder and enhancing face details using coarse and detail training
2022	FENerf [71]	DCNN MLP GAN	The reconstructed 3D face is highly accurate and has a strong sense of realism	Slow reconstruction, requires high performance GPUs	FFHQ	FID:28.2	Generating face semantic masks and textures in 3D voxels using decoupled latent encodings to supervise the training of Nerf generators
2022	HeadNerf [70]	DCNN MLP	The reconstructed 3D face is sufficiently detailed and realistic, and hair can be reconstructed.	Slow reconstruction, requires high performance GPUs	FFHQ	PSNR:24.9	Based on the a priori information of the face, the feature map is predicted using Nerf for differential rendering
2023	HRN [43]	Encoder Decoder	Reconstructs detailed face models in both single and multiple views	Produces inaccurate results for occluded faces	FaceScape	NMSE:0.065 rad	Facial Geometry Decoupling via Hierarchical Representation Learning while Incorporating a Detailed 3D Prior to Improve Reconstruction Accuracy and Visualization, and Proposing De-Retouching Module to Mitigate Ambiguity between Geometry and Appearance

1) FACE PUPPET

Face puppetry refers to the technology of driving a virtual image through real face video input, which aims to migrate the expressions and emotions of a real user to a virtual facial puppet. An example of a face puppet is shown in Fig. 26. which is implemented in two main ways: expression coefficients with semantics or a dense motion field between the user and the virtual puppet. The first approach directly learns a linear mapping from the user's expression base to the target's expression base [94], while the dense motion field-based approach extracts the 2D dense motion field of the face from the original video and maps it to the 3D geometric motion field of the puppet image [95].

FIGURE 26. Real-time face puppets [94].

2) SPEAKER VIDEO GENERATION

Mapping the speech signal input to the corresponding speaker video sequence through neural networks, this type of method requires the help of 3D models or facial landmarks as intermediate modalities to establish the link between the speech signal and the final video, Thies et al. [96] establish the mapping between speech input and face generation with the help of expression coefficient-controlled 3D face models, and render realistic speaker video results through delayed rendering techniques. The algorithmic framework for cross-modal video generation is illustrated in Fig. 27. Chen et al. [97] mapped speech input to facial 2D landmarks and generated network sequences using Generative Adversarial Networks, but the method was unable to control the facial poses in the generated video.

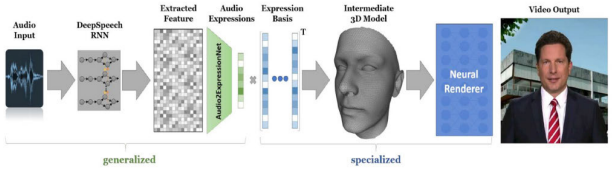

FIGURE 27. The algorithm of Thies et al. input a speech signal and synthesize the corresponding video sequence of the speaker with the help of 3D model [96].

D. FACE ATTRIBUTE EDITING

A 3D face model with new features is generated by changing some attributes of the face (pose, age, shape, texture, etc.) and it is used in some specific scenarios: face replacement, expression migration, virtual make-up, etc.

1) FACE REPLACEMENT

Face replacement is a technique for replacing a face in a target video with another face based on parameters such as identity trajectory, facial features and facial expression, where the difficulty lies in synthesizing a realistic video sequence while maintaining temporal continuity. Depending on the technique used, these methods can be categorized into 3D face reconstruction based algorithms and image based algorithms. Among them, model-based algorithms need to reconstruct the pose, expression, etc. of the source and target faces, and render the source face according to the characteristics of the target attributes [98], such as in Fast and Furious 7, where the face of the deceased protagonist is swapped to the face of the stand-in actor to complete the subsequent shooting of the movie. The image-based algorithm combines image retrieval and face migration techniques by selecting a key frame from the source face video and morphing that frame to the target face [99].

2) EXPRESSION TRANSFER

Face expression transfer means an editing technique to migrate the expression of a source face to a target face. Researchers generally need to reconstruct the identity, expression, posture, illumination and albedo parameters of the source and target faces, and then render the expression parameters of the source face to the target face. Thies et al. [96] proposed the first algorithm for real-time expression substitution based on an RGB-D camera, which utilizes a parametric the algorithm utilizes a parametric model to extract the pose, illumination, and other parameters of the face and uses them as a priori information to reconstruct the facial movements of the source and target faces to achieve the result of expression transfer. Wu et al. [100] accomplished the transfer of source to target face expressions using a self-encoder-based model, and the results are shown in Fig. 28.

