IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2020, accepted October 1, 2020, date of publication October 27, 2020, date of current version December 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3031886

Full Face-and-Head 3D Model With Photorealistic Texture

YANGYU FAN[1], YANG LIU[2], GUOYUN LV[1], SHIYA LIU[2], GEN LI[1], AND YANHUI HUANG[1]

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
[2]Content Production Center of Virtual Reality, Beijing 101318, China

Corresponding author: Guoyun Lv (lvguoyun101@nwpu.edu.cn)

This work was supported in part by the Key Program of Research and Development Plan of Shaanxi Province under Grant 2020ZDLGY04-09 funded by the Department of Science and Technology of Shaanxi Province. This work was supported in part by the National Natural Science Foundation of China under Grant 62071384.

ABSTRACT In the recent period, significant progress has been achieved towards reconstructing the 3D face model from face image. With the support of the render engines and sufficient data, the reconstruction results are fine in detail. Nevertheless, the research on the 3D face reconstruction with texture from a single unrestricted face image is imperfect. The rebuild process lacks essential structure and texture information in the profile and the craniofacial region. To address this problem, we present a method of creating a 3D full face-and-head model with photorealistic texture from a single ''in-the-wild'' face image in this paper. To this end, we introduce a pipeline to integrate the highly-detailed face model into the basic model. Specifically, the basic model was built by multilinear optimization, and the highly-detailed face model which represents the facial features generated by constrained illumination distribution. Additionally, to infer the invisible region texture information corresponding to the input face image, we design an effective architecture with the generative adversarial network (GAN) for panoramic UV texture generation. The final results after UV texture mapping were visualized in the experiment, which demonstrates that the model faithfully recovers the photorealistic details in arbitrary perspective. Furthermore, compared to the state-of-the-art facial modeling techniques and existing commercial solutions, our method takes less input and performs better in surface detail.

INDEX TERMS 3D face reconstruction, full face-and-head model, 3D morphable model, UV texture, generative adversarial networks.

I. INTRODUCTION

The three-dimensional (3D) face-and-head reconstruction technology has important theoretical significance and practical value in our daily life, especially in Internet communications and artificial intelligence applications. Due to their powerful ability to represent human feature and the face contour, the 3D face-and-head models are intensively applied to face reconstruction [1], [2], computer graphics [3], [4], biometrics [5]–[7], and texture blending [8]–[10]. The research purpose of the 3D face-and-head model is to make it as realistic as possible. Observers can visually distinguish the model from texture and capture the tiny details. Furthermore, the source of the inputs is ''in-the-wild'' facial images, which can be acquired conveniently.

We employ the synthetic 3DMM to reconstruct the full face-and-head model from a single ''in-the-wild'' face image. The cranium and the face region are constructed with different statistical correlations. The resolution requirement and feature details of the two parts is distinctive, since the cranium part just needs to show the large-scale contour representation of the human head, whereas the facial part requires detailed feature information and better spatial resolution.

Our crucial contributions are: (1) we present a method that integrates the basic model with the highly-detailed face region to obtain a suitable full face-and-head model, (2) we introduce an effective network architecture for UV texture generation. Using Variational Auto-Encoder (VAE) as the identity constraints to control the GAN network training process, (3) we collect sufficient UV texture maps, including some UV textures from the useful dataset, and the high-fidelity textures obtained by our laboratory using professional 3D scanning equipment, a total of 1060 identities, (4) we perform an intact experiment to map the finished UV texture to the synthetic face-and-head model. Visible results are more realistic than models from the state-of-the-art

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

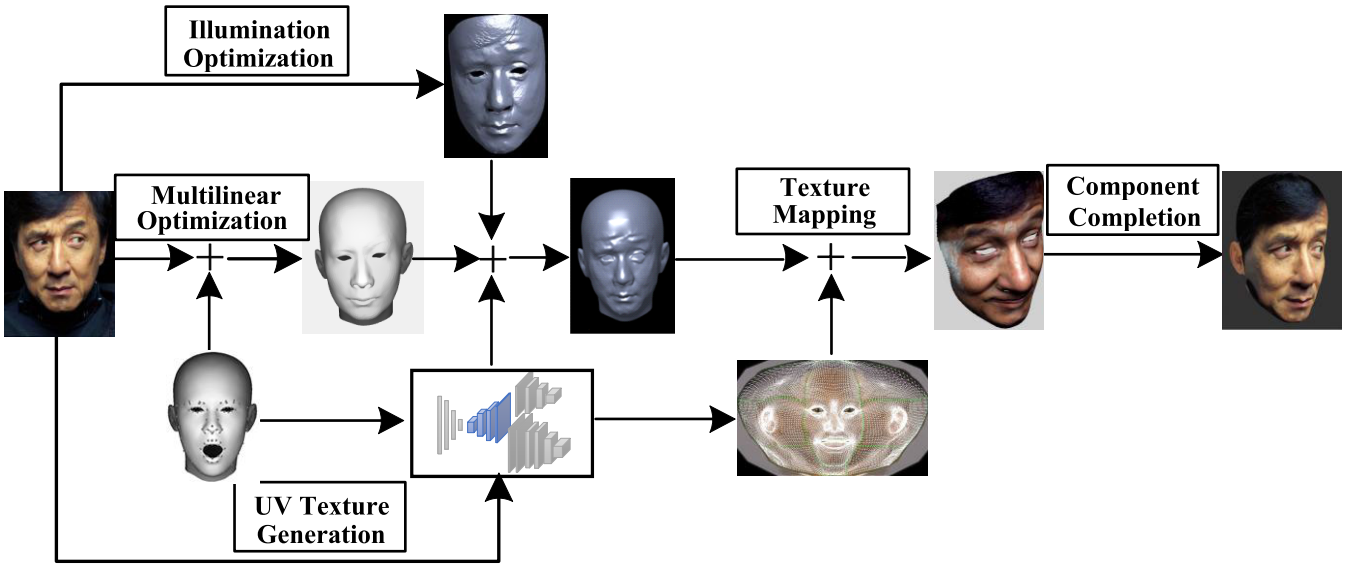

FIGURE 1. Overview of our proposed method. We build the basic model by joint optimizing of the input images and the database and integrate it with the highly-detailed model to obtain the full face-and-head mesh model. The UV texture can be generated by our network architecture with a single "in-the-wild" face image. The complete full face-and-head model is acquired after the texture mapping and the component completion.

modeling methods. The brief of the processing overview is shown in Fig. 1.

II. RELATED WORKS

The 3D face-and-head modeling has been extensively explored in the past twenty years and has been widely used in numerous applications. This technology can represent the face and head physiological characteristics, and mainly includes the 3D Morphable Model (3DMM) and the avatar digitization. Empirically speaking, 3DMM is generally used to represent the large-scale contour information of the human head and the appearance features of the facial region. Therefore, 3DMM technology is primarily applying to the fields of face recognition and 3D face reconstruction. Avatar digitization is mostly used for the game animation and social communication, which tends to the global exterior features and the surface texture resembles the digital real-time rendering. We summarize the relevant research work of the 3D face-and-head models according to the different application fields.

