3DCMM: 3D Comprehensive Morphable Models with UV-UNet for Accurate Head Creation

Jie Zhang, Kangneng Zhou, Yan Luximon, Tong-Yee Lee, *Senior Member, IEEE*, and Ping Li, *Member, IEEE*

***Abstract*—In recent studies of 3D shape modelling and reconstruction, the focus has primarily been on the 3D face region. However, accurately creating the entire 3D head opens up a wide range of applications, including headwear design, cranial diagnosis, and avatar design. Therefore, we present our newly developed method of constructing 3D comprehensive morphable models (3DCMM) specifically tailored for human heads, along with a novel 3DCMM-based stepwise pipeline for creating accurate full 3D heads. Within our 3DCMM framework, we constructed a powerful 3D morphable face model with UV-UNet to generate the 3D face and predict the 3D scalp, resulting in a complete representation of the head. Additionally, our 3DCMM-based self-learning approach incorporates novel facial boundary-aware and structure-aware losses for highly accurate overall reconstructions of the entire facial region. Experimental evaluations demonstrate that our 3DCMM exhibits superior face representation power and achieves higher head prediction accuracy than existing models. Consequently, our 3DCMM-based 3D head creation method from a single image demonstrates outstanding performance capability on both face and head benchmarks. Our project is publicly available at: https://github.com/Easy-Shu/HeadUV-UNet.**

***Index Terms*—3D morphable model, 3D face reconstruction, 3D scalp completion, 3D head creation.**

I. Introduction

3D face creation is a significant research area with the objective of recovering accurate 3D facial geometry from unconstrained 2D images. Previous approaches have predominantly utilized learning-based methods to estimate shape coefficients of 3D statistical morphable face models [1]. However, the task of full 3D head creation poses greater challenges and complexities, yet it offers numerous novel applications and opportunities to overcome the constraints associated with existing 3D face creation and reconstruction methods based solely on face images. In the field of computer graphics, a comprehensive understanding of the entire head holds significant value for designers, enabling them to effectively model hairstyles and generate high-fidelity avatars [2]. Furthermore, in ergonomics design, a detailed 3D representation of the full head enables the evaluation and customization of headwear products like helmets and headphones, ensuring optimal fit and comfort for individuals [3], [4]. Additionally, in cranial diagnosis, a comprehensive 3D representation of the full head can facilitate the detection of craniofacial deformities and changes [5]. In contrast to previous studies that predominantly concentrated on closely cropped inner facial regions [1], our research addressed the challenge of creating a full 3D head from 2D face images, encompassing both the facial and scalp regions.

Although many powerful 3D statistical morphable models (3DMMs) of human faces [7], [8], [9] or heads (e.g., low-resolution FLAME [10] and LYHM (mostly Caucasian)) have been constructed for 3D face/head generation and reconstruction, limited studies exist on predicting the full 3D head from the generated 3D facial regions [11]. Furthermore, compared with 3D face reconstruction from a single 2D image [1], [12], [13], [14], [15], [16], [17], full 3D head reconstruction methods are much less common. State-of-the-art self-supervised-learning-based methods for 3D head reconstruction utilize deep convolutional neural networks (CNNs) to regress the head shape or texture parameters of 3DMMs in an analysis-by-synthesis scheme, which offers the advantage of not requiring ground-truth 3D data for training. Examples of such methods include RingNet [14] and DECA [18]. However, a single-face image can only provide facial information and does not capture scalp information due to the hair-occlusion problem, resulting in inaccurate reconstruction of the scalp region.

To address the limitations mentioned above, our research developed 3D comprehensive morphable models (3DCMM) to enable the stepwise creation of full 3D heads from a single 2D face image (see Fig. 1). The process involves initially reconstructing the face region (see Fig. 1 (b) and Fig. 1 (c)) and subsequently predicting the full head (see Fig. 1 (d) and Fig. 1 (e)). In this study, we developed an automatic pipeline that facilitates accurate full 3D head creation from a single image, which consists of three main stages: (1) construction of a 3D human head dataset, (2) development of 3DCMM with a UV map-based UNet (UV-UNet) for face-to-head prediction, and (3) reconstruction of the 3D face region and creation of the full head using 3DCMM. Compared to our

Manuscript received 24 January 2024; revised 24 June 2024. This work was supported in part by the Research Grants Council of Hong Kong under Grant PolyU 15603419 and Grant PolyU 15606321, in part by the National Science and Technology Council under Grant 111-2221-E-006-112-MY3 and Grant 110-2221-E-006-135-MY3, Taiwan, and in part by The Hong Kong Polytechnic University under Grant P0048387, Grant P0042740, Grant P0044520, Grant P0043906, Grant P0049586, and Grant P0050657. *(Corresponding Authors: Yan Luximon and Ping Li.)*

Jie Zhang is with the with the Faculty of Applied Sciences, Macao Polytechnic University, Macau 999078, China (e-mail: peterzhang1130@163.com).

Ping Li is with the Department of Computing and the School of Design, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

Kangneng Zhou is with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: elliszkn@163.com).

Yan Luximon is with the School of Design, The Hong Kong Polytechnic University, Hong Kong, and also with the Laboratory for Artificial Intelligence in Design, Hong Kong (e-mail: yan.luximon@polyu.edu.hk).

Tong-Yee Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 70101, Taiwan (e-mail: tonylee@ncku.edu.tw).

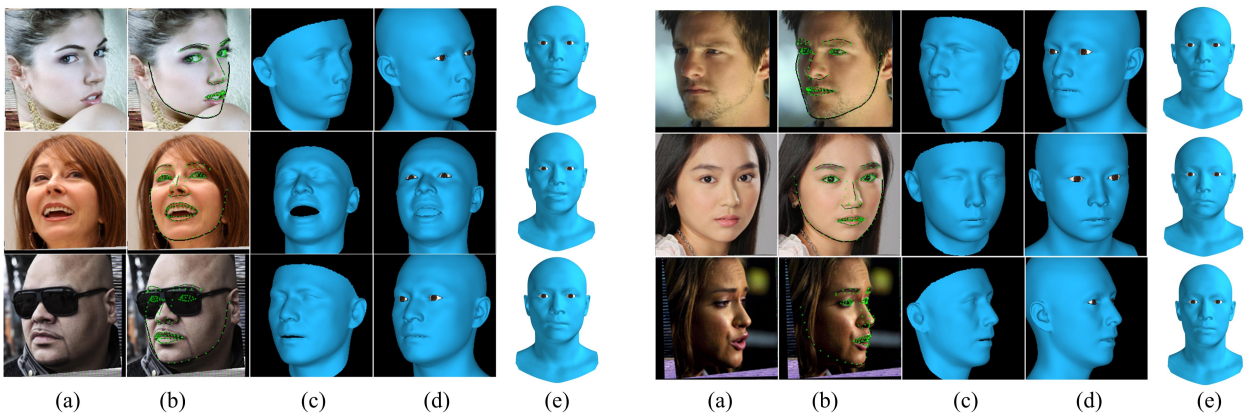


Fig. 1: Our 3D full head creation from a single-face image. (a) 2D face images from the CelebA dataset [6], (b) face alignment results with projected 150 landmarks from (c) our 3D reconstructed face, (d) 3D created full head, (e) 3D full head without expressions.

previous work, which proposed a linear face-to-scalp model transformation approach, this study represents significant advancements in creating full 3D heads. We constructed a more robust 3D morphable model by combining datasets of Chinese and Caucasians subjects, developed a nonlinear UV-UNet for predicting the full head from face regions, and successfully applied these methods to achieve full 3D head creation from a single 2D image. The main contributions of our research are:

- We constructed powerful 3DCMMs with UV map-based UNet via large-scale datasets of real 3D human heads to generate 3D faces and predict 3D scalps for full heads.
- We proposed a novel and accurate stepwise 3DCMM-based full-head creation pipeline and fully exploited image-level information by incorporating two new and effective losses in 3DCMM-based 3D face reconstruction: facial boundary-aware and structure-aware losses.
- We demonstrated the superiority of our 3D comprehensive morphable models with UV map-based UNet through several qualitative and quantitative comparisons and introduced some new applications for our model.

II. Related Work

A. 3D Head Datasets

The availability of 3D face/head datasets plays a vital role in constructing 3D statistical models and training deep networks with relevant data. While several 3D human face datasets, including Facewarehouse [9], MeIn3D [7] and FaceScape [8], exist, there is a scarcity of large-scale, precise 3D human head datasets. The only dataset of this kind is HeadSpace [19], which includes 1,519 subjects, primarily of Caucasian ethnicity. Existing 3D face datasets are typically created for face-related tasks, such as face reconstruction and recognition, while 3D head datasets have broader applications [19], such as headwear design and craniofacial diagnosis. The dearth of large-scale 3D head datasets poses challenges in head scan collection due to hair occlusion. In this paper, we present our effort to construct a comprehensive 3D head dataset on a large scale for creating 3D statistical models and training head prediction networks.

B. 3D Morphable Models

3DMMs are statistical models designed to capture the primary components that represent variations in shape and texture within a given training dataset, forming a foundation for tasks related to 3D face and head generation and reconstruction. Numerous powerful 3DMMs for human faces and heads have been developed, including BFM [20], LSFM [7], FaceScape [8], FaceWarehouse [9], LYHM [19], and HiFi3D++ [21]. However, few 3DMMs are constructed using large-scale datasets with a balanced representation of both Caucasian and non-Caucasian populations spanning different age groups, from children to the elderly. Therefore, in this paper, we describe our more comprehensive and powerful 3DMM of human faces, which we created by combining our large-scale Chinese head datasets with an existing large-scale Caucasian head dataset [19]. For a more extensive exploration of 3DMMs, we recommend referring to a recent survey by Egger et al. [22].

