

A Hierarchical Representation Network for Accurate and Detailed Face Reconstruction from In-The-Wild Images

Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, Xuansong Xie
DAMO Academy, Alibaba Group

{biwen.lbw, jianqiang.rjq, mengyang.fmy, miaomiao.cmm}@alibaba-inc.com,
xingtong.xxs@taobao.com

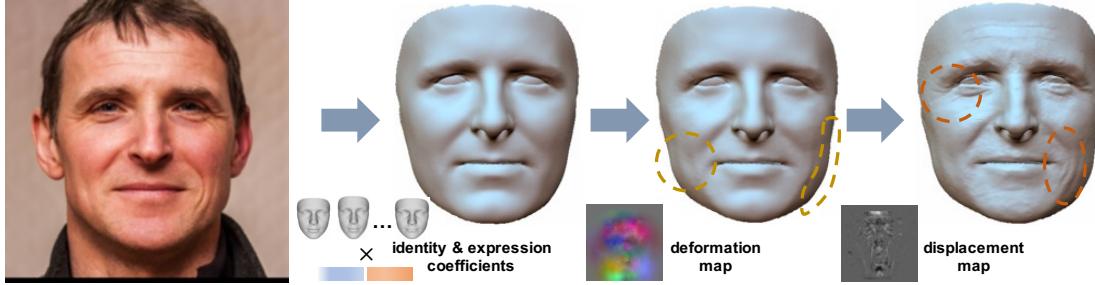


Figure 1. Example of low-frequency geometry, MF, HF facial details and the hierarchical representation.

Abstract

Limited by the nature of the low-dimensional representational capacity of 3DMM, most of the 3DMM-based face reconstruction (FR) methods fail to recover high-frequency facial details, such as wrinkles, dimples, etc. Some attempt to solve the problem by introducing detail maps or non-linear operations, however, the results are still not vivid. To this end, we in this paper present a novel hierarchical representation network (HRN) to achieve accurate and detailed face reconstruction from a single image. Specifically, we implement the geometry disentanglement and introduce the hierarchical representation to fulfill detailed face modeling. Meanwhile, 3D priors of facial details are incorporated to enhance the accuracy and authenticity of the reconstruction results. We also propose a de-retouching module to achieve better decoupling of the geometry and appearance. It is noteworthy that our framework can be extended to a multi-view fashion by considering detail consistency of different views. Extensive experiments on two single-view and two multi-view FR benchmarks demonstrate that our method outperforms the existing methods in both reconstruction accuracy and visual effects. Finally, we introduce a high-quality 3D face dataset FaceHD-100 to boost the research of high-fidelity face reconstruction. The project homepage is at <https://younglbw.github.io/HRN-homepage/>.

1. Introduction

High-fidelity 3D face reconstruction finds a wide range of applications in many scenarios, such as AR/VR, medical

treatment, film production, etc. While extensive works already achieved excellent reconstruction performance using specialized hardware like LightStage [2, 11, 35], estimating highly detailed face models from single or sparse-view images is still a challenging problem. Based on 3DMM [8], a statistical model learned from a collection of face scans, many works [16, 22, 23, 32] attempt to reconstruct the 3D face from a single image and achieve impressive results. However, limited by the nature of the low dimensional representational ability of the 3DMM, these methods can not recover the detailed facial geometry.

Recently, some methods [13, 24, 38] devote to capturing high-frequency facial details such as wrinkles by predicting a displacement map. They achieve realistic results, however, fail to model the mid-frequency details, such as the detailed contour of the jaw, cheek, etc. To this end, some works try to capture the overall details by introducing latent encoding of details [19] or non-linear operations [20, 44]. Nevertheless, it is hard to make a trade-off when handling the mid- and high-frequency details simultaneously. Another challenge is how to obtain accurate shapes and detailed 3D facial priors considering multifarious lightings and skins for different images. [10, 13] resort to the wrinkle statistics computed from 3D face scans to fulfill realistic high-frequency details, but still fail to model the mid-frequency details.