3DMM [11], [12] technology was first proposed by Blanz and Vetter in 1999s, the model (100 male and 100 female) were built from 200 scans. This method is the most outstanding and widely-used, many other ideas are also improved and optimized based on this method. The Basel Face Model (BFM) is built by Paysan [13], they register the scans by utilizing the Non-rigid Iterative Closest Point (NICP) algorithm. Classic Principal Component Analysis (PCA) is employed as a dimension descent method to construct 3DMM. With the improvement of the hardware scanning capability, recent works have established some [6], [14] database through a large number of face scans. Available face 3DMM database can refer to FaceWareHouse [15] and LSFM [16].

In [17], Ploumpis proposes a pipeline for combining current 3DMM of the human head to face and some other parts. He uses a regressor to accomplish missing parts of one model, the Gaussian Process framework is applied to blend covariance matrices from multiple models. However, this work has a flaw in texture reconstruction which may lead to the full head model to discard the partial surface information in the profile and top of the head region. A similar problem also appeared in [18]. An approach is proposed to improve the nonlinear 3D modeling in additional side-step regularization and leverages to increase detailed shape in [19]. This work builds a model with mid and high-level details in the face, but it is separately trained from 2D images which may lack geometry information in profile and back areas. In[20], a framework is presented to build a new face-and-head shape model that combines the variability and facial detail of the LSFM with the full head modeling of the LYHM. But this method ignores the texture parameters, and the new model cannot match the texture well.

Recently, some research work based on data-driven has achieved suitable progress on 3D face reconstruction. In [21], Ranjan has introduced a versatile mesh model that adopted a face non-linear representation with spectral convolutions on the mesh model surface. The data sample consists of 20466 meshes of extreme expressions captured over 12 subjects. In [22], a novel framework that learns a generative 3D face model using autoencoder architecture has been proposed. It combines the convolutional networks of being robust to corrupted data with the multilinear models of effectively modeling and decoupling shape variations. In [23], Cheng proposes the first intrinsic GANs architecture directly operating on the 3D mesh model. The network can generate expressions for random identities from latent spaces where identity and expression are mixed. In [3], the multi-frame video-based self-supervised training of a deep network learns a face identity model both in shape and appearance, while jointly learning to rebuild the 3D face. The approach produces high-quality monocular reconstructions of facial geometry

from ''in-the-wild'' data. In [24], Wu addressed the problem of recovering the 3D geometry of a humankind face from multi-view facial images and proposed a method to regress 3D model parameters from multi-view input with end-to-end trainable Convolutional Neural Network (CNN).

However, those data-driven methods demand a large data set, complex network architecture, and the training cost, all of these will limits the apply scope. Some other novel methods in 3D face reconstruction, such as [25] proposed by Yao. He designs a new representation called UV position map which records the 3D shape of a complete face in UV space. This method does not rely on any prior model information and the network is light-weighted in the experiment.

In the domain of avatar digitization, the destination is to rebuild the personal avatar resembles the captured subject. Existing 3D avatar creation systems rely on multiple images to build a more precise texture map [26]. In [8], a system for creating the fully rigged 3D avatar from a hand-held video is introduced. The system recovers the expression of the objects by adapting the blendshape template to video using an optimization, which includes optical flow and Shape from Shading (SFS). In [27], Hu introduces an end-to-end framework to create a complete 3D avatar from a single image for real-time rendering. Their methods and experiments are impressive, but the results are imperfect in details, and the surface display like ''animated face'' from a subjective perspective.

These methods we mentioned above are all related to the 3D face-and-head modeling. On the other hand, the face texture reconstruction mainly based on the texture statistical model as the prior information. Traditionally speaking, 3DMMs texture by the UV map which can assign 3D pixel data into 2D plane with per-pixel alignment. Nonetheless, the texture statistical model is to scan under strict controllable conditions to acquire the low-high frequency and albedo information. This kind of texture model is hardly obtained and it is also difficult for ''in-the-wild'' image analysis. In [28], a data-driven inference method that can synthesize the texture map of a complete 3D face model from the 2D view image. The middle layer feature can be extracted from a deep convolutional neural network, and the texture map is synthesized by iteratively optimizing. In [29], Deng has proposed a framework for training DCNN to finish the facial UV map extracted from partial face images. This architecture learns an identity-preserving UV completion model and minimizes the pose discrepancy during the training process. In [30], Gecer reconstructs the facial texture and shape from single images by GAN and DCNN and optimize the parameters with the supervision of pre-trained deep identity features through the end-to-end differentiable framework. In [18], the texture map is generated using raw texture images from five views, which means the face texture reconstruction is done under the restricted conditions. This pixel embedding method can maintain the same pixel resolution as the texture map and the raw texture images. In recent works, differentiable renderers were employed to learn the relationship between the facial identity features and the parameters of a 3DMM for shape and texture.

In conclusion, those methods based on 3DMM have some ill-conditioned matters in the multi-parameter information fitting process. The method based on data-driven is limited by the amount of data of the statistical models under controlled conditions and are prone to local optimization and gradient explosion problems. Additionally, with the demand of data, the latest methods based on GAN have uncontrollable training processes and feature couple in non-linear spaces. However, the large data cost and the complex device requirement makes these approaches above are unsuitable for applications. Furthermore, some other approaches lacking the ability to represent fine-scale features. Particularly, when we inspect the model in specific angles, there is an obvious blurring of texture on the profile, back, and bottom of the 3D model.

In contrast to the methods above, our full face-and-head models are built separately based on the display characteristics of each part, and more complete in structure than the current 3D face reconstruction methods. Additionally, the texture generation module uses the specially designed VAEGAN [31] framework to generate the corresponding UV texture map make the information in the invisible region available. Under the equivalent condition, our final results are better than the state-of-the-art methods in effect.

III. 3D FULL HEAD BASIC MODEL

According to the representative structure of the statistical model, the 3D face Morphable Model can be divided into a 3D full head model and a 3D face model. In general, the full head model is used to show human head contours and large-scale geometric features. This model has relatively few vertices, but the surface is smooth and can be rendered rapidly. The 3D full head Morphable Model [12] can be defined by the PCA model, and the identity parameters were extracted from the statistical model. The shape of the PCA model and the texture of the PCA model is respectively expressed as:

$$S = \overline{S} + \sum_{i=1}^{K} \alpha_i P_i \tag{1}$$

$$T = \overline{T} + \sum_{i=1}^{K} \beta_i T_i \tag{2}$$

where $\overline{S}$ and $\overline{T}$ represent the mean shapes and the mean textures of the face samples. P_i and T_i are the shape and texture eigenbasis, which represents the principal component of the face shape and texture models respectively. α_i and β_i are the shape and texture parameters.