C. 3D Face Reconstruction

With the assistance of 3DMMs, self-learning-based methods [12], [13], [15], [16], [23] have gained popularity in predicting 3D face meshes from 2D face images. These methods utilize the shape or texture coefficients of a 3DMM in an analysis-by-synthesis scheme. Compared to supervised CNN-based methods [24], [25], [26], [27], [28], which require large-scale datasets of 2D face images and reference 3D face shapes, the main advantage of self-learning-based methods is their ability to train solely on 2D face images, eliminating the need for 3D face shape references. Commonly encountered image-level losses in these methods include photometric loss, perceptual loss, and landmark loss. However, other largely unexplored image-level losses exist that have the potential to affect the accuracy of 3D face reconstruction, e.g., facial boundary-aware and structure-aware losses. Hence, we comprehensively considered these image-level losses in 3D face reconstruction. A more comprehensive review of 3D face reconstruction can be found in a recent survey [1].

D. 3D Head Reconstruction

While previous studies primarily focused on reconstructing the 3D face [16], [13], [12], or even a tightly cropped facial region [15], there are limited studies that directly reconstruct the full 3D head from 2D images [14], [18]. A facial image alone provides only facial region information, and lacks scalp information due to hair occlusion. In these 3D weekly supervised full-head reconstruction approaches [14], [18], the 3DMM of human heads is used instead of the 3DMM of human faces, where the scalp shape is often influenced by regularization loss. Hence, in our paper, we adopted a different pipeline for creating the full 3D head in a stepwise manner: 3D face reconstruction using the 3DMM of human faces followed by 3D head completion using a supervised learning network.

E. 3D Head Prediction

In previous studies, 3D head prediction has been defined as an estimation of the scalp region from the face region to generate a full head, which is the same as 3D scalp completion [2], partial data reconstruction [19], and 3D cranium prediction [29]. Since 3D scalp regions, different from 3D face regions, cannot be captured directly using scanners, it is significant and useful to be able to predict the full 3D head. Some previous studies [29], [11] computed a model-coefficients mapping matrix between 3DMMs of human faces and scalps/heads to achieve 3D scalp prediction. It is assumed that this model mapping relationship is linear in their studies. Some previous studies [2], [19] also used 3DMMs of human heads to fit the 3D face regions to estimate the model coefficients, then produced the full 3D head. However, the accuracy of scalp shape prediction is susceptible to the regularization term in this 3DMM-based fitting approach. To address this limitation, we adopted an alternative approach. By utilizing a large-scale dataset of paired real face and head meshes as a training dataset of UV maps, we developed a UV-map-based UNet (UV-UNet) architecture to achieve accurate 3D full-head prediction.

III. Full 3D Head Creation

A. Overview

Our pipeline for creating a full 3D head consists of two primary steps, depicted in Fig. 1: (1) 3D face reconstruction from a single 2D image and (2) 3D head prediction based on the reconstructed 3D face. The foundation of this pipeline is our 3DCMM, which comprises the 3DMM for human faces used in 3D face reconstruction and the UV-UNet utilized for 3D head prediction. Consequently, the construction of the 3DCMM is the initial step. Our 3DCMM is built upon a large-scale dataset of real human heads, offering significant advantages in terms of its ability to facilitate accurate 3D face reconstruction and 3D scalp prediction for 3D head creation. Furthermore, compared with other model-based 3D face reconstructions [12], [13], [18], [16], a significant difference is that our method uses two critical novel facial boundary-aware and structure-aware losses, thereby producing more accurate facial contours and structure consistency.

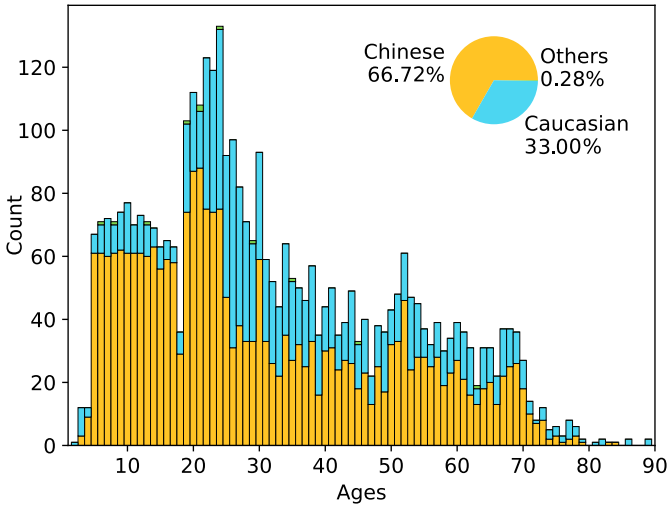


Fig. 2: Ethnics and age information of the subjects in the 3D head database for our 3DCMM construction.

B. 3D Comprehensive Morphable Model Construction

1) Full-Head Dataset: To establish a comprehensive 3DCMM, we combined our Chinese head database (i.e., Adult-Heads [31] and Children-Heads [32]) with a Caucasian head dataset (HeadSpace [19]) for a total of 3,846 subjects (including 49.90% females and 50.10% males). Their ethnicity and age information are shown in Fig. 2. Compared to previous datasets [8], [9], our dataset has balanced representation of both Chinese and Caucasian populations spanning different age groups, from children to the elderly. Once we had created the full combined database, we identified 51 landmarks as the registration constraints on the head scan surface using the Face Alignment Network (FAN) [33] (see Step ❶ of Fig. 3) and registered these 3D scans into parameterized heads by applying the widely-used non-rigid iterative closest points (NICP) algorithm [34] with 3D face [35] and head [21] templates (see Step ❷ of Fig. 3), respectively. Each registered 3D facial/head mesh has 53,215/20,052 vertices and 105,840/39,984 triangles. The position relationship between the face and head is shown in Fig. 4, where Region-F and Region-S indicate the facial and scalp regions in the full head.

2) Human Faces' 3DMM: Based on these registered facial meshes, we use the General Procrustes Analysis (GPA) [36] to unify their size, poses and positions, and then Principal Component Analysis (PCA) to extract their principal components (PCs) (see Step ❸ of Fig. 3). Each registered mesh geometry is represented as a shape-vector, S=(x_1, y_1, z_1, …, x_m, y_m, z_m)^T $\in \mathbb{R}^{3m}$, that contains the x, y, z coordinates of its m vertices. Consequently, a novel morphable face shape S_f (see Step ❻ of Fig. 3) can be described and constructed with the average shape $\overline{S_f}$, the extracted PCs P_f, and the facial shape representation coefficient vector α_f:

$$S_f = \overline{S_f} + \sum_{i=1}^{n} \alpha_{f,i} P_{f,i} = \overline{S_f} + P_f \alpha_f, \tag{1}$$

where n is the number of the facial shape PCs. The probability $p(\alpha_f)$ of coefficients α_f is given by $p(\alpha_f)$ $\sim$

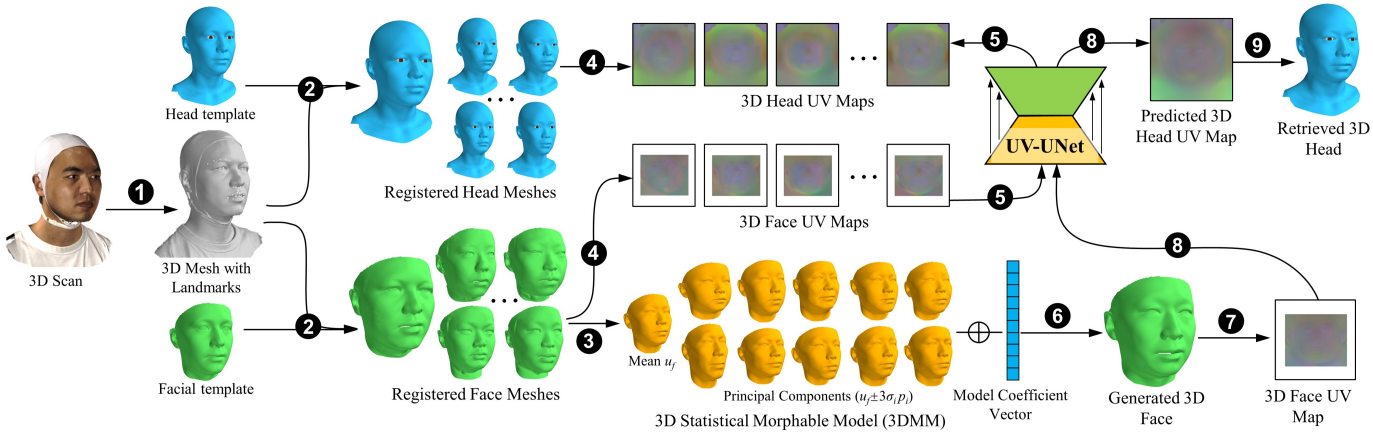


Fig. 3: Overview of our automatic 3DCMM construction pipeline for 3D full head creation, including two main components: (1) a 3DMM to generate 3D human face, and (2) a UV-UNet to predict 3D full head. ❶ 3D facial landmark detection, ❷ 3D face/head registration, ❸ human face's 3DMM establishment, ❹ 3D face/head UV map generation, ❺ UV-UNet training, ❻ 3D face generation, ❼ 3D face UV map generation, ❽ 3D head UV map prediction, ❾ 3D full head retrieval

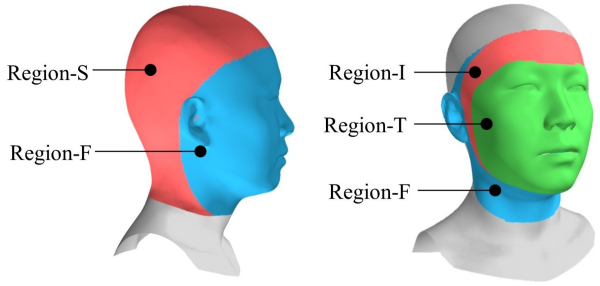


Fig. 4: Our predefined face and head segmentation regions. Facial regions [30]: Region-T ∈ Region-I ∈ Region-F. Scalp regions [11]: Region-S.