FIGURE 28. Expression transfer using ReenactGAN [100].

3) VIRTUAL MAKE-UP

Face virtual makeup refers to the technique of altering the texture features of certain regions of a 3D face model and then re-rendering them to achieve makeup changes. Garrido et al. [101] reconstructed a 3D face model from monocular RGB data, and then added virtual makeup to the face image based on texture mapping. Li et al. [102] simulated the effect of makeup on the face by physically based face texture editing. The face make-up technique is shown in Fig. 29.

FIGURE 29. Synthesised virtual tattoos technology [101].

VIII. FUTURE PROSPECTS

In recent years, with the development of deep learning, a variety of representative algorithms have emerged, which have pushed forward the progress of 3D face reconstruction technology, but still face many challenges:

1) Data sets are scarce and difficult to obtain: deep learning based 3D face reconstruction requires the use of a large number of samples as training data, and these samples also need to have 2D images and their corresponding 3D geometries, although it is possible to take the fitting of 3DMMs to real images to generate the training samples, but this method of augmenting the data with synthetic samples reduces the accuracy of the reconstruction model.

2) The reconstructed face model has poor individualized details: the reconstruction method that introduces 3DMM as a priori information has linear limitations, and the reconstructed model tends to lose some personalized features of the face (e.g. wrinkles, moles, etc.).

3) Poor network generalization and poor robustness: when the test image contains faces with large poses or self-obscuring, it leads to reconstruction failure.

4) High algorithm complexity and low reconstruction efficiency: 3D face reconstruction based on hybrid learning and based on neural radiation field has the problem of more complex algorithms, and the neural radiation field needs to be optimized scene by scene, and the reconstruction efficiency is low.

5) Inability to reconstruct areas other than the face: most algorithms today only focus on reconstructing details of the face, and are unable to reconstruct areas such as hair, mustache, and teeth.

6) Lack of standardized evaluation metrics: In the comparison of 3D face reconstruction accuracy, AvatarMe and 2DASL have a normalized mean error of 3.53, both of which show good alignment accuracy, but the former aims to reconstruct 3D faces with high-frequency details, while the latter aims to improve the shortage of data annotations between the image and the model, and most of the existing evaluation metrics compare the input image with the rendered synthetic

image and it is difficult to evaluate the reconstructed 3D face model.

7) Models are difficult to run on mobile devices: Most of the 3D face reconstruction methods rely on high-performance GPUs for training and testing and cannot be used for entertainment applications on mobile devices.

To address the above problems, the following points can be improved or solved:

1) To address the problem of insufficient 3D datasets, unsupervised or self-supervised training can be adopted in the algorithm design, and a more efficient loss function can be designed to constrain the model. In the future, we should explore more efficient data collection methods, produce and open source more datasets, which should contain changing face images (occlusion, lighting, scene, etc.) and different ethnic attributes (yellow, black, white, etc.).

2) To address the problem of poor personality details, we can refer to Lin et al. [50] to introduce GCN to deal with non-Euclidean structures such as faces, or we can refer to the method of GANFIT [53], which combines with GAN to generate realistic textures, and we can use face semantic masks and attention mechanisms to improve the network's attention to the details of the face parts, e.g., FOCUS [103], SADRNet [104], and so on.

3) To address the problem of poor network generalization and robustness, some attribute changes (lighting, skin color transformations, etc.) can be made to the face samples in the data to achieve the effect of enhancing the data, e.g., StyleGAN [82], Adv-MakeUP [105], and at the same time, some special samples (large pose, self-obscuration, weak lighting, etc.) can be added to the training data to train the model.

4) To address the problems of high algorithm complexity, low efficiency, and inability to be deployed on mobile devices, we can consider introducing some scenario a priori information to improve the optimization efficiency, e.g., Plenoxles [106], mip-Nerf [107], and adopting some common means of lightweighting (distillation, model pruning, and low-rank decomposition, etc.) in the network structure, e.g., 3DDFA_V2 [28].

IX. CONCLUSION

In this paper, we provide a detailed review of 3D face reconstruction based on a single image. In the beginning of the paper, the relevant physical knowledge in 3D face reconstruction is introduced, followed by a review of the relevant methods for 3D face reconstruction in the past time, and three main approaches are delineated: traditional reconstruction methods, hybrid learning-based methods, and face reconstruction based on explicit regression. Some representative algorithms are highlighted and the advantages and disadvantages between each algorithm are compared. After that, some commonly used face datasets, and some metrics used to measure 3D face reconstruction are introduced, in addition, applications of 3D face reconstruction, including 3D face recognition, digital entertainment, face video editing, and face attribute editing, are also introduced. Finally, some problems in the current 3D face reconstruction task are discussed, and some ideas for solving and improving them are given. In summary, 3D face reconstruction based on a single image is an open research area.