The PCA model above can be used to reconstruct the face shape and the texture, but it disregards the internal relationship between the identity and the expression. We arrange face samples according to different axial features to form those data samples into third-order tensor. The first dimension is the coordinate vertex of the face, the second dimension represents the identity feature and the third dimension represents the

facial expression feature. The tensor of face data reflects the internal association of the shape variation. The PCA model of the face texture will be used for the UV texture generation in section V.

A. FACE DATABASE AND LANDMARK PROCESSING

To establish the 3D full head basic model, sufficient face samples are needed for statistical analysis. We integrate FaceWareHouse [15], BU-3DFE, and 200 extra female models into our face database, which consists of 650 different identities and 47 different expressions. Registration in the database has been finished already. We use the average sample of the statistical model to recover the craniosacral region and ear components, make those places are roughly matched to the human head contour.

The 3D face landmark points are supposed to be handled before multilinear optimization. The normal vector of all candidate points is calculated according to the current face posture in space[32]. In each candidate point row, the vertex with the maximum angle between the normal vector and the Z-axis (i.e., perpendicular to the face plane) is defined as the alternative contour points.

B. MULTILINEAR OPTIMIZATION

We use the multilinear model based-PCA to compress the dimension of the face sample since it presents the linear combination of facial geometry. The feature dimension can be extracted based on identity and expression. The face sample S^* can be represented by a multilinear combination of the third-order tensors:

$$S^* = \mathbf{Cr} \times_{\mathrm{id}} \omega_{\mathrm{id}} \times_{\mathrm{exp}} \omega_{\mathrm{exp}} \tag{3}$$

where Cr is the third-order tensor kernel of the corresponding dimension of the orthogonal matrix, $\times_{\mathrm{id}} \times_{\mathrm{exp}}$ respectively represent the modular multiplication on the dimension of identity and expression, ω_{id} is the identity feature vector and ω_{exp} is the expression feature vector of the face. The basic model can be expressed as:

$$F = \boldsymbol{R} \cdot (\mathbf{Cr} \times_{\mathrm{id}} \omega_{\mathrm{id}} \times_{\mathrm{exp}} \omega_{\mathrm{exp}}) + \boldsymbol{T} \tag{4}$$

where $\boldsymbol{R}$ and $\boldsymbol{T}$ are respectively represent the rotation matrix and translation vector of the current face.

C. BASIC MODEL

After 2D coordinate projection, we can represent the basic model by:

$$E_{\mathrm{data}} = \sum_{i=1}^{L} \left\| Q(f) \cdot (\boldsymbol{R} \cdot (\mathbf{Cr} \times_{\mathrm{id}} \omega_{\mathrm{id}} \times_{\mathrm{exp}} \omega_{\mathrm{exp}}) + \boldsymbol{T})^{(i)} - q_i \right\|^2 \tag{5}$$

where $Q(f)$ indicates 3D coordinates projection set, $\{q_i\}$ is 2D coordinates of facial feature points in the image, and L is the number of the points. We indicate (5) by the optimization values of parameters:

$$\boldsymbol{R}^*, \boldsymbol{T}^*, f^*, \omega_{\mathrm{id}}^*, \omega_{\mathrm{exp}}^* = \underset{R,T,f,\omega_{\mathrm{id}},\omega_{\mathrm{exp}}}{\arg\min} \; E_{data} \tag{6}$$

FIGURE 2. The illustration of basic model after multilinear optimization.

equation (4) is to minimize the objective function and use the coordinate descent method to figure out each variable.

The basic model after multilinear processing still has some bias, which may be related to misalignment between 2D landmark points and 3D landmark points. Therefore, we utilize the Laplace deformation to correct this deviation:

$$v_1', v_2', \ldots v_M' = \underset{v_1', v_2', \ldots v_M'}{\arg\min} \left\{ \sum_{i=1}^{M} \|Hv_i\|^2 + w_1 \cdot \sum_{v} \|Q(f) \cdot (\boldsymbol{R} \cdot v_i + \boldsymbol{T}) - q_i\|^2 \right\} \tag{7}$$

where $v_1', v_2', \ldots v_M'$ are constrained vertices of the 3D face model, H is the Laplace coefficient matrix obtained from coordinate vertices of the current model, and v_i is 3D model feature points. w_1 is the weight of the restrain term and we select $w_1 = 0.8$ after tested. The full head basic models after visualized exhibit in Fig. 2, where the first row is the input face images, and the corresponding basic model displayed in the second row.

IV. FULL FACE-AND-HEAD MODEL

The full head basic model obtained in section III lacks facial details, and can only be used to represent the large-scale geometric contour. But, the facial region has features and surface details, and the 3D face-and-head model requires precise facial details to verify the identity and the expression.

The facial details in the geometric model are shown as tiny-scale surface deformation, which can hardly be represented by the statistical model. We use the face reconstruction algorithm based on (Shape From Shading) SFS to generate high-resolution point cloud model [33]. We use the facial region of the full head basic model yielded in Section 2 as a template for point cloud matching since the high-resolution point cloud model lacks standard grid topology,. After that, the highly-detailed face model is obtained as the face representation of our full face-and-head model.

A. ILLUMINATION OPTIMIZATION

For the SFS algorithm needs the illumination changes of the target object to recover the 3D geometric shape, which satisfies the input requirements of our single ''in-the-wild'' image. The principle of SFS is to reversely deduce the rendering equation from image information, meanwhile recover surface normal vector and depth information of the image. We build

the high-resolution point cloud model by SFS and propose an optimized algorithm for face detail restoration.

The primary light reflection of "in-the-wild" images are the diffuse reflection after absorbed by the skin and the specular reflection generated by the skin oily components [5]. In SFS rendering equation, some basic conditions are assumed [34] as:

– Face surface does not emit itself, the reflection in skin is assumed to be Lambert reflection and specular reflection, and the albedo is known;

– The light spot is unique, and its luminance and position are known;

– The image condition is the orthogonal projection.

The Lambert-light model [34] can be represented by:

$$I(x, y) = l \cdot \rho \cdot (\mathbf{u} \cdot \mathbf{n}) \tag{8}$$

where $I(x, y)$ is the pixel value of the image, $l$is the luminance of the light spot, and ρ is the albedo of surface. $\mathbf{u} = (u_x, u_y, u_z)$ is the light incident direction, and $\mathbf{n} = (n_x, n_y, n_z)$ is the normal vector to the surface. According to the spherical harmonic function, the Lambert model pixel value in the image can be obtained by:

$$I_1(x, y) = \rho(x, y) \cdot \vec{L}_1 \cdot \vec{Y}(\mathbf{n}(x, y)) \tag{9}$$

where $\vec{L}_1$ is the Spherical harmonic function of incident direction and the light spot luminance, and we represent the vector product of the surface normal as $\vec{Y}(\mathbf{n}(x, y))$. Similarly, the specular model pixel value can be represented by:

$$I_2(x, y) = k \cdot l \cdot ((2\mathbf{n} \cdot L) \cdot \mathbf{n} - L)^{\mathrm{s}} \tag{10}$$