$\exp\left[-\frac{1}{2}\sum_{i=1}^{n-1}\left(\frac{\alpha_{f,i}}{\sigma_{f,i}}\right)^2\right]$, where $\sigma_{f,i}^2$ is the eigenvalues of the shape covariance matrix.

The FaceScape dataset [8] comprises Chinese face scans that exhibit highly precise texture details captured using a multi-view system under controlled illumination. Furthermore, the HeadSpace dataset [19] consists of head scans with texture information primarily from Caucasian subjects. Leveraging these datasets, we utilized 399/497 available face scans from FaceScape/HeadSpace to establish a 3DMM of facial textures, which involved employing a similar model construction pipeline. To describe and construct a novel face texture T_f, we combined the average texture $\overline{T_f}$, the PCs Q_f with facial shape representation coefficient vector δ_f, using the following formulation:

$$T_f = \overline{T_f} + \sum_{i=1}^{n} \delta_{f,i} Q_{f,i} = \overline{T_f} + Q_f \delta_f. \tag{2}$$

3) Face-to-Head UV-UNet: For any given 3D facial mesh, we proposed a UV-UNet to predict the full head mesh through a 2D UV map. To apply the 2D convolutional neural networks (CNNs) to the 3D facial/head meshes, it is necessary to transform the 3D meshes into 2D UV maps using predefined UV coordinates. Several subsequent steps were performed to achieve the full head prediction.

(a) Mesh Alignment. To achieve face pose normalization, we employed Procrustes analysis (PA) to align the face scan to the face template. The objective of PA is to compute a linear transformation matrix $\mathbf{R}$ that minimizes the total distance between the vertices of the facial template ($\tilde{S}_f \in \mathbb{R}^{m\times 3}$) and the registered facial meshes ($S_{f,i} \in \mathbb{R}^{m\times 3}$), as given by:

$$\min_{\mathbf{R}} \sum_{i=1}^{m} \left\| \mathbf{R} S_{f,i} - \tilde{S}_{f,i} \right\|_F^2, \tag{3}$$

where, $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}$ is a constraint on the transformation matrix, and $\|\cdot\|_F$ denotes the Frobenius norm, which corresponds to the element-wise Euclidean distance. The resulting aligned facial mesh is computed as $\ddot{S}_f = \mathbf{R}S_f$. Similarly, each corresponding head mesh S_h can be aligned to the facial template using the same linear transformation matrix $\mathbf{R}$, yielding $\ddot{S}_h = \mathbf{R}S_h$.

(b) UV Map Generation. The differences D_f between the aligned facial meshes and the facial template were computed and scaled to 0∼1: $D_f = \left(\left(\ddot{S}_f - \tilde{S}_f\right)/D_{max} + 1\right)/2$, and then rendered into 2D UV maps (256 × 256 pixels) using predefined UV coordinates (see Step ❹ of Fig. 3). Here, D_{max}=25, which was confirmed as the maximal difference between all aligned facial meshes and the template. We also used a similar approach to compute the 2D head UV maps (256 × 256 pixels): $D_h = \left(\left(\ddot{S}_h - \tilde{S}_h\right)/D_{max} + 1\right)/2$.

(c) Network Development and Training. ResNet34 [37], [38] was used as the backbone for constructing the UV-UNet to predict the head UV maps from the facial UV maps (see Step ❺ of Fig. 3). The training loss $L(x)$ in supervised learning is computed as the mean absolute difference between the predicted head UV map $D^{h,p}$ and the ground truth $D^{h,t}$:

$$L(x) = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} \|D_{i,j}^{h,p} - D_{i,j}^{h,t}\|, \tag{4}$$

Here, $W = H = 256$. For our study, we divided the UV

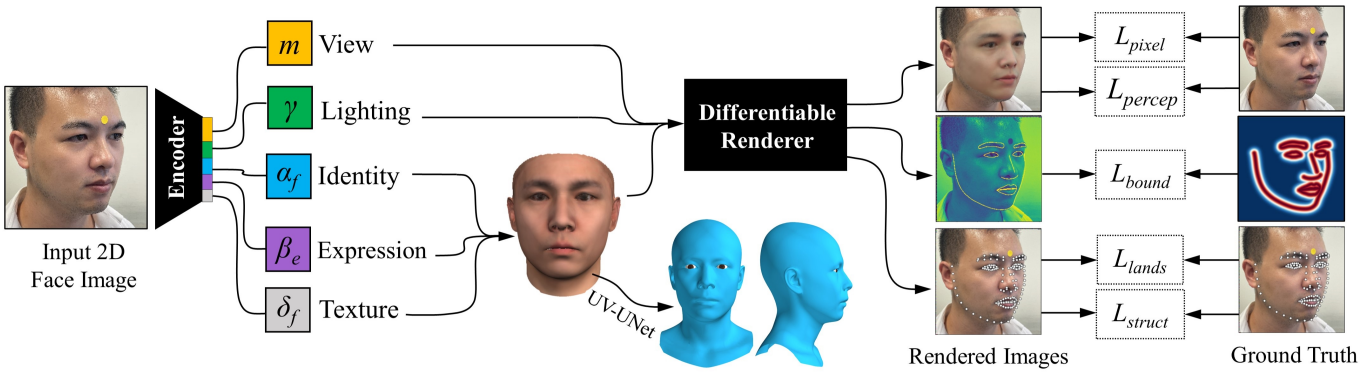


Fig. 5: Overview of our 3DCMM-based face reconstruction procedure with self-supervised learning and head prediction using UV-UNet. Besides the above image-level losses (including pixel-wise $L_{pixel}(x)$, perceptual identity L_{percep}, boundary-aware L_{bound}, facial landmark L_{lands}, and structure-aware L_{struct} losses), there are still two other commonly used losses: mesh skin variance loss L_{text} and regularization term L_{regual}.

map dataset into three subsets: 90% was used for training, 5% for validation, and 5% for testing purposes. To implement and train our model, we employed PyTorch [39] along with the Adam optimizer [40] and conducted approximately 500K iterations with a batch size of 16 and an initial learning rate of 1e-4.

(d) 3D Head Prediction and Retrieval. A bilinear interpolation method with predefined UV coordinates was adopted to retrieve the vertices' positions of the 3D head mesh from the generated 2D UV map: $S_{h,i} = \tilde{S}_{h,i} + \left(D_h^{i \to k} \times 2 - 1\right) \times D_{max}$, where $D_h^{i \to k}$ is the interpolated value in the predicted UV map corresponding to the ith vertex in the head mesh (see Steps ❼, ❽ and ❾ of Fig. 3).

C. Model-Based 3D Face Reconstruction

1) Differentiable Renderer: Given an unconstrained 2D face image, our goal is to train an encoder (ResNet50 [37]) to take a 2D facial image input and decompose it into outputs of 3D facial shape S_f, facial albedo T_f, illumination l, and viewpoint w, as illustrated in Fig. 5. The input image I^R can be reconstructed from these four components in two steps of lighting Λ and reprojection Π, as: $I^R = \Pi(\Lambda(S_f, l, T_f), S_f, w)$.

To deal with the varying facial expressions in 2D images, we used our 3DMM of facial textures (see Equation 2) to estimate the facial albedo, and integrated the 3D face expression bases P_e (built from FaceWarehouse [9]) with our 3DCMM-identity shape bases P_f (see Equation 1) into a complete 3D face model, as follows: $S_f = \overline{S_f} + P_f \alpha_f + P_e \beta_e$, where β_e is the shape coefficients for face expression bases P_e. For the lighting model, since a Lambertian surface was assumed for the face [18], [12], [13], we approximated the scene illumination using Spherical Harmonic (SH [41]) basis functions with the first three bands (the parameters $\gamma \in \mathbb{R}^{27}$). For the camera model, similar to previous studies [12], we employed a global camera model of perspective projection to project the 3D face model onto the 2D image plane, where the camera position is determined by an estimated camera matrix m, which can be computed from a rotation vector $\nu \in \mathbb{R}^3$ and a translation vector $t \in \mathbb{R}^3$.

2) Network Objective Function: To decompose the 2D images successfully, we calculated image-level losses (see Fig. 5), facial skin color variances $L_{text}(x)$ and regularization term $L_{regual}(x)$ as the training losses and sought a CNN-based encoder to minimize them. Image-level losses consist of pixel-wise $L_{pixel}(x)$, facial landmark $L_{lands}(x)$, facial structure-aware $L_{struct}(x)$, facial boundary-aware $L_{bound}(x)$, and perceptual identity $L_{percep}(x)$ discrepancies between the input and rendered 2D images. All our training losses $L(x)$ are shown as follows:

$$\begin{aligned} L(x) =& w_1 L_{pixel}(x) + w_2 L_{percep}(x) + \\ & w_3 L_{lands}(x) + w_4 L_{struct}(x) + \\ & w_5 L_{bound}(x) + w_6 L_{text}(x) + w_7 L_{regual}(x), \end{aligned} \tag{5}$$

where w_i (i=1,2,...,7) are hyperparameters balancing the weights of different losses. L_{text} and L_{regual} are computed using the same formulates as the previous method [12].