Based on the observations above, we introduce a hierarchical representation network (HRN) for accurate and detailed face reconstruction from single image, as shown in

Fig. 2. Firstly, we decouple the facial geometry into low-frequency geometry, mid-frequency (MF) details, and high-frequency (HF) details. Then, in a hierarchical fashion, we model these parts with face-wise blendshape coefficients, vertex-wise deformation map, and pixel-wise displacement map, respectively (shown in Fig. 1). Concretely, we employ two image translation networks [27] to estimate the corresponding detail maps (deformation and displacement map), and further employ them to generate the detailed face model in a coarse-to-fine manner. Moreover, we introduce the 3D priors of MF and HF details by fitting face scans with our hierarchical representation to facilitate accurate and faithful modeling. Inspired by [33], we propose a de-retouching module to adaptively refine the base texture to overcome the ambiguities between skin blemishes and illuminations. Extensive experiments show that our method outperforms the existing methods on two large-scale benchmarks, exhibiting excellent performance in terms of detail capturing and accurate shape modeling. Thanks to the detail disentanglement strategy and the guidance of detail priors, we extend HRN to a multi-view fashion and achieve accurate FR from only a few views. Finally, to boost the research of sparse-view and high-fidelity FR, we introduce a high-quality 3D face dataset named FaceHD-100.

Our main contributions in this work are as follows:

- (A) We present a hierarchical modeling strategy and propose a novel framework HRN for single-view FR task. Our HRN produces accurate and highly detailed FR results and outperforms the existing state-of-the-art methods on two large-scale single-view FR benchmarks.
- (B) We introduce detail priors to guide the faithful modeling of hierarchical details and design a de-retouching module to facilitate the decoupling of geometry and appearance.
- (C) We extend HRN to a multi-view fashion to form MV-HRN, which enables accurate face modeling from sparse-view images and outperforms the existing methods on two large-scale multi-view FR benchmarks.
- (D) To boost the research on sparse-view and high-fidelity FR tasks, we introduce a high-quality 3D face dataset FaceHD-100, containing 2,000 detailed 3D face models and corresponding high-definition multi-view images.

2. Related Work

Single-View Face Reconstruction. Recovering 3D face from a single image is an ill-posed problem, but the advent of the 3D morphable model (3DMM) [8, 39] has made it possible. The 3DMM provides strong prior knowledge and can represent complicated face geometry with low-dimensional coefficients. In this formulation, current methods can be categorized into either optimization-based [7, 8, 25, 31, 51] or learning-based [15, 40, 43, 52]. Optimization-based approaches usually need costly analysis-by-synthesis processes and are sensitive to initialization, while learning-

based methods directly train neural networks to regress the low-dimensional coefficients of 3DMM and recover 3D face through efficient forward inference.

However, the original 3DMM models inherently lie in low-dimensional linear space and lack fine details. Many works [14, 19, 42, 45, 48] are proposed to overcome this limitation. Tran *et al.* [45] present a nonlinear 3DMM model and achieve more powerful representational abilities. Sela *et al.* [42] employ image-to-image network to generate pixel-based geometric representation for high quality reconstructions. In addition to static face geometry details, Feng *et al.* [19] present an animatable displacement model to generate dynamic expression-depended wrinkles. Yang *et al.* [48] predict displacement maps via pix2pixHD network and combine them according to blendshape weights for dynamic details synthesis. Compared with these approaches, our method further introduces facial detail priors and can recover high fidelity facial details with hierarchical geometry representations.

Multi-View Face Reconstruction. Traditional multi-view stereo (MVS) methods [6, 9, 21] are designed for 3D reconstruction given a set of multi-view images, but they heavily rely on the precision of camera calibration, and can hardly recover intact geometry in the sparse-view situation. To address these problems, many face-specialized multi-view reconstruction methods [5, 17, 26, 36, 37, 46, 47] are proposed. Ramon *et al.* [37] introduce siamese neural networks to extract relevant features from each view, and learn the 3D shape and the individual camera poses simultaneously. Wu *et al.* [46] exploit both 3DMM and multi-view geometric constraints by estimating the alignment loss between multi-view inputs. Bai *et al.* [5] leverage non-rigid multi-view stereo optimization to explicitly enforce multi-view appearance consistency, which is able to capture medium-level details. With the emergence of implicit 3D representation, Xiao *et al.* [47] propose to learn an implicit function to recover detailed geometry from calibrated multi-view images, but the implicit function learning is time-consuming which needs dozen of seconds and is sensitive to camera count and pose estimation.