where k is the specular reflection coefficient, L is the unit vector of incident light which proportioned to the luminance and distance of the light spot, and s is the specular index related to the intensity of the highlight area of the face [5]. To simplify calculating, we use the energy function:

$$E_l = \sum\nolimits_{(x,y)\in face} \left\| I(x, y) - \rho(x, y) \cdot \vec{L} \cdot \vec{Y} \right\|^2 \tag{11}$$

$$E_s = \sum\nolimits_{(x,y)\in face} \left\| I(x, y) - k(x, y) \cdot l \cdot ((2\vec{Y} \cdot L) \cdot \vec{Y} - L)^{\mathrm{s}} \right\|^2 \tag{12}$$

where E_l represents the energy function of the Lambert reflection model, and E_{s} is the Specular reflection energy function. The light rendering model R is acquired by:

$$R = (1 - w)E_l + wE_{\mathrm{s}} \tag{13}$$

where w is the luminance weight and $w \in [0, 1]$. Additionally, the input image pixel values are consistent in this circumstance, and the SFS rendering model is obtained by:

$$R = \arg\min \sum_{(x,y)} \left\{ (1 - w) \left\| I(x, y) - \rho(x, y) \cdot \vec{L} \cdot \vec{Y} \right\|^2 + w \left\| I(x, y) - k(x, y) \cdot l \cdot ((2\vec{Y} \cdot L) \cdot \vec{Y} - L)^{\mathrm{s}} \right\|^2 \right\} \tag{14}$$

The depth corresponding to the pixel coordinates can be computed by partial differential equations after obtaining the normal vector corresponding to each pixel in the face region. We have optimized the lighting rendering model in the process of high-resolution facial model generation. It has a great effect on the display of facial details (wrinkles, expressions, *et al.*), which is better than the current advanced algorithms in runtime.

FIGURE 3. Some illustrations of the 3D face mesh model after mesh subdivision.

B. HIGHLY-DETAILED FACE MODEL

To make the point cloud model match better, the orthogonal projection is used to stretch the parameterized face plane extracted from the basic model, and the basic model is converted into the high-resolution model by mesh subdivision algorithm. The Butterfly Subdivision algorithm is employed to increase the vertices on the basic model while the original grid vertices were kept unchanged. The added vertices were located in the triangle of the plane grid, it can be found from the original model that contained the points of the triangular grid. The mesh model after subdividing is shown in Fig. 3, and we utilize various characters and expressions to demonstrate the effect of mesh subdivision.

Besides, the high-resolution face model after mesh subdivided needs to be padding with the point cloud structure to show the facial details. The spatial point cloud needs to be matched which is obtained by mesh subdivision and the Iteration Closest Point (ICP) algorithm, the process can be expressed as:

$$D_{p-\mathrm{m}} = \sum_i \left\| v_i - v_p \right\|^2 \tag{15}$$

$$E_{\mathrm{laplacian}} = \sum_i \left\| Hv_i \right\|^2 \tag{16}$$

where $D_{p-\mathrm{m}}$ represents the minimum distance from the vertex in the template to the vertex matched to point cloud, $v_i$symbolize the vertices in the face model, and v_p is the vertices in the point cloud. Furthermore, $E_{\mathrm{laplacian}}$ is the Laplace regular term of the mesh without deformation. The function means the vertex will be iterated and converged until matched. The specific ICP step is to find the nearest point in the cloud and reverse the corresponding point in the model. We can see the cloud point padding effect in Fig. 4, the face mesh model after subdivided accurately matches facial details and local deformation details.

C. MODEL INTEGRATION

Once we have obtained the full head basic model and the highly-detailed model in the former section, the assignment of this step is to integrate the highly-detailed face model with the full head basic model.

FIGURE 4. An illustration of the cloud point padding effect. Figure (a) shows the high-resolution model, Figure (b) is the overlapped cloud point, Figure (c) shows the highly detailed model after padding, and Figure (d) is the highly-detailed model after texture mapping.

FIGURE 5. The process of the high-detailed face model and the basic full head model integrating. The first figure is the full head basic model after displaced the face region, and the second figure is the highly-detailed face model. The third figure shows the integrating process, and the last figure is the integrated model.

Our method is supposed to remove the face region in the full head basic model and replace this area with the highly-detailed face model. The discarded part (see in Fig. 5) is adjusted according to the scale of the detailed model, it can be referred to the extracted area in Fig. 4. To ensure the consistency of the vertices between those two models, a transitional mesh gap is appended between the full head basic model and the highly-detailed model. The face boundary is extracted and mapped the two boundary vertices to the plane. This gap is generated by the Delaunay method to integrate two models with triangulation, and the vertices in the gap are processed by Laplace smoothing algorithm. The processing diagram is shown in Fig. 5, we can see that the integrated model is smooth and the face details finely maintained.

V. TEXTURE GENERATION

After integrating the full face-and-head model, mapping the panoramic texture to the mesh model will increase the realism of the model. However, it is difficult to synthesize the panoramic UV texture from a single face image since the essential texture information such as profile and the craniofacial region were lost due to the perspective occlusion. To solve this, we design an effective framework to generate the panoramic UV texture with the optimized generative adversarial network. Moreover, we have also collected sufficient UV texture data, some textures are selected from access databases, and some others are scanned by our 3D professional scanner. In this section, we mainly introduce our texture generation network structure and the experimental configuration.

A. UV TEXTURE GENERATION FRAMEWORK

Some 3D face alignment methods establish the connection between 3DMM and UV map, the facial UV map can be generated by sampling 2D image with a fitted 3D face model. Besides, methods about using framework based-GAN to extract the facial UV map [29] from images are proposed recently. Spired by these approaches, we design a modified-GAN framework to generate corresponding UV panoramic texture maps from ''in-the-wild'' images. This network structure includes a VAE module, a generator module, and two discriminators with different purposes.

In the GAN model, the data can be straightforward sampled without a present distribution, therefore, the real data can be approximate from the training process. Nevertheless, the training process of the GAN model is uncontrollable, the samples generated from the GAN are often different from the real image, especially for the UV texture images with complex pixels and high feature entanglement. The feature of the VAE is to add constraints to the encoder, and the encoder is subject to the latent variable of the unit Gaussian distribution [31]. On the other hand, the output image generated by the VAE is blurred because of the input noise and coarse loss function MSE (Mean Square Error). We also use the unsupervised adversarial training of the GAN network to improve the constrained effect of the VAE module and create the instances more realistic.

Our texture generation module mainly consists of four parts: the encoder module is used to map the input image to the latent vector; the generator network module generates the face texture image according to the latent vector; the discriminator D is used to identify the authenticity of the generated UV texture; and the classifier network module is used to detect the consistency of the face landmark in the local with the input texture parameters. Fig. 6 illustrates our network architecture.

1) GENERATION MODULE

We project the input face image to a latent vector by the encoder module, and the generator is used to reconstruct the original pixels which match the characteristics of the input image with the latent vector. The relationship between the latent space and the input image can be established by the Encoder and the GAN module.