Pixel-Wise Loss We utilized the accurate face parsing mask (produced using MaskGAN [42]) to gain robustness to facial occlusions and defined the pixel-wise loss L_{pixel} between the input raw image I and its rendered counterpart I^R as follows:

$$L_{pixel}(x) = \frac{1}{N_E} \sum_{i \in M} P_i \| I_i - I_i^R \|^2, \tag{6}$$

where P is the parsing mask with different values in different regions and N_P is the sum of non-zero pixels in the parsing mask P. In our method, only Region-I (see Fig. 4 (b)) of the rendered facial mesh, M, was used to produce a new 2D image.

Perceptual Identity Loss To incorporate this loss, we utilized a pre-trained face recognition network f (FaceNet [43]) to extract perceptual features from the input image I and the rendered image I^R. The cosine distance between these features was then computed as the perceptual identity loss L_{percep}, which quantifies their perceived similarity as:

$$L_{percep}(x) = 1 - \frac{f(I) f(I^R)}{\| f(I) \|_2 \cdot \| f(I^R) \|_2}. \tag{7}$$

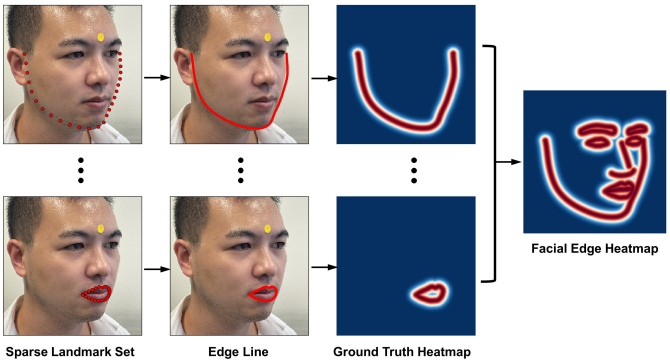


Fig. 6: An illustration of the ground-truth facial boundary heatmap generation process, where the standard deviation σ is set as 7. Each row represents the process of one specific facial boundary, including outer facial contour, outer/inner mouth, left/right eyebrow, upper/lower eyelid, and nose bridge/base.

Facial Landmark Loss A publicly available facial landmark detector (Baidu) was used to identify 150 points as sparse ground-truth facial landmarks for the training images. Similar to [14], [18], we modelled the facial landmarks as either dynamic or static 3D landmarks and defined the facial landmark loss L_{lands}, as follows:

$$L_{lands}(x) = \frac{1}{N_L}\sum_{k=1}^{N_L} w_k \|q_k - q_k^R\|^2, \tag{8}$$

where q and q^R are the landmarks in the input image and the corresponding vertices in the projected mesh, respectively. w_k is the landmark weight, which is experimentally set as 20/0.8/1.0 for nose/boundary/other landmarks.

Structure-Aware Loss The widely used 2D sparse facial landmark loss only considers point-to-point distances, making it easily unrobust to extreme facial poses and highly sensitive to facial occlusion. In reality, the human face has a fairly consistent structure with the facial components maintaining fairly stable relative distances [44]. Hence, we added a face structure-aware loss to constrain landmark positions in a global context.

As a preprocessing step, we defined a graph structure of a template face with a small, opened mouth using Delaunay triangulation of sparse landmarks. We computed the distance between the ground-truth edges e and projected facial edges e^R as the structure-aware loss L_{struct}:

$$L_{struct}(x) = \frac{1}{N_E}\sum_{i=1}^{N_E} w_i \|e_i - e_i^R\|^2, \tag{9}$$

where N_E is the sum of edges in the graph structure, and w_i is the edge weight, which is experimentally set as 0.8 for the edges with landmarks in the facial contour and 1.0 for the rest of the edges.

Boundary-Aware Loss Image boundaries provide information about 2D shape independently of the texture and illumination [45]. Facial boundary lines help with 2D face alignment significantly [46], which inspired us to utilize facial boundaries to improve 3D face reconstruction. Hence, we proposed a novel facial boundary-aware loss L_{bound} as:

$$L_{bound}(x) = \frac{1}{N_B}\sum_{i=1}^{N_B} \|1 - H(\sum_{k=1}^{3} w^{i,k} S_f^{i,k})\|^2. \tag{10}$$

where N_B is the number of boundary points. To compute this loss, two steps were performed to produce the ground-truth heatmap H and reconstructed facial boundary V_b:

(a) For the input image, the facial boundaries were interpolated from sparse facial landmarks to obtain a dense boundary line. Then, a series of heatmaps was generated by applying a 2D Gaussian function with a standard deviation of σ pixels centred on the location of each point successively. Finally, all heatmaps were fused into a single boundary heatmap H as the input ground truth by selecting only the maximum of each pixel with the same position in all heatmaps (see Fig. 6).

(b) For the reconstructed mesh, the facial boundaries were predesigned using dynamic dense points S_b. These dense points S_b are computed by interpolating at each corresponding vertex $S_f^{i,k}$ and their barycentric coordinates with the weights w^k: $S_b^i = \sum_{k=1}^{3} w^{i,k} S_f^{i,k}$, then their heat values $H(S_b)$ were retrieved from the boundary heatmap using a bilinear interpolation method based on their X and Y values. Theoretically, the boundary heat values should be very large and nearly 1.0.

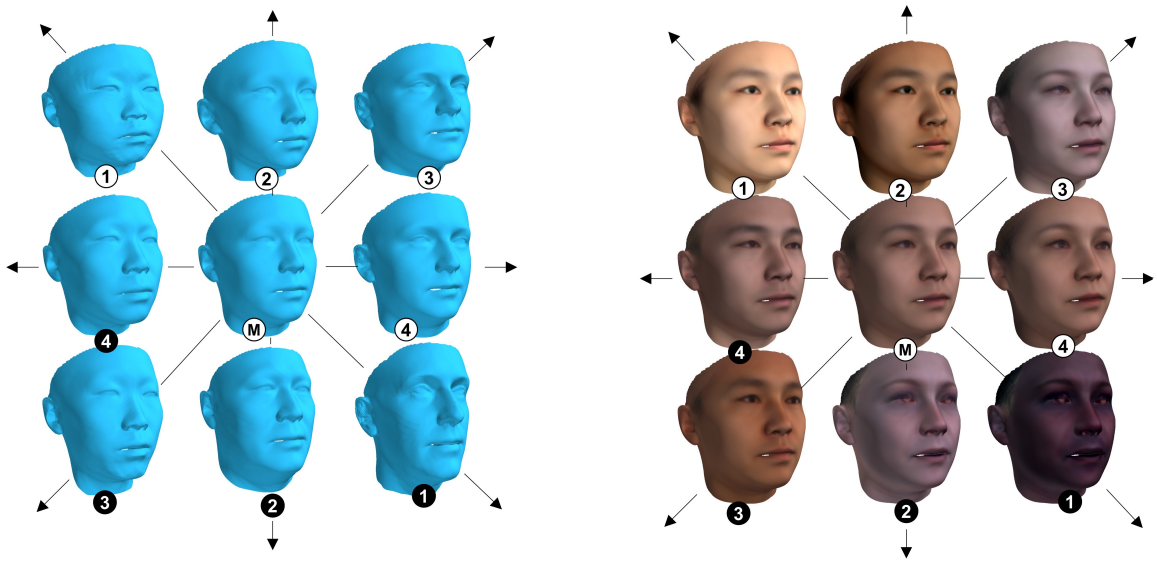
3) Implementation Details: To train our CNNs, we collected 2D face images from multiple sources, including FFHQ [48], AAF [49], CACD [50], 300W-LP [24] and SCUT-FBP5500 [51]. We balanced the race and pose distributions and obtained approximately ∼150k 2D face images as the training and validation dataset. The encoder needs to regress 254 parameters, including the model identity ($\alpha_f \in \mathbb{R}^{100}$), expression ($\beta_e \in \mathbb{R}^{57}$) and texture ($\delta_f \in \mathbb{R}^{64}$) coefficients, lighting ($\gamma \in \mathbb{R}^{27}$), and view ($m \in \mathbb{R}^6$) parameters. The input image size is 224×224 pixels. We implemented our model using PyTorch [39] and PyTorch3D [52]. For network training, we employed the Adam optimizer [40] with an initial learning rate of 1e-4. The training process consisted of approximately 500K iterations, with a batch size of 16.