REFERENCES

[1] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt, "State of the art on monocular 3D face reconstruction, tracking, and applications," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 523–550, May 2018.

[2] S. Sharma and V. Kumar, "3D face reconstruction in deep learning era: A survey," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 3475–3507, Aug. 2022.

[3] J. T. Wang and H. B. Li, "Review of single-image 3D face reconstruction methods," *Comput. Eng. Appl.*, vol. 59, no. 17, Oct. 2022.

[4] A. C. Sauve, A. O. Hero, W. L. Rogers, S. J. Wilderman, and N. H. Clinthorne, "3D image reconstruction for a Compton SPECT camera model," *IEEE Trans. Nucl. Sci.*, vol. 46, no. 6, pp. 2075–2084, Dec. 1999.

[5] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Portland, OR, USA, Jun. 2013, pp. 3444–3451.

[6] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 497–500.

[7] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.

[9] F. I. Parke, "Measuring three-dimensional surfaces with a two-dimensional data tablet," *Comput. Graph.*, vol. 1, no. 1, pp. 5–7, May 1975.

[10] L. Kovacs, A. Zimmermann, G. Brockmann, M. Gühring, H. Baurecht, N. A. Papadopulos, K. Schwenzer-Zimmerer, R. Sader, E. Biemer, and H. F. Zeilhofer, "Three-dimensional recording of the human face with a 3D laser scanner," *J. Plastic, Reconstructive Aesthetic Surgery*, vol. 59, no. 11, pp. 1193–1202, Nov. 2006.

[11] T. Bell, B. Li, and S. Zhang, "Structured light techniques and applications," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, 1999, pp. 1–24.

[12] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[13] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Mach. Vis. Appl.*, vol. 12, no. 1, pp. 16–22, Jul. 2000.

[14] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[15] R. Zhang, P. S. Tsai, C. E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[16] W. A. P. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1914–1930, Dec. 2006.

[17] J.-F. Hu, W.-S. Zheng, X. Xie, and J. Lai, "Sparse transfer for facial shape-from-shading," *Pattern Recognit.*, vol. 68, pp. 272–285, Aug. 2017.

[18] M. Castelán and E. R. Hancock, "Acquiring height data from a single image of a face using local shape indicators," *Comput. Vis. Image Understand.*, vol. 103, no. 1, pp. 64–79, Jul. 2006.

[19] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Seminal Graphics Papers: Pushing the Boundaries*, vol. 2, 2023, pp. 157–164.

[20] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5553–5562.

[21] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1503–1512.

[22] Y. Luo, X. Tu, and M. Xie, "Learning robust 3D face reconstruction and discriminative identity representation," in *Proc. IEEE 2nd Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, Weihai, China, Sep. 2019, pp. 317–321.

[23] X. Fan, S. Cheng, K. Huyan, M. Hou, R. Liu, and Z. Luo, "Dual neural networks coupling data regression with explicit priors for monocular 3D face reconstruction," *IEEE Trans. Multimedia*, vol. 23, pp. 1252–1263, 2021.

[24] W. Zhu, H. T. Wu, Z. Chen, N. Vesdapunt, and B. Wang, "ReDA: Reinforced differentiable attribute for 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4957–4966.

[25] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 146–155.

[26] N. Chinaev, A. Chigorin, and I. Laptev, "MobileFace: 3D face reconstruction with efficient CNN regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV), Workshops*, 2018.

[27] P. Wang, Y. Tian, W. Che, and B. Xu, "Efficient and accurate face shape reconstruction by fusion of multiple landmark databases," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 335–339.

[28] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 152–168.

[29] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1585–1594.

[30] F. Liu, R. Zhu, D. Zeng, Q. Zhao, and X. Liu, "Disentangling features in 3D face shapes for joint face reconstruction and recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5216–5225.

[31] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1031–1039.

[32] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Genova, Italy, Sep. 2009, pp. 296–301.

[33] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, Mar. 2014.

[34] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 896–903.

[35] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt, "MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3735–3744.

[36] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 285–295.