We add texture parameters of the texture PCA model in section III as the conditional attributes before the encoder which can promote the accuracy of the discriminator, as well as making the GAN network more controllable. The optimizing process of the generator module can be formulated as minimizes the per-pixel Euclidean distance between the synthetic texture and the sample texture corresponding to the latent parameters. The new instances can be constructed by pixel-wise $l2$ norm as the generation loss:

$$L_G = \frac{1}{2}(\|I^*(x,y) - I(x,y)\|_2^2 + \|f_D(I^*) - f_D(I)\|_2^2 + \|f_C(I^*) - f_C(I)\|_2^2) \quad (17)$$

where $I(x,y)$ and $I^*(x,y)$ respectively symbolize the pixel of the generic texture and the pixel of the synthetic texture. f_C and f_D are the features of an intermediate layer of global discriminator network and local face discriminator network.

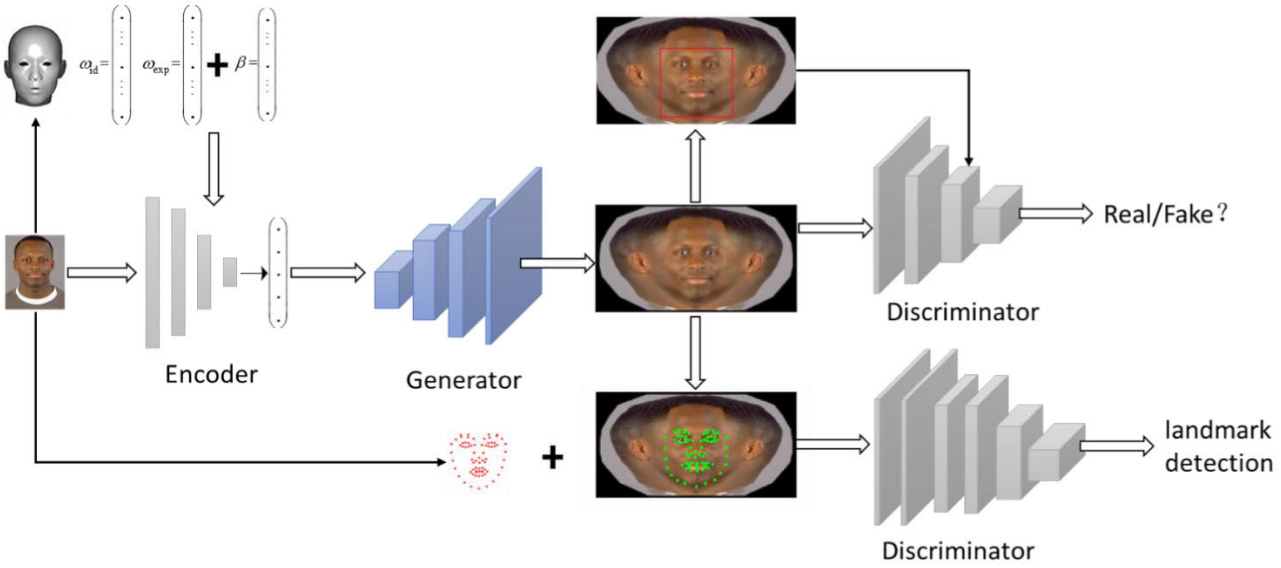

FIGURE 6. The pipeline of our texture generation network architecture. It consists of one Encoder, one Generator, and two Discriminators. The Encoder takes the "in-the-wild" face image and some texture parameters extracted from (2) and (6) in section III as input. The Generator outputs a UV texture based on the mixed parameter vector. Two discriminators have different characteristics and are respectively used to distinguish feedback.

2) DISCRIMINATION MODULE

In the pre-trained GAN network, the discriminators with different features are learned to validate the genuineness of the generated texture [29]. The discriminator module consists of a global discriminator and a face landmark classifier, where the global discriminator is used to estimate the authenticity of the UV texture images, and the landmark classifier is used to verify the identity of the main face region. The criterion of the global discriminator evaluation process is that the UV texture generated according to the texture parameters is true and consistent with the UV map standard. Meanwhile, the combined area of the central face and the edge is consistent and diverse. To achieve this, the correlated distribution of the discriminator can be formulated as:

$$LD = -\mathbb{E}_{x\sim Pd(x),y\sim Pd(y)}[\log D(x,y)] - \mathbb{E}_{c\sim P(c),y\sim Pd(y)}[1-\log D(G(c,y),y)] \quad (18)$$

where $P_{d(x)}P_{d(y)}P_{(c)}$ are the distribution of global UV texture, local UV texture, and the PCA texture parameters correspondingly.

3) LANDMARK DETECTION CLASSIFICATION

We employ the face recognition network [35] to obtain the identity-related features of the face region in the generated UV texture map. The network is pre-trained and used to detect whether the central face area in the generated UV texture is consistent with the ground truth. The cosine distance between the landmark of the input face image and the generated UV texture face region is calculated as the identity loss function:

$$L_{id} = -\sum_{i=1}^{m}\log\frac{e^{\|x_i\|+y_i}}{\sum_{i=1}^{n}e^{\|x_i\|+y_i}} + \frac{\lambda}{2}\sum_{i=1}^{m}\|x_i - c_{yi}\|_2^2 \quad (19)$$

where m is the batch size, n is the number of the training samples. $x_i \in \mathbb{R}^{512}$ and $y_i \in \mathbb{R}^{512}$ indicates the feature vector of the input face image and the feature vector generated by the network model. c_{yi} is the y_i-th class feature center and the λ is the weight of the center loss. We take $\lambda = 0.1$ in our experiment since it is beneficial to leverage this loss to maintain identity in the synthetic texture [29]. The face landmark detect score of a pair of testing image is calculated according to cosine similarity between the two feature vectors can be referred to [13].

For the principle of discriminator and classifier network structures, please refer to [31] and [29]. However, CVAEGAN recovers the image in a specific category with optional drawn values on a latent attribute vector by transforming the fine-grained category label into the resulting generative model [31]. UVGAN is to fill the incomplete textures with random noise and distinguish real and fake. In the experiment, we preserve the face landmark consistent with the "in-the-wild" face image. In other parts (i.e., ears, profile, part of the hair, and the neck region), the network generates the novel content based on the original attributes.

4) LOSS FUNCTION OF THE NETWORK

The final loss function of the UV texture GAN model is a sum of the aforementioned losses:

$$L = \lambda_G L_G + \lambda_D L_D + \lambda_{id} L_{id} \quad (20)$$

λ_G, λ_D, λ_{id} are the weights to balance the final loss function leverage the different modules. We empirically fix the weights as $\lambda_G = 1$, $\lambda_D = 5 \times 10^{-2}$, $\lambda_{id} = 10^{-3}$.