IV. Experimental Results

A. Morphable Model Evaluation

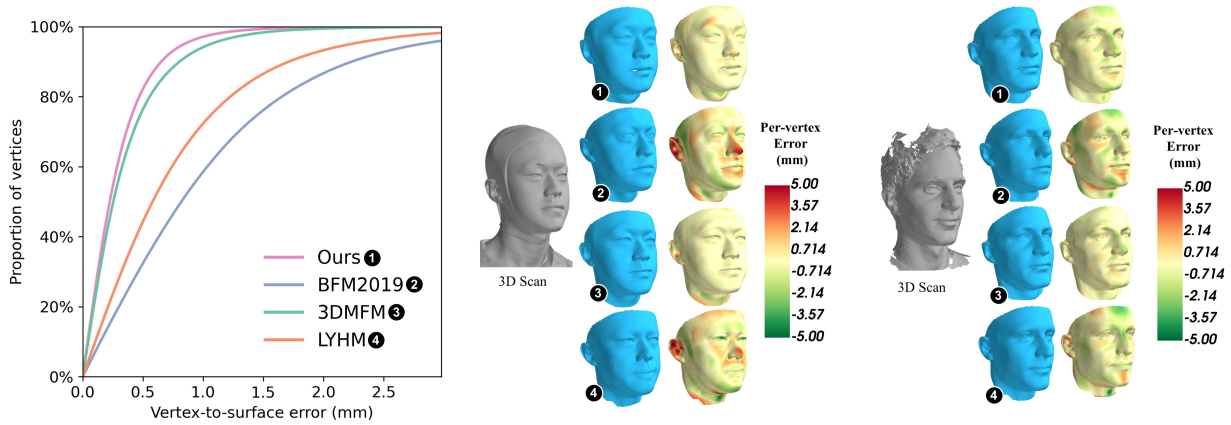
We developed a 3DCMM that encompasses facial shapes and textures, as illustrated in Fig. 7 (a) and Fig. 7 (b), respectively. To ensure consistent sizes of the registered meshes, we applied Generalized Procrustes Analysis (GPA). This approach allowed the eigenvector loading distributions of the PCs to solely represent mesh shapes, independent of sizes. In our 3DCMM, the first 100/64 PCs accounted for 98.20%/93.05% of the explained variances of the 3D facial shapes/textures in the training datasets. This indicates that the PCs utilized in our 3D face reconstruction provided a concise representation of the training datasets.

To highlight the advantages of our model, we compared it with BFM2019 [47], LYHM [19], and our previous 3D Morphable Face Model (3DMFM) [11] (see Fig. 8, only the vertices within the facial region were considered). We assessed the generalization ability of each model using 30 Chinese [11] and 30 Caucasian [53] facial meshes with neutral expressions



(a) (b)

Fig. 7: Our comprehensive 3D morphable models: (a) Facial shape 3DMM and (b) Facial texture 3DMM. The figures show the mean shape M (centre) and the first four PCs with weights of 3/-3 σ_i (white/black circle label), where σ_i represents the standard deviation of the PCs.



(a) (b) (c)

Fig. 8: 3DMM generation ability evaluations. (a) Quantitative comparisons. (b) Qualitative comparisons of a Chinese subject. (c) Qualitative comparisons of a Caucasian subject. (1) Our constructed 3DCMM. (2) BFM 2019 [47], (3) our previous 3DMFM [11], and (4) LYHM [19].

(aged 18-60 years). Employing the same model-fitting method, we obtained the closest facial mesh and calculated the mean error. To ensure fairness, only the first 100 PCs were used for facial reconstruction across all models. The mean reconstructed face errors (ME) for our 3DCMM, 3DMFM, LYHM, and BFM2019 were 0.30 $\pm$ 0.29, 0.36 $\pm$ 0.36, 0.78 $\pm$ 0.72, and 1.04 $\pm$ 0.91 mm, respectively, as shown in Fig. 8 (a). Specifically, the qualitative comparisons in Fig. 8 (b) and Fig. 8 (c) reveal significant errors in the nose, forehead, and mouth regions for the Chinese subject in the BFM2019 and LYHM results. In conclusion, our 3DCMM outperforms other 3DMMs in terms of generation capability.

B. Face Reconstruction Evaluation

We evaluated our 3D face reconstruction qualitatively and quantitatively using a 2D face image dataset (CelebA [6]), and two 3D face datasets (FaceWareHouse [9], FaceScape [8]). We then compared our results with publicly available methods, (i.e., PRNet [25], 2DASL [54], 3DDFA-V2 [26], Deep3DFace [12], MGCNet [16], MoFa [15], InverseFaceNet [58], Tewari et al. [57], 3DDFA-V3 [55], and HRN [56]). In the qualitative evaluation, we followed the previous protocol [30], [59], [14] using PA, with a set of corresponding landmarks to align (rotate, translate, and scale) the reconstructed face to the ground-truth face initially, then, perform a scan-to-mesh distance-based rigid alignment on them, and finally compute the mean per-vertex distance (ME) as the evaluation metric.

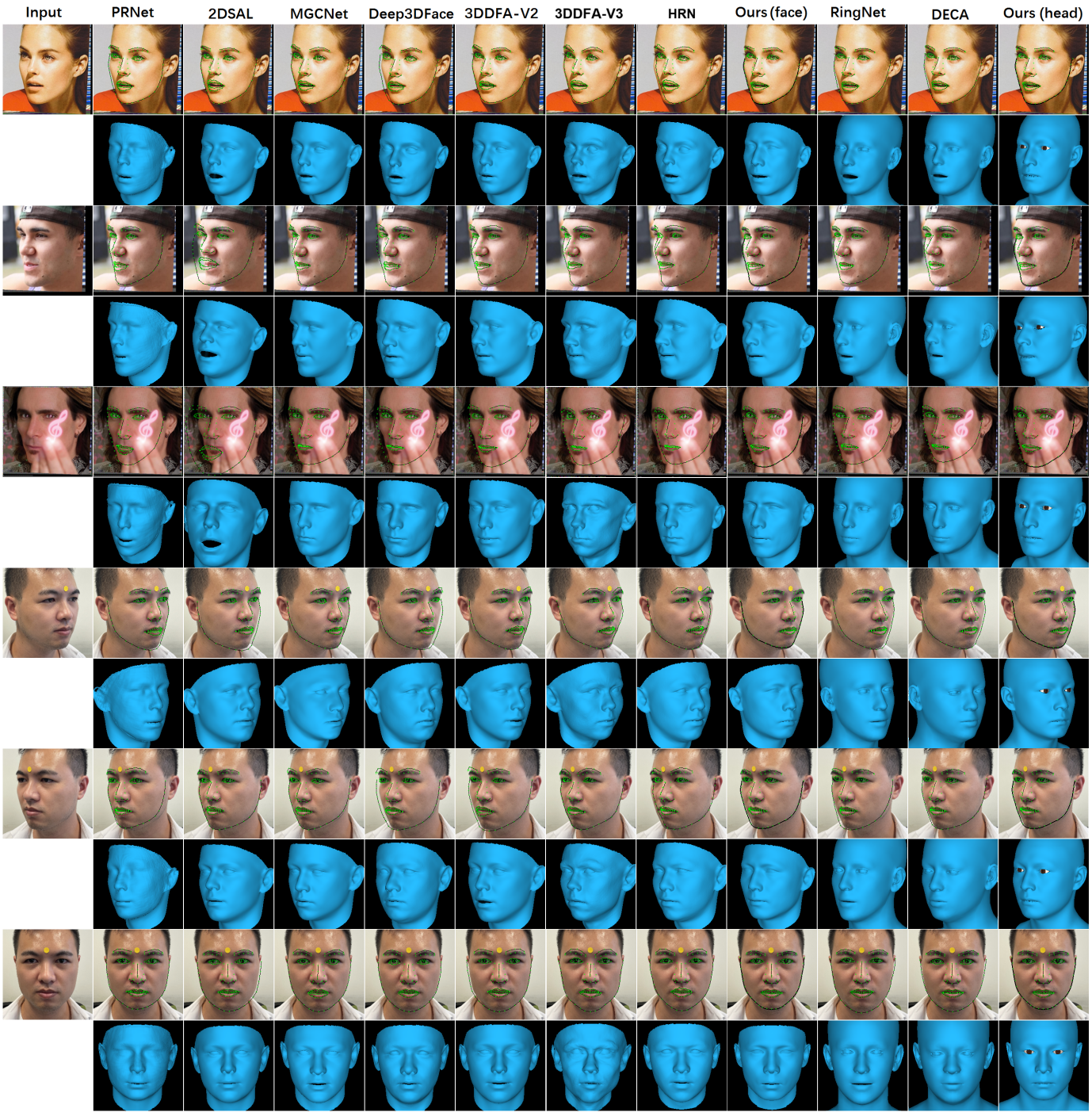


Fig. 9: Qualitative comparisons of 3D face reconstruction and head creation on face images from CelebA [6] and our head image datasets: PRNet [25], 2DASL [54], MGCNet [16], Deep3DFace [12], 3DDFA-V2 [26], 3DDFA-V3 [55], HRN [56], ours (face), RingNet [14], DECA [18], and ours (head). Note that our method can regress facial shape with higher facial boundary and structure consistency and complete the accurate 3D scalp region to produce full head geometry, which is robust to facial occlusion and extreme poses.