Rather than a specific design for multi-view inputs, our single-view model can be easily transferred to sparse-view face reconstruction task by adding hierarchical detail consistency between different views. Our method is robust to the calibration error of cameras thanks to the coarse-to-fine learning scheme.

3. Methods

3.1. Overview

In this paper, we propose a novel hierarchical representation network for accurate and detailed face reconstruction from single image. Fig. 2 illustrates the overview of our

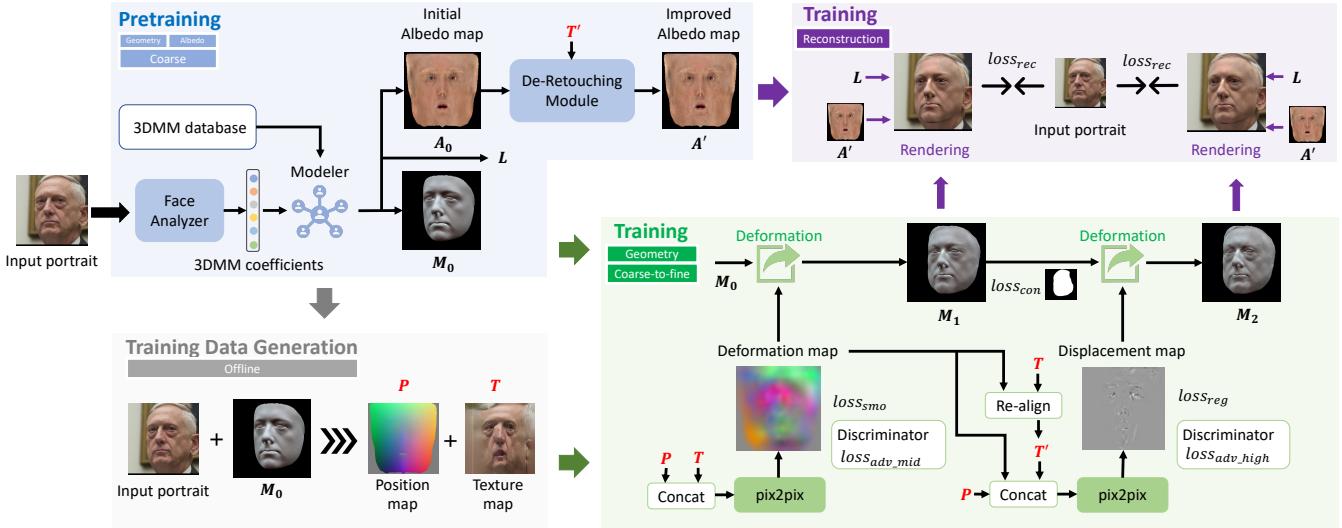


Figure 2. Overview of the proposed hierarchical representation network (HRN).

framework. We first employ 3DMM to predict a coarse mesh and albedo (blue area in Fig. 2). Then we develop a hierarchical modeling (Sec. 3.2) strategy to handle the complex facial details in a coarse-to-fine manner (green and purple area). To facilitate the accurate and faithful modeling of the hierarchical details, the 3D priors are incorporated through adversarial and semi-supervised learning (Sec. 3.3). Besides, we propose a de-retouching module (Sec. 3.4) to fulfill better decoupling of the geometry and appearance, alleviating the ambiguities between the various skin texture and illuminations. Moreover, we extend our framework to a multi-view fashion (Sec. 3.5) and introduce a high quality 3D face dataset (Sec. 3.6) to boost the research on sparse-view face reconstruction. For simplification, we specify the related loss functions and training strategy in each section.