B. CONFIGURATION AND DATASETS

We collect the UV textures in useful datasets, including the WildUV dataset which contains nearly 2000 texture images of individuals with different identities and 5638 unique facial UV texture maps. Besides the WildUV dataset, the other part of texture data is obtained by our laboratory and partner companies using professional 3D scanning equipment.

About 400 testers with different identities (250 males, 150 females) which provides 2000 various UV texture maps. Moreover, we deploy the data augmentation on perfect texture images in the database. In the end, there are a total of 10143 texture samples in the experiment.

In the network structure, the encoder and generator module follow [36] and employs the fully convolutional cascaded form structure, and skip connections between corresponding layers in the encoder and generator module are made. Subsampling is used in Conv and Pooling layers for the input images and the latent vector dimension is fixed to 256. We cascade two contiguous layers at the end of the encoder for merging the texture parameters with the sub-sampled input image. The generator and the encoder are linked by two fully connected layers, followed by 6 up-sampled deconvolution layers and corresponding filters where the convolution layer includes 256,128 and 64 channels [37]. We resize the input of the Alexnet framework [38] to 256*256 as our discriminator module and utilize the batch normalization layer after the convolution layers.

For the processing of the UV texture, our initial input is ''in-the-wild'' face image in optional size. In this experiment, the test image size we select from the CelebA-HQ is 512*512. We resize the image to 256*256 as the input since the standard size of the UV texture samples is 597*377. The face region is cropped in 128*128 which is centered with the nose tip. The network is deployed with Tensorflow. We train the network for 200 epochs, batch size is 16, and the learning rate is 0.001.

C. MAPPING OPTIMIZATION

The full face-and-head model require rendering once we have obtained the panoramic UV texture. Our method is to calculate the texture coordinates of each vertex in the 3D model and map the texture to the vertex based on UV mapping.

The texture information is stored in coordinate form in UV texture plane, and each coordinate T is corresponding to the vertex in 3D space. We can define a mapping function through the correspondence among the coordinates of the UV texture plane and the vertices of the 3D model, and the UV texture pixels can be matched with the 3D full head model by the mapping function [39]. The mapping function can be formulated as:

$$\begin{aligned} T &= \wp(I_{\mathrm{uv}}, t_{\mathrm{c}}) \\ \wp &: I_{\mathrm{uv}} \rightarrow t_{\mathrm{c}} \end{aligned} \tag{21}$$

I_{uv} is the texture image, t_c is the texture coordinate in the UV image plane, $\wp : I_{uv} \rightarrow t_c$ represent the project correlation between the corresponding coordinates between the UV space and the texture image.

Since the profile and back region in our panoramic UV texture is generated by inference, some deviations may appear between the textures to the mesh model. We make some appropriate adjustments before UV mapping to avoid this deviation. Expanding the mesh model under the panoramic texture, and align the expanded mesh according to some landmark point which is labeled by us. The texture coordinates corresponding to the vertices in the 3D mesh model are acquired by projection. It can decrease the project offset between the UV texture and the mesh model, and the projection effect will be better.

FIGURE 7. The illustration of the UV texture mapping. The first figure is the plane projection of our mesh model, the second figure is the panoramic UV texture generated by our network, the last figure shows the process of the face organic components padding.

To make the high-fidelity model, the artificial organic components are used in the 3D model. Since the ear and the teeth part are unimportant biological feature regions, we conduct the ear model as a generic model by statistical optimization from the full head basic model in section 3.

The component of the facial organs (including eyes, teeth, *et al.*) have been accomplished manually. The texture of the iris is received by reshaping the largest ellipse inside the projection of the eye region to the most frontal input face image [17]. The main process of the grid expansion is shown in Fig. 7.

VI. EXPERIMENTS AND EVALUATIONS

A. EXPERIMENTS

Our full face-and-head model is a multi-aspects synthesis, including full-head model reconstruction, highly-detailed face model reconstruction, and panoramic texture map generation. We first show the generated models in each process, subsequently, evaluate the final results with various methods. Under the hardware configuration of this experiment (memory 16G, CPU 4.0 Hz, GPU NVIDIA 2080TI), the average time to recover a full face-and-head model from a single ''in-the-wild'' face image is about 12500 ms. We show the output of each step and compare it with the state-of-the-art method in Fig. 8 and Fig. 9.

B. EVALUATIONS

In this section, we follow the brief description of the pipeline (refer to Fig. 1) to conduct our experiments. The results of each step are visualized as illustration, we can check the output from the illustration respectively.

We mainly estimate the results of panoramic texture generation, and then mapped the texture to the synthesized full face-and-head model for visual comparison. In the experiment, we use the single frontal face image in the CelebA dataset as the input, and the result generated from each segment of the experiment is derived from the same frontal face image. To prove the efficiency of our texture generation network architecture, we first compare the results of UV texture generation in the following 1) section. Apart from this, we use the standard metrics to quantitatively evaluate the texture results from the various methods in 2) section.