[37] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, and J. Feng, "3D face reconstruction from a single image assisted by 2D face images in the wild," *IEEE Trans. Multimedia*, vol. 23, pp. 1160–1172, 2021.

[38] Y. Wu and L. Dong, "3D face shape and texture reconstruction based on weak supervised learning," *Appl. Comput. Syst.*, vol. 29, no. 11, pp. 183–189, 2020.

[39] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5548–5557.

[40] J. Zhang, L. Lin, J. Zhu, and S. C. H. Hoi, "Weakly-supervised multi-face 3D reconstruction," 2021, *arXiv:2101.02000*.

[41] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," 2022, *arXiv:2204.06607*.

[42] H. Li, B. Wang, Y. Cheng, M. Kankanhalli, and R. T. Tan, "DSFNet: Dual space fusion network for occlusion-robust 3D dense face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada. IEEE, Jun. 2023, pp. 4531–4540.

[43] B. Lei, J. Ren, M. Feng, M. Cui, and X. Xie, "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Canada, BC, Canada: IEEE, Jun. 2023, pp. 394–403.

[44] L. Tran and X. Liu, "Nonlinear 3D face morphable model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7346–7355.

[45] X. Li, Z. Weng, J. Liang, L. Cei, Y. Xiang, and Y. Fu, "A novel two-pathway encoder–decoder network for 3D face reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 3682–3686.

[46] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3D face morphable model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA, Jun. 2019, pp. 1126–1135.

[47] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, Aug. 2021.

[48] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3D face decoding over 2500FPS: Joint texture & shape convolutional mesh decoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1097–1106.

[49] C. Yuan, K. Li, Y.-K. Lai, Y. Liu, and J. Yang, "3D face reprentation and reconstruction with multi-scale graph convolutional autoencoders," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*. Shanghai, China: IEEE, Jul. 2019, pp. 1558–1563.

[50] J. Lin, Y. Yuan, T. Shao, and K. Zhou, "Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5890–5899.

[51] Y. Qiu, X. Xu, L. Qiu, Y. Pan, Y. Wu, W. Chen, and X. Han, "3DCaricShop: A dataset and a baseline method for single-view 3D caricature face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 10231–10240.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, Quebec City, QC, Canada, 2014, pp. 2672–2680.

[53] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1155–1164.

[54] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou, "AvatarMe: Realistically renderable 3D facial reconstruction 'in-the-wild,'" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 757–766.

[55] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 704–720.

[56] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "MeshGAN: Non-linear 3D morphable models of faces," 2019, *arXiv:1903.10384*.

[57] G.-H. Lee and S.-W. Lee, "Uncertainty-aware mesh decoder for high fidelity 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6099–6108.

[58] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3D: Generative neural texture rasterization for 3D-aware head avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 20991–21002.

[59] J. Piao, K. Sun, Q. Wang, K.-Y. Lin, and H. Li, "Inverting generative adversarial renderer for face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 15614–15623.

[60] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Proce. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 534–551.

[61] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[62] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, and L. Bao, "Self-supervised learning of detailed 3D face reconstruction," *IEEE Trans. Image Process.*, vol. 29, pp. 8696–8705, 2020.