Qualitative Evaluation In the qualitative evaluation of facial geometry, we used face images from CelebA [6] to reconstruct the 3D facial shape and compared our results with two state-of-the-art supervised-learning-based face reconstruction methods, (PRNet [25] and 3DDFA-V2 [26]), and five state-of-the-art self-learning-based face reconstruction methods, (Deep3DFace [12], MGCNet [16], 2DASL [54], 3DDFA-V3 [55], and HRN [56]), as shown in Fig. 9. From this, it is clear that, compared with other methods, our method achieves better facial contour alignment and constructs a more

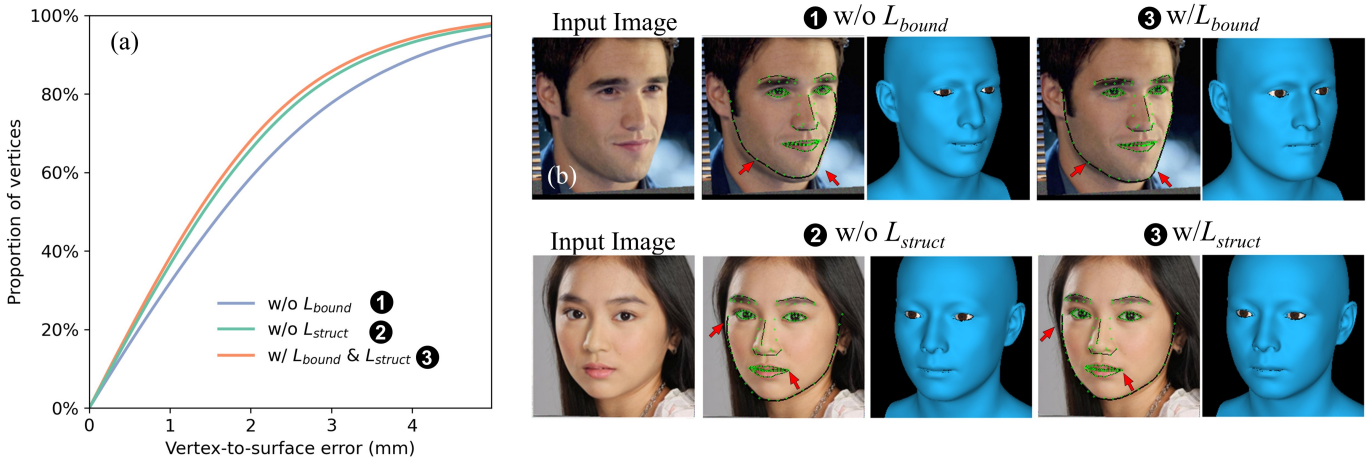


Fig. 10: Ablation experiment. (a) Quantitative results. (b) Qualitative effect of L_{bound} (guaranteed to construct more consistent facial contours for face images). (c) Qualitative effect of L_{struct} (improved generation of stable global facial structures, such as landmarks in the mouth, lower/upper inner lips, and facial contours).

TABLE I: Quantitative comparisons of 3D face reconstruction on 180 meshes of nine subjects with 20 expressions from FaceWarehouse [9]:ME ± standard deviation (mm).

Region	Tewari et al. [57]	MoFa [15]	InverseFaceNet [58]	MGCNet [16]	Deep3DFace [12]	Ours
Region-T	1.84±0.38	2.19±0.54	2.11±0.46	2.18±0.35	1.74±0.29	**1.60±0.29**
Region-I	-	-	-	2.23±0.39	1.78±0.32	**1.63±0.29**
Region-F	-	-	-	2.47±0.36	2.18±0.47	**2.12±0.31**

TABLE II: Quantitative comparisons of 3D face reconstruction on 900 meshes of nine subjects with 20 expressions and five facial poses from FaceScape [8]: ME ± standard deviation (mm).

Region	2DASL [54]	PRNet [25]	3DDFA-V2 [26]	Deep3DFace [12]	Ours
Region-T	2.78±0.61	2.68±0.52	2.95±0.56	2.11±0.50	**2.12±0.45**
Region-I	2.95±0.62	2.91±0.55	2.98±0.58	2.21±0.50	**2.13±0.46**
Region-F	3.39±0.62	3.32±0.54	3.52±0.65	2.71±0.53	**2.67±0.50**

TABLE III: Quantitative comparisons of 3D head creation on 75 meshes of 25 subjects with three different head poses: ME ± standard deviation (mm).

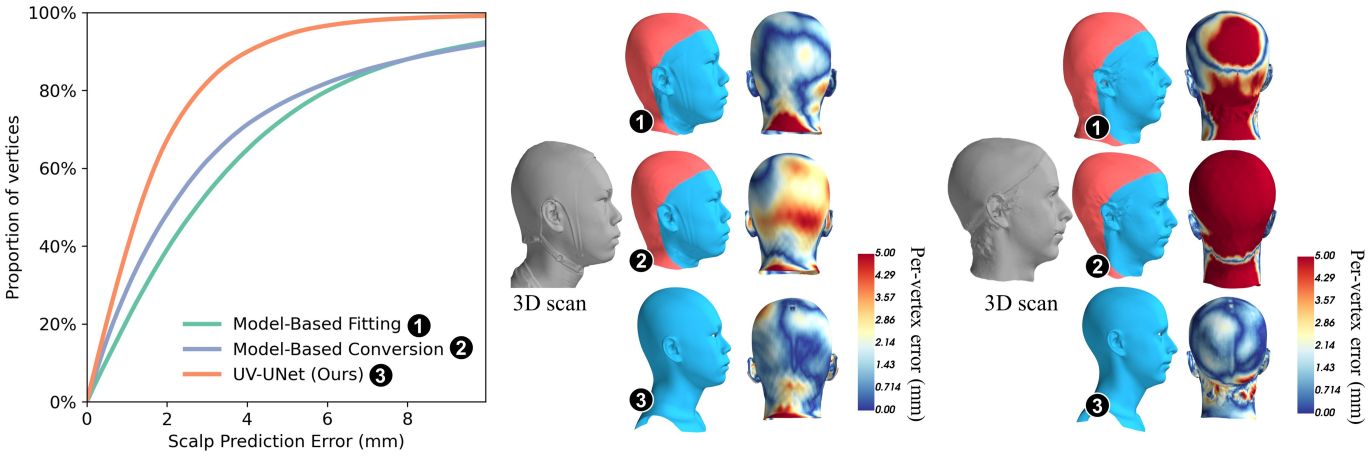
Region	RingNet [14]	DECA [18]	Ours
Region-I	3.09±2.24	**1.97±1.60**	2.02±1.46
Region-F	3.46±2.63	2.33±2.00	**2.25±1.78**
Region-S	5.00±4.06	4.49±3.87	**4.24±3.89**

accurate overall face shape, even for face images with extreme poses or facial occlusions.

Quantitative Evaluation For the quantitative evaluation of facial geometry, we used 180 facial meshes (nine identities, 20 expressions each) selected from the FaceWareHouse dataset [9] created in a previous study [57]. Three face regions were evaluated as shown in Fig. 4 (b): the tightly cropped face region (Region-T, same as in [57]), inner facial region (Region-I) with more cheek area (same as in [12]) and full face (Region-F) with neck and ears. We compared our results with five model-based methods: Deep3DFace [12], MoFa [15], MGCNet [16], Tewari et al. [57], and InverseFaceNet [58]. The MEs of face reconstruction methods are shown in Table I; our method is a significant improvement upon the previous ones.

To evaluate the face pose influences, we compared our results with those of two supervised-learning-based methods: PRNet [25], and 3DDFA-V2 [26], and two self-learning-based methods: Deep3DFace [12] and 2DASL [54], as shown in Table II. We used 900 facial meshes (nine identities, 20 expressions each, five different facial poses each) from the FaceScape dataset [8]. Table II shows that the mean face reconstruction error of our method is slightly greater than that of Deep3DFace [12] for Region-T, but less than that of other methods for Region-I and -F, indicating that our method made notable improvement on the overall reconstruction of facial shape.

Ablation Study To showcase the effectiveness of our proposed novel losses, we performed an ablation study, which



(a) (b) (c)

Fig. 11: Head prediction method comparisons. (a) Quantitative comparisons. (b) Qualitative comparisons of a Chinese subject. (c) Qualitative comparisons of a Caucasian subject. (1) 3DMM-based conversion method [11], (2) 3DMM-based fitting method [2] and (3) Our UV-UNet-based method.

involved training our network with and without the boundary-aware/structure-aware loss components. In this evaluation, we utilized 180 facial meshes (nine identities, 20 expressions) selected from the FaceWareHouse dataset [9]. The quantitative evaluation results are shown in Fig. 10 (a). The reconstruction errors for Region-I for the training networks with L_{bound} & L_{struct}, without L_{bound}, and without L_{struct} are 1.63 $\pm$ 0.29, 1.98 $\pm$ 0.37, and 1.71 $\pm$ 0.28, respectively. These results indicate that the inclusion of boundary-aware and structure-aware losses improves reconstruction accuracy. Furthermore, in the qualitative comparisons in Fig. 10 (b), the boundary-aware loss L_{bound} generates more consistent ground-truth facial boundaries for overall facial geometry. Additionally, in Fig. 10 (c), the qualitative comparisons reveal that structure-aware loss enhances the relative positions of facial landmarks, such as the closed eyes and mouth.

C. Head Prediction Evaluation

We first compared our 3D head creation from 3D facial regions with our previous 3DMM-based transformation method [11] and 3DMM-based fitting method [2] qualitatively and quantitatively, and then compared our 3D head creation from a single image with the publicly available methods (RingNet [14], DECA [18]) qualitatively and quantitatively on a 2D face image dataset (CelebA [6]) and our established 3D head datasets.

3D Scalp Prediction We compared our UV-UNet-based head prediction results from 3D facial regions with our previous 3DMM-based transformation method [11] and 3DMM-based fitting method [2] using 30 Chinese identities [11] and 30 Caucasian identities from HeadSpace [19], as shown in Fig. 11. The mean scalp prediction errors of our UV-UNet, 3DMM-based transformation, and 3DMM-based fitting methods are 1.87$\pm$0.82, 3.58$\pm$1.88, and 3.97$\pm$1.61 mm, respectively (see Fig. 11 (a)). Hence, our method is shown to predict scalp regions more accurately from 3D facial regions for 3D head creation.