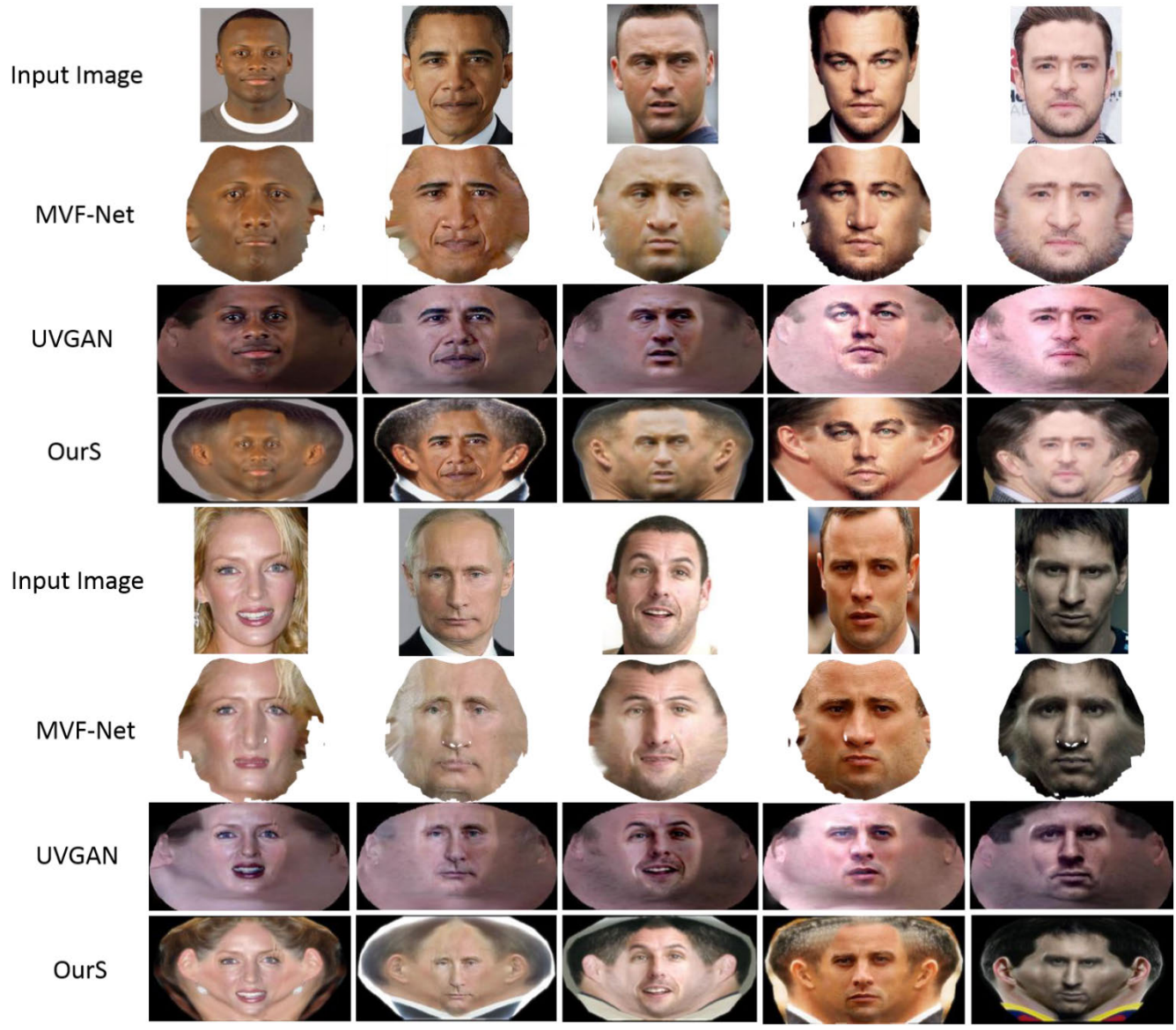

FIGURE 8. Comparison of the complete textures between our framework and other texture generation methods.

For the effect evaluation of the complete model with texture, we visualize the final results under multi-perspectives from other methods and applications in the following 3) section.

1) UV TEXTURE EVALUATION

We compare the UV texture generated in section V with the results produced by UVGAN [29] and MVF-Net [24]. Among them, the object of UVGAN is similar to ours to generate panoramic UV textures. In [24], output is the frontal texture generated by the state-of-the-art 3DMM regression method. The MVF-Net texture is just appropriate for the facial region generated by the 3DMM method and unable generalize to the full face-and-head model. We use it as a comparison sample of the facial region recover effect. For the subjective evaluation of our results with other state-of-the-art methods, we can refer to Fig. 8. As shown, the first row is the input image, we take some ''in-the-wild'' frontal face image as input to compare the texture maps obtained by different methods. The second row below the input images is the texture acquired from the MVF-Net, those texture maps fit well in albedo and facial feature details. However, the texture map obtained by the MAF-Net only has the information of the visible area in the input image and is no longer applies to the full head model mapping. In addition, some missing spots and deformation occur in the shadow and self-occlusion regions. Output of the UVGAN are shown in the third row, the panoramic UV texture can be used in full face-and-head model mapping and the details in the missing part is inferred to be more precise. Nonetheless, this method has defects in the surface albedo, which caused the skin color of the obtained UV texture to be inconsistent with the original input. This might be caused by the lack of face texture parameters

FIGURE 9. Results of the visual comparison with other methods and the commercial application. The first column is the input images, and the other columns are the results of the corresponding method in multi-view.

in the UVGAN prediction process. Our results are shown in the fourth row, the panoramic UV texture generated by our network maintains the identity and expression, as well as precisely predicts the texture information of the invisible region. Compared to UVGAN, our method create a high-fidelity model, and recovers a wider region (e.g., the hair and the neck part). The overall effect is more realistic, surface details and the skin in particular.

2) QUANTITATIVE EVALUATION

To quantitatively evaluate the texture results from various methods, we employ the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics to evaluate the quality of the textures. The PSNR measures the difference between the pixel values, and the SSIM estimates the similarity between the two results. We compute the metrics between the generated textures and the ground truth to compare the reliability of the invisible region which is represented by PSNR (UV) and SSIM (UV) in Tab.1. Meanwhile, we measure the metrics of the facial center area among the outputs and the original images which is indicated by PSNR (center) and SSIM (center) in Tab. 1. Note that the top two rows of the 3DMM column are empty as the texture generated by 3DMM is incomplete. In Tab.1, we show the two metrics between each methods, it can be seen that the texture generated by our method is better than other methods in both metrics.

3) EFFECT EVALUATION

To demonstrate the effect of our full face-and-head model is better than the results of other face modeling methods, we visually compare the complete model from

TABLE 1. Average PSNR and SSIM comparison

Algorithm	MVF-Net	UVGAN	Ours
PSNR(UV)	-	28.5	29.4
SSIM(UV)	-	0.912	0.930
PSNR(center)	29.3	28.7	29.5
SSIM(center)	0.923	0.919	0.952

multi-perspective. In Fig. 9, we can see that both GANFIT, Tran [4] methods, and the commercial application FaceGen recovers the frontal face model well. Those models generated by the GANFIT and Tran are incomplete models, which just has a facial part. Under the perspective in profile, some unexpected irregular spots emerged on the nose corner from the GANFIT model. The reason may be information loss which caused by the self-occlusion and the light shadow of the original image. The result built by Tran method also performs well in the frontal facial area, but some pixel blur and texture stretching occurs in the profile. In addition to the component of the facial organs (eyes, teeth, and tongue, *et al.*), those two comparative methods adopt UV texture as the covering layer rather than the artificial components (GANFIT discards the mouth components) which may lack stereoscopy. Furthermore, we join a comparison with the current commercial solution FaceGen. The result built by FaceGen is a full face-and-head model that is relatively complete in structure. However, FaceGen just recovers the texture of the face center region, it lacks the essential surface information, as well as in the cheek and forehead parts where are visible in the input image. Moreover, those models from FaceGen are identical in structure and have the same face shape and size in Fig. 9, because of the modeling process is based on a general model. In contrast, our models maintain individual structure features well, and the panoramic texture perfectly fulfills the entire surface of the model even in the profile, back, and the top of the head region. In general, our result is more fine and photorealistic than other comparative methods.

VII. CONCLUSION

In this paper, we introduce a method that generates the full face-and-head model with photorealistic texture from a single ''in-the-wild'' face image. The pipeline incorporates multiple effective processes to extract input image information for 3D model reconstruction, and the invisible region texture information is recovered by our data-driven VAEGAN network with a large-scale UV texture dataset. From the evaluation, we demonstrate that our full face-and-head model performs better than other ''state-of-the-art'' methods under the equivalent circumstances. Our model exhibits an unprecedented level of detail and realism in the experiment, it will provide some artistic inspirations to this field.