3.2. Hierarchical Modeling

3DMM exhibit great performance in expressing various shape of the face, while the low-dimensional representation itself severely stem its learning on the details, leading to the imperfect alignment to the face or the over-smoothed results. Some methods extend 3DMM by introducing displacement maps to reconstruct some details such as wrinkles and bumps. However, focused on the high-frequency part, a simple displacement map still fails to handle some larger-scale details, such as the contours of the jaw and cheek. Based on such observation, we decouple the facial geometry into three components: (1) low-frequency part, which provides the coarse shape roughly aligned to the input face; (2) mid-frequency details, which describe the details of the contour and local shape relative to the low-frequency part; (3) and high-frequency details, such as wrinkles, micro bumps, etc. As shown in the Fig. 1, the scales decrease from the low-frequency part to the high-

frequency detail, while the fineness increases in turn.

We design the hierarchical representation to model the above three components respectively. For the low-frequency part, we adopt the BFM as our base model and output the low-dimensional coefficients to fulfill a coarse reconstruction of the input face. Then we introduce a three-channel deformation map, which lies in the UV space and indicates the offset of each vertex relative to the coarse result. Worked as the representation of the mid-frequency details, the deformation map provides a flexible way to manipulate the geometries. We use the size of 64×64 to represent the deformation map to balance the fineness and smoothness of the mid-frequency details. For the high-frequency details, we employ the displacement map following [19], which is a one-channel map (256×256) denoting the geometry deformation along the direction of the normals. The displacement map is converted to detailed normals used in the rendering process in a pixel-wise manner to exhibit all the tiny details, breaking the limitation of the vertex density of the base model. Accordingly, we are enabled to describe an arbitrary complex face with these representations.

As shown in Fig. 2, given a portrait image I , we firstly utilize a regression network as the face analyzer to predict the BFM coefficients and obtain the coarse aligned mesh M_0 and albedo A_0 using the corresponding basis from the 3DMM database. Combined I and M_0 , we are able to acquire the inpainted texture T in UV space by applying the differentiable rendering with a coarse-to-fine strategy. And we concatenate the position map P and T as the input of the following modules for hierarchical details learning. We adopt two pix2pix [27] networks to synthesize the deformation map and displacement map in sequence. Note that, considering the deformation map will change the facial geometry and lead to the misalignment between T and the

deformed mesh, we generate the realigned texture \mathbf{T}' as the input of second pix2pix network by projecting the three-channel deformation map to the 2D space and transform it to a reversal flow \mathbf{F} to re-align \mathbf{T} . Taking advantage of the abundant details from \mathbf{T} and the pixel-wise learning strategy, we manage to obtain the accurate detail maps which are further employed in a coarse-to-fine manner to generate the detailed face mesh \mathbf{M}_1 and \mathbf{M}_2 . Finally, combined with the lighting \mathbf{L} and the refined albedo generated from the de-retouching module (Sec. 3.4), we accomplish detailed face reconstruction from the single image.

Overall, the framework is trained in a self-supervised manner guided with 3D detail priors learned from face scans (Sec. 3.3). To reduce the training complexity, we adopt the pre-trained encoder and MLP from the [15] as the face analyzer to predict the coefficients and generate the corresponding \mathbf{P} and \mathbf{T} for the following details learning. The two image translation networks are trained jointly and the related loss functions are composed of three components:

Reconstruction Loss. The reconstruction loss is calculated between the rendered face and the input face and consists of the photometric loss L_{photo} , perception-level loss L_{per} and landmark loss L_{lan} following [15]. Thanks to the delighted albedo and the illumination system of 3DMM, the photometric loss will enforce the deformation of the facial geometry to fit the various shadows and highlight areas of the input face. It is crucial that we apply the reconstruction loss on both images rendered from the \mathbf{M}_1 and \mathbf{M}_2 , which benefit the disentanglement of the MF and HF details.

Details Loss. We apply the total variation loss L_{tv} [29] to encourage the smoothness of the deformation map, and use the L1 regularization loss L_{reg} to limit the scale of the displacement map.