3D Head Reconstruction In the qualitative evaluation of head geometry, we used the 2D face images from CelebA [6] to predict full 3D head shape and compared our results with two state-of-the-art self-supervised-learning-based head reconstruction methods (RingNet [14] and DECA [18]), as shown in Fig. 9. Compared with other methods, our method achieved better facial contour consistency. Furthermore, from the first row, it is clearly visible that our method can model a more accurate mouth closure, compared with DECA [18]. For the quantitative evaluation of full-head geometry, especially the scalp region, we collected an additional 75 full heads from 25 identities with three different head poses, ranging from frontal to profile view. We furthermore compared our results from a single image with RingNet [14] and DECA [18] using the same evaluation metrics [30], as shown in Table III. We evaluated three head regions: the inner facial region (Region-I), the full facial region (Region-F), and the scalp region (Region-S), as shown in Fig. 4 (a). Table III shows that our method had a similar rate of errors as DECA [18] in Region-I, but performs favorably against both [18] and [14] in Region-F and -S.

D. Runtime Analysis

The runtime performance of the proposed method was evaluated on a computer equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The analysis included the time taken for 2D face image reading and preprocessing (including facial landmark detection for alignment and cropping utilizing FAN [33]), 3D face estimation from 2D face images (using ResNet50 [37]), and 3D head prediction from 3D face regions (using our UV-UNet).

For 2D face images (size: 256 pixels $\times$ 256 pixels), the image reading and preprocessing time, including facial landmark detection, is approximately 0.321 seconds. This time, though, is influenced by the size of the input image. The time required for 3D face estimation from a 2D face image (size: 224 pixels $\times$ 224 pixels) is approximately 0.007 seconds. The 3D head

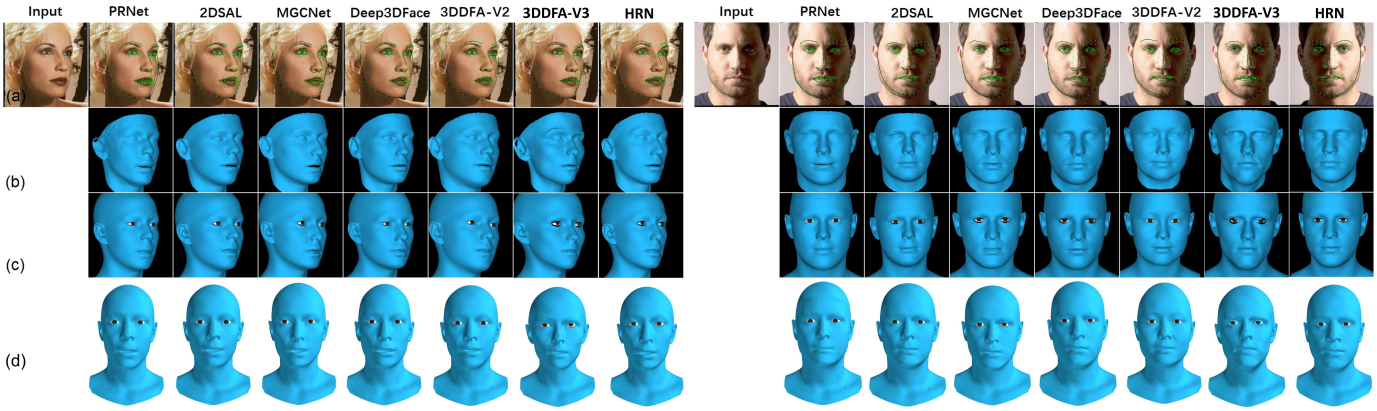


Fig. 12: Our full head prediction from face regions reconstructed using different methods. (a) Face alignment results with 150 projected landmarks from (b) reconstructed facial regions. (c)/(d) Predicted full head.

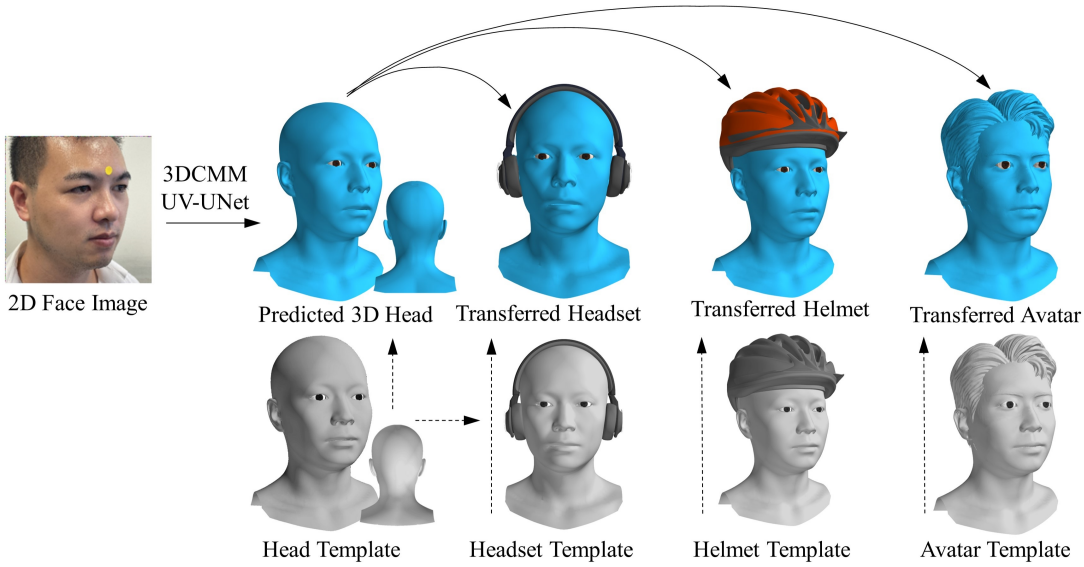


Fig. 13: Diverse application scenarios for our 3DCMM, encompassing headset and helmet customization, and avatar creation. When provided with a headwear product template meticulously crafted by professional designers, our system seamlessly transfers the product from the template head to the accurately predicted specific head.

prediction from the face regions takes around 0.205 seconds. This time includes the steps of projecting the 3D face (with 53,215 points) into a 2D UV position map (size: 256 pixels $\times$ 256 pixels), predicting the 2D head UV map, and retrieving the 3D head mesh (with 56,804 points) from the 2D UV map. In total, the computation time for our proposed method was approximately 0.526 seconds. These runtime measurements provide an overview of the performance of our method on the specified hardware setup and the actual runtime may vary depending on the hardware specifications and the complexity of the input data.

E. Application Scenarios

Compared to existing 3DMMs that primarily focus on the frontal face region, including the forehead [7], [8], [9], [47], the scalp region is often overlooked. Fortunately, our developed UV-UNet-based method demonstrates the capability to predict the full head based on 3D faces generated or synthesized by other 3DMMs [7], [8], [20], [47] (see Fig. 12). This expansion of functionality effectively broadens the applications of previous studies and holds great promise for future tasks related to 3D face/head reconstruction and prediction. This represents a key advantage of our UV-UNet-based head prediction method, even when compared to our previous model-based transformation approach [11]. Notably, the accuracy of the created full-head mesh is influenced by the quality of the reconstructed face from images or scans (see Fig. 12 (d)).

An accurate full-head mesh, encompassing not only the face but also the scalp region, brings about significant advancements in various applications beyond face reconstruction and manipulation. These applications include the creation of realistic avatars [2], [60], the customization of headwear products [4], and the facilitation of headwear virtual try-ons [3], as depicted in Fig. 13. The inclusion of the scalp region in the full-head representation holds particular importance for

ergonomic headwear design, as it directly impacts the shape design of the contact area between the headwear and the scalp. By incorporating the scalp region, precise measurements and simulations can be conducted, resulting in enhanced headwear comfort, fit, and overall design.

V. LIMITATIONS

In this study, there were still limitations that could affect the efficiency of our 3D face reconstruction and 3D head prediction. The first limitation is the impact of hair contamination on 3D head capture and modelling. In our Chinese head database, participants were required to wear a tight, custom-designed latex cap during scanning to mitigate surface distortions caused by hair [31], [32]. However, despite these precautions, compressed hair can still affect the scanned scalp regions, particularly in individuals with longer hair. This is a recognized issue, with previous studies [61] reporting average hair thickness offsets of 3.6 mm for males and 5.8 mm for females in Australian adults, even when participants' hair was compressed by wearing a wig cap. To address this concern in future research, we intend to explore the utilization of 3D head data obtained through Computed Tomography (CT) imaging [62]. CT imaging offers a more detailed representation of internal structures, enabling the removal of hair and cap effects from 3D head meshes. By adopting this approach, we aim to enhance the accuracy and reliability of our methods by eliminating biases introduced by hair and caps during the scanning process. The second limitation is the impact of extreme face poses or expressions on 3D face reconstruction from 2D images. The detected facial landmarks can only cover facial contours, eyes, noses, and mouths. Thus, ensuring accurate reconstruction of the forehead (see Region-I in Fig. 4), which is greatly related to head completion, is challenging. Inaccuracies in the reconstructed face (especially the forehead regions), particularly under extreme poses or expressions, can propagate to the final 3D head model and result in inaccuracies in the predicted scalp regions (see Fig. 12 (a)). During extreme poses, the limited visibility of certain facial regions and the occlusion of landmarks by other facial components can further contribute to inaccuracies in head reconstruction. Under extreme expressions, facial geometry can undergo significant changes, leading to challenges in accurately capturing the shape and details of the head. In the future, we plan to collect more face images with a variety of extreme poses and expressions as a supplementary training dataset to improve the robustness and accuracy of our proposed method.