[63] Y. Kao, B. Pan, M. Xu, J. Lyu, X. Zhu, Y. Chang, X. Li, and Z. Lei, "Towards 3D face reconstruction in perspective projection: Estimating 6DoF face pose from monocular image," *IEEE Trans. Image Process.*, vol. 32, pp. 3080–3091, 2023.
[64] Z. Lin, J. Lin, L. Li, Y. Yuan, and Z. Zou, "High-quality 3D face reconstruction with affine convolutional networks," in *Proc. 30th ACM Int. Conf. Multimedia*. Lisboa Portugal: ACM, Oct. 2022, pp. 2495–2503.
[65] H. Bai, D. Kang, H. Zhang, J. Pan, and L. Bao, "FFHQ-UV: Normalized facial UV-texture dataset for 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 362–371.
[66] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3D objects from images in the wild," 2019, *arXiv:1911.11130*.
[67] X. Zeng, X. Peng, and Y. Qiao, "DF2Net: A dense-fine-finer network for detailed 3D face reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 2315–2324.
[68] G. Gafni, J. Thies, M. Zollhöfer, and M. Nießner, "Dynamic neural radiance fields for monocular 4D facial avatar reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 8645–8654.
[69] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 5795–5805.
[70] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, "HeadNeRF: A realtime NeRF-based parametric head model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 20342–20352.
[71] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang, "FENeRF: Face editing in neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 7662–7672.
[72] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 233–254, Apr. 2018.
[73] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schoenborn, and T. Vetter, "Morphable face models—An open framework," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 75–82.
[74] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, Dec. 2017.
[75] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[76] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.
[77] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc.13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 6–12.
[78] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. Santiago, Chile: IEEE, Dec. 2015, pp. 3730–3738.
[79] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, p. 41.
[80] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, Oct. 2008.
[81] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 2129–2138.
[82] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 4396–4405.
[83] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 211–216.
[84] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.
[85] Y. Baocai, S. Yanfeng, W. Chengzhang, and G. Yun, "BJUT-3D large scale 3D face database and information processing," *J. Comput. Res. Develop.*, vol. 6, p. 4, Jan. 2009.
[86] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 947–954.
[87] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 3438–3446.
[88] H. A. Le and I. A. Kakadiaris, "UHDB31: A dataset for better understanding face recognition across pose and illumination variation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2555–2563.
[89] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*. Paris, France: IEEE, Oct. 2014, pp. 343–347.
[90] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear wavelets: A statistical shape space for human faces," in *Proc.13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, 2014, pp. 6–12.
[91] H. Dai, N. Pears, W. Smith, and C. Duncan, "A 3D morphable model of craniofacial shape and texture variation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3104–3112.
[92] S. Zhou and S. Xiao, "3D face recognition: A survey," *Hum. Cent. Comput. Inf. Sci.*, vol. 8, no. 1, p. 35, Dec. 2018.
[93] J. Lin, Y. Yuan, and Z. Zou, "MeInGame: Create a game character face from a single portrait," in *Proc. AAAI*, May 2021, vol. 35, no. 1, pp. 311–319.
[94] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014.
[95] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "What makes tom hanks look like tom hanks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. Santiago, Chile: IEEE, Dec. 2015, pp. 3952–3960.
[96] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., 2020, pp. 23–28.
[97] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7824–7833.
[98] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister, "Video face replacement," in *Proc. SIGGRAPH Asia Conf.*, Dec. 2011, pp. 1–10.
[99] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormaehlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 4217–4224.
[100] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, "ReenactGAN: Learning to reenact faces via boundary transfer," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 603–619.
[101] P. Garrido, L. Valgaert, C. Wu, and C. Theobalt, "Reconstructing detailed dynamic face geometry from monocular video," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–10, Nov. 2013.
[102] C. Li, K. Zhou, and S. Lin, "Simulating makeup through physics-based manipulation of intrinsic image layers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4621–4629.
[103] C. Li, A. Morel-Forster, T. Vetter, B. Egger, and A. Kortylewski, "Robust model-based face reconstruction through weakly-supervised outlier segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 372–381.
[104] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. Wang, "SADRNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction," *IEEE Trans. Image Process.*, vol. 30, pp. 5793–5806, 2021.

[105] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," 2021, *arXiv:2105.03162*.
[106] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5491–5500.
[107] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5835–5844.

HAOJIE DIAO received the B.S. degree from Sichuan University of Science and Engineering, Sichuan, China, in 2022, where he is currently pursuing the degree with the School of Automation and Electrical Information. His current research interests include 3D face reconstruction and deep learning.

XINGGUO JIANG received the M.S. degree from Chongqing University, Chongqing, China, in 2003, and the Ph.D. degree from the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China, in 2007. He is currently an Associate Professor with the School of Automation and Electrical Information, Sichuan University of Science and Engineering, Sichuan, China. Prior to that, he was an Associate Professor with the School of Information and Communication, Guilin University of Electronic Technology, Guangxi, China. His current research interests include image processing, intelligent information processing, and deep learning.

YANG FAN received the B.S. degree from Sichuan University of Science and Engineering, Sichuan, China, in 2021, where he is currently pursuing the degree with the School of Automation and Electrical Information. His current research interests include image processing and deep learning.

MING LI received the B.S. degree from Jiamusi University, Heilongjiang, China, in 2022. He is currently pursuing the degree with the School of Automation and Information Engineering, Sichuan University of Science and Engineering, Sichuan, China. His current research interests include image processing and deep learning.

HONGCHENG WU received the B.S. degree from Sichuan University of Science and Engineering, Sichuan, China, in 2022, where he is currently pursuing the degree with the School of Automation and Electrical Information. His current research interests include image processing and deep learning.