LIMITATIONS AND FUTURE WORK

In the texture generation module, our network preserves the texture of the facial area well. But for some hair and neck regions, the network is sensitive to the pixel variations (e.g., illumination and gray hair, *et al.*), this may result in the predicted texture in those regions are quite different from the original image. We will try to append some specific noise in the training process to address this matter in the future work. On the other hand, our full face-and-head model does not involve the reconstruction of the hair. We utilize the texture to restore the visual effect of some male hairstyles like short hair or bald condition, but our approach performs weird when the hairstyle is complex. For the different physical characteristics of the hair and the head, the topology structure is also inconsistent, and those two parts are modeled separately in the current method. Our future work will concern on the hair modeling, combining the hair part with our full face-and-head model. One possible solution could be mesh modeling based on the hairstyle data-driven and mapping with the hair texture. We will verify the feasibility of this inspiration and hope to use it for the 3D avatar creation.

REFERENCES

[1] L. Hu, ''Avatar digitization from a single image for real-time rendering,'' *ACM Trans. Graph.*, vol. 36, no. 6, p. 195, 2017.

[2] 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. 1576–1585.

[3] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Perez, M. Zollhofer, and C. Theobalt, ''FML: Face model learning from videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10804–10814.

[4] 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)*, Jun. 2019, pp. 1126–1135.

[5] 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, Nov. 2017.

[6] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, ''A 3D morphable model learnt from 10,000 faces,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5543–5552.

[7] J. R. Tena, F. De la Torre, and I. Matthews, ''Interactive region-based linear 3D face models,'' *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011.

[8] A. E. Ichim, S. Bouaziz, and M. Pauly, ''Dynamic 3D avatar creation from hand-held video input,'' *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–14, Jul. 2015.

[9] T. Bagautdinov, C. Wu, J. Saragih, P. Fua, and Y. Sheikh, ''Modeling facial geometry using compositional VAEs,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3877–3886.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, ''Deep learning face attributes in the wild,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[11] V. Blanz and T. Vetter, ''Face recognition based on fitting a 3D morphable model,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[12] V. Blanz and T. Vetter, ''A morphable model for the synthesis of 3D faces,'' in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 187–194.

[13] 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.*, Sep. 2009, pp. 296–301.

[14] D. S. Ma, J. Correll, and B. Wittenbrink, ''The chicago face database: A free stimulus set of faces and norming data,'' *Behav. Res. Methods*, vol. 47, no. 4, pp. 1122–1135, Dec. 2015.

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

[16] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, ''Large scale 3D morphable models,'' *Int. J. Comput. Vis.*, vol. 126, pp. 233–254, Apr. 2018.

[17] S. Ploumpis, E. Ververas, E. O' Sullivan, S. Moschoglou, H. Wang, N. Pears, W. Smith, B. Gecer, and S. P. Zafeiriou, "Towards a complete 3D morphable model of the human head," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 29, 2020, doi: 10.1109/TPAMI.2020.2991150.

[18] 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.

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

[20] S. Ploumpis, H. Wang, N. Pears, W. A. P. Smith, and S. Zafeiriou, "Combining 3D morphable models: A large scale Face-And-Head model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10926–10935.

[21] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 725–741.

[22] V. F. Abrevaya, S. Wuhrer, and E. Boyer, "Multilinear autoencoder for 3D face model learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1–9.

[23] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "MeshGAN: Non-linear 3D morphable models of faces," in *Proc. CVPR*, 2019. [Online]. Available: https://arxiv.org/abs/1903.10384

[24] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K. N. Ngan, and W. Liu, "MVF-net: multi-view 3D face morphable model regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 959–968.

[25] 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. Eur. Conf. Comput. Vis.*, 2018, pp. 534–551.

[26] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, "Real-time facial animation with image-based dynamic avatars," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016.

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

[28] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li, "Photorealistic facial texture inference using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2326–2335.

[29] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7093–7102.

[30] 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)*, Jun. 2019, pp. 1155–1164.

[31] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: fine-grained image generation through asymmetric training," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2764–2773.

[32] W. Gao, X. Zhao, Z. Gao, J. Zou, P. Dou, and I. A. Kakadiaris, "3D face reconstruction from volumes of videos using a mapreduce framework," *IEEE Access*, vol. 7, pp. 165559–165570, 2019.

[33] P. Huber, G. Hu, R. Tena, P. Mortazavian, and J. Kittler, "A multiresolution 3D morphable face model and fitting framework," in *Proc. VISAPP*, 2016, pp. 79–86.

[34] C. Li, K. Zhou, and S. Lin, "Intrinsic face image decomposition with human face priors," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 218–233.

[35] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351. Computer Science Department and BIOSS Center for Biological Signalling Studies, 2015, pp. 234–241.

[37] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," presented at the 33rd Int. Conf. Mach. Learn., Mach. Learn. Res., 2016. [Online]. Available: http://proceedings.mlr.pres

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[39] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3D morphable model regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8377–8386.

YANGYU FAN received the M.S. degree in electromechanical engineering from the Shaanxi University of Science and Technology, Xi'an, China, in 1992, and the Ph.D. degree in acoustics signal processing from Northwestern Polytechnical University, Xi'an, in 1999. He is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. He has authored or coauthored numerous papers that appeared in various publications, including *Neurocomputing*, *Signal Processing*, *Image and Vision Computing*, IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, *Multimedia Tools and Applications*, and so on. His research interests include image processing, pattern recognition, and virtual reality.

YANG LIU received the M.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information. His research interests include image processing, virtual reality, computer vision, machine learning, and their applications to 3D face reconstruction.

GUOYUN LV received the B.S. degree in applied physics from the National University of Defense Technology, and the Ph.D. degree in computer application from Northwestern Polytechnical University, in 2008. From 1996 to 2006, he was a Research Scientist in different research institutes and companies for signal processing, pattern recognition, sound and photoelectric warning system, communications and electronic information systems, embedded systems, and so on. Since 2008, he has been an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University. His research interest includes signal and information processing, audio and video image processing, artificial intelligence, target detection and identification, 3D virtual reality and augmented reality, as well as multiple heterogeneous information fusion.

SHIYA LIU is currently the Director of the Virtual Reality Content Production Center, the Director of the Qingdao Star Shark Virtual Reality Technology Research Institute, and a Researcher of the United Nations Institute of Digital Economy. He focuses on the integration, innovation, and application of high and new technologies in the field of electronic information and communication, such as VR/AR, 4K/8K, AI, 5G, and microelectronics. He is a member of the Information and Communication Economy Expert Committee of the Ministry of Industry and Information Technology.

GEN LI received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2018, where he is currently pursuing the M.S. degree with the School of Electronics and Information. His research interests include virtual reality, image processing, computer graphic, and their applications to hair reconstruction.

YANHUI HUANG received the B.S. degree in electronics and information engineering, the M.S. degree in electronics and information engineering, and the Ph.D. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2008, 2011, and 2018, respectively. He is currently an Assistant Researcher with miHoYo company, Shanghai, China. His research interests include computer graphics, virtual reality, computer vision, machine learning, and their applications to 3D avatar creation.