VI. CONCLUSION

We constructed comprehensive 3D morphable models of human heads by leveraging large-scale real 3D human head data, which included incorporating 3D morphable models of human faces and our UV-UNet-based head prediction model. Additionally, we developed a 3DCMM-based 3D face reconstruction method using self-supervised learning techniques. We maximized the utilization of image-level information and introduced two innovative losses: facial boundary-aware and structure-aware losses. The incorporation of these losses improved the consistency of facial boundaries and structures. Through model comparison experiments, we have demonstrated that our model surpasses existing state-of-the-art 3DMMs in terms of generalization ability and accuracy. Furthermore, thorough evaluation experiments have shown that our head creation method achieves outstanding results in 3D face reconstruction and scalp prediction.

REFERENCES

[1] A. Morales, G. Piella, and F. M. Sukno, "Survey on 3D face reconstruction from uncalibrated images," *Computer Science Review*, vol. 40, pp. 100 400:1–100 400:35, 2021.

[2] L. Bao, X. Lin, Y. Chen, H. Zhang, S. Wang, X. Zhe, D. Kang, H. Huang, X. Jiang, J. Wang, D. Yu, and Z. Zhang, "High-fidelity 3D digital human creation from RGB-D selfies," *ACM Transactions on Graphics*, vol. 41, no. 1, pp. 3:1–3:21, 2021.

[3] J. Zhang, Y. Luximon, P. Shah, and P. Li, "3D statistical head modeling for face/head-related product design: a state-of-the-art review," *Computer-Aided Design*, vol. 159, pp. 103 483:1–103 483:24, 2023.

[4] J. Zhang, Y. Luximon, P. Shah, K. Zhou, and P. Li, "Customize my helmet: A novel algorithmic approach based on 3D head prediction," *Computer-Aided Design*, vol. 150, pp. 103 271:1–103 271:10, 2022.

[5] M. Schaufelberger, C. Kaiser, R. Kühle, A. Wachter, F. Weichel, N. Hagen, F. Ringwald, U. Eisenmann, J. Hoffmann, M. Engel, C. Freudlsperger, and W. Nahm, "3D-2D distance maps conversion enhances classification of craniosynostosis," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 11, pp. 3156–3165, 2023.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[7] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 2018.

[8] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 598–607.

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

[10] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 194:1–194:17, 2017.

[11] J. Zhang, Y. Luximon, L. Zhu, and P. Li, "3DCMM: 3D comprehensive morphable models for accurate head completion," in *Proceedings of the ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, 2023, pp. 13:1–13:8.

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 285–295.

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

[14] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7763–7772.

[15] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3735–3744.

[16] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan, "Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 53–70.

[17] W. Yang, Y. Zhao, B. Yang, and J. Shen, "Learning 3D face reconstruction from the cycle-consistency of dynamic faces," *IEEE Transactions on Multimedia*, pp. 1–14, 2023.

[18] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 88:1–88:13, 2021.

[19] H. Dai, N. Pears, W. Smith, and C. Duncan, "Statistical modeling of craniofacial shape and texture," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 547–571, 2020.

[20] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, 1999, pp. 187–194.

[21] Z. Chai, H. Zhang, J. Ren, D. Kang, Z. Xu, X. Zhe, C. Yuan, and L. Bao, "REALY: Rethinking the evaluation of 3D face reconstruction," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 74–92.

[22] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter, "3D morphable face models–Past, present, and future," *ACM Transactions on Graphics*, vol. 39, no. 5, pp. 157:1–157:38, 2020.

[23] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3D morphable model regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8377–8386.

[24] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition*, 2016, pp. 146–155.

[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 *Proceedings of the European Conference on Computer Vision*, 2018, pp. 534–551.

[26] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 152–168.

[27] Y.-P. Cao, Z.-N. Liu, Z.-F. Kuang, L. Kobbelt, and S.-M. Hu, "Learning to reconstruct high-quality 3D shapes with cascaded fully convolutional networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 616–633.

[28] X. Chai, J. Chen, C. Liang, D. Xu, and C.-W. Lin, "Expression-aware face reconstruction via a dual-stream network," *IEEE Transactions on Multimedia*, vol. 23, pp. 2998–3012, 2021.

[29] S. Ploumpis, E. Ververas, E. O'Sullivan, S. Moschoglou, H. Wang, N. Pears, W. A. P. Smith, B. Gecer, and S. Zafeiriou, "Towards a complete 3D morphable model of the human head," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4142–4160, 2020.

[30] Z.-H. Feng, P. Huber, J. Kittler, P. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. Rätsch, "Evaluation of dense 3D reconstruction from 2D face images in the wild," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 780–786.

[31] J. Zhang, H. Iftikhar, P. Shah, and Y. Luximon, "Age and sex factors integrated 3D statistical models of adults' heads," *International Journal of Industrial Ergonomics*, vol. 90, pp. 103 321:1–103 321:13, 2022.

[32] J. Zhang, F. Fu, Y. Shi, and Y. Luximon, "Modeling 3D geometric growth patterns and variations of children's heads," *Applied Ergonomics*, vol. 108, pp. 103 933:1–103 933:11, 2023.

[33] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

[34] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[35] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

[36] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[38] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, A. Dovzhenko, O. Tietz, C. Dal Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, and O. Ronneberger, "U-Net: deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, pp. 67–70, 2019.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015, pp. 1–15.

[41] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 117–128.

[42] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[44] J. Morton and S. Lee, "Floating-point precision and deformation awareness for scalable and robust 3D face alignment," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 2019, pp. 25:1–25:10.

[45] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 986–993.

[46] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2129–2138.

[47] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Lüthi, S. Schönborn, and T. Vetter, "Morphable face models-an open framework," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 75–82.

[48] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[49] J. Cheng, Y. Li, J. Wang, L. Yu, and S. Wang, "Exploiting effective facial patches for robust gender recognition," *Tsinghua Science and Technology*, vol. 24, no. 3, pp. 333–345, 2019.

[50] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 768–783.

[51] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," in *Proceedings of the IEEE International Conference on Pattern Recognition*, 2018, pp. 1598–1603.

[52] J. Johnson, N. Ravi, J. Reizenstein, D. Novotny, S. Tulsiani, C. Lassner, and S. Branson, "Accelerating 3D deep learning with PyTorch3D," in *Proceedings of the ACM SIGGRAPH Asia Courses*, 2020, pp. 10:1–10:1.

[53] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2D/3D hybrid face dataset," in *Proceedings of the Joint ACM Workshop on Human Gesture and Behavior Understanding*, 2011, pp. 79–80.

[54] 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 Transactions on Multimedia*, vol. 23, pp. 1160–1172, 2021.

[55] Z. Wang, X. Zhu, T. Zhang, B. Wang, and Z. Lei, "3D face reconstruction with the geometric guidance of facial part segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1672–1682.

[56] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 394–403.

[57] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2549–2559.

[58] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt, "Inversefacenet: Deep monocular inverse face rendering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4625–4634.

[59] H. Pottmann, Q.-X. Huang, Y.-L. Yang, and S.-M. Hu, "Geometry and convergence analysis of algorithms for registration of 3D shapes," *International Journal of Computer Vision*, vol. 67, no. 3, pp. 277–296, 2006.

[60] J. Zhang, K. Zhou, Y. Luximon, T.-Y. Lee, and P. Li, "MeshWGAN: Mesh-to-mesh wasserstein GAN with multi-task gradient penalty for 3D facial geometric age transformation," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–14, 2023.
[61] T. Perret-Ellena, S. L. Skals, A. Subic, H. Mustafa, and T. Y. Pang, "3D anthropometric investigation of head and face characteristics of Australian cyclists," *Procedia Engineering*, vol. 112, pp. 98–103, 2015.
[62] Z. Liu, Y. Luximon, W. L. Ng, E. Chung, and J. Zhang, "Anatomical landmark-guided deformation methods for cranial modeling," *Proceedings of the International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences*, pp. 46–52, 2023.

Jie Zhang received the Ph.D. degree from The Hong Kong Polytechnic University, Kowloon, Hong Kong in 2023. He is currently a Lecturer with the Faculty of Applied Sciences, Macao Polytechnic University, Macau, China. He has published over 30 peer-reviewed journal articles (including IEEE TVCG, CAD, IJHCS, IJHCI), 3 textbooks, patents and international conference papers. His current research interests include AI for design and graphics, 3D generative models, color science, data visualization, ergonomic design, human-computer interaction, and 3D digital human modeling/editing.

Kangneng Zhou received the B.Sc. degree in internet of things and the M.Eng. degree in computer science and technology from the University of Science and Technology Beijing, Beijing, China, in 2020 and 2023, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science, Nankai University, Tianjin, China. His current research interests include generative models, 3D head synthesis, computer graphics, and deep learning.

Yan Luximon received the Ph.D. degree in ergonomics from The Hong Kong University of Science and Technology, Hong Kong, in 2006. She is currently an Associate Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. She is also the Leader for Asian Ergonomics Design Lab and the Deputy Discipline Leader for B.A. Product Design. She has published over 100 peer-reviewed journal articles, book chapters, patents and international conference papers. Her current research interests include computer graphics, 3D digital human modeling and CAD, AI in design and visualization, 3D head and face reconstruction, deep learning, and virtual reality.

Tong-Yee Lee (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Washington State University, Pullman, in 1995. He is currently a Chair Professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University (NCKU), Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, NCKU (http://graphics.csie.ncku.edu.tw). His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a Senior Member of the IEEE and a Member of the ACM. He is an Associate Editor of the *IEEE Transactions on Visualization and Computer Graphics*.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 200 top-tier scholarly research articles (e.g., TMM, TPAMI, TVCG, TIP, TNNLS, TMI, TCSVT, TCYB, TBME, TSMC, TII, AAAI, CVPR, ICCV, NeurIPS, Nature Metabolism), pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, computational art, and creative media.