Contour-aware Loss. We propose a novel contour-aware loss L_{con} to fulfill accurate modeling of the face contour. The L_{con} works on \mathbf{M}_1 and aims to pull the vertices of edge to align the face contour. As shown in Fig. 3, we firstly project vertices of \mathbf{M}_1 into the image space. Then we predict the face mask \mathbf{M}_{face} using the pre-trained face matting network [34] and implement post process to obtain the left side and right side points for each row. Given a vertex p and the corresponding projected point p' on \mathbf{M}_{face} , we obtain the vector \mathbf{l}_p and \mathbf{r}_p (from p' to the edge points in the horizontal direction). Then L_{con} can be describe as:

$$L_{con} = \frac{1}{N_p} \sum_{p \in M_1} (f(p) \mathbb{1}[y(p') > \delta]), \quad (1)$$

$$f(p) = |h(\frac{\mathbf{l}_p \cdot \mathbf{r}_p}{\max(\|\mathbf{l}_p\|, \|\mathbf{r}_p\|)} + \lambda) - \lambda|, \quad (2)$$

where h is the ReLU function, and λ denotes a soft margin relative to the face contour ($\lambda = 0.01$ as default), $\mathbb{1}[y(p') > \delta]$ indicates whether p' is on the lower part of the image ($\delta =$

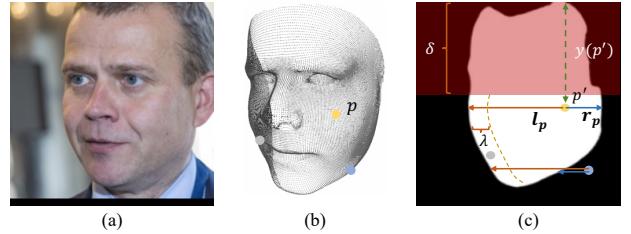


Figure 3. The details of the proposed contour loss. (a) input image, (b) the projected vertices, (c) the predicted face mask.

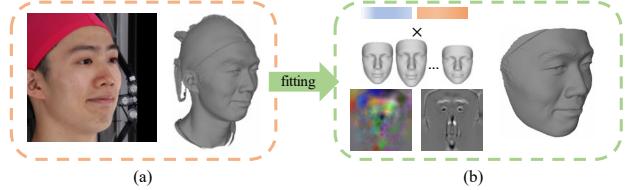


Figure 4. An example of the 3D priors of facial details. (a) the raw image and scan, (b) the hierarchical representation and the corresponding fitted mesh.

100 as default). As we can see, L_{con} punish the vertices outside the soft margin (such as the blue and gray points in Fig. 3) of the face and pull them to the face contour, while keeping the vertices inside the face intact. Combined with L_{tv} of the deformation map, L_{con} will avoid the unsmooth effect near the face contour. Note that we only focus on the lower part of the face contour to avoid the distraction of the hair. Compared to the common segmentation loss, L_{con} gives a more straightforward direction for optimizing the face contour and is easier for training. We conduct an ablation study to reveal the effectiveness of L_{con} in Sec. 4.4.

3.3. 3D Priors of Facial Details

Although facial details can be roughly learned from single image using the reconstruction loss (Sec. 3.2), it suffers from unreality and ambiguousness due to its ill-posed essence. Adding additional regularization may help to narrow the solution space, but also lead to severe degradation in detail accuracy and fidelity.

To address this problem, we exploit the 3D priors of facial details derived from face scans and corresponding multi-view images in our framework. Firstly, given an raw image and its corresponding scan, we transform the raw scan to align to the image in BFM space (the details can be found in the supplemental files). Then we can obtain the ground-truth deformation map and displacement map for each image by fitting the image and scan using the loss functions mentioned in Sec. 3.2 with additional supervision on vertices distance following [3]. Thanks to the powerful hierarchical representation, the details of scans can be accurately captured. See Fig. 4 for example.

We take advantage of the 3D priors of details on two

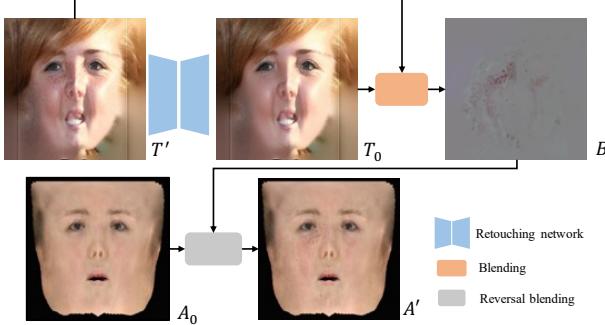


Figure 5. The details of the de-retouching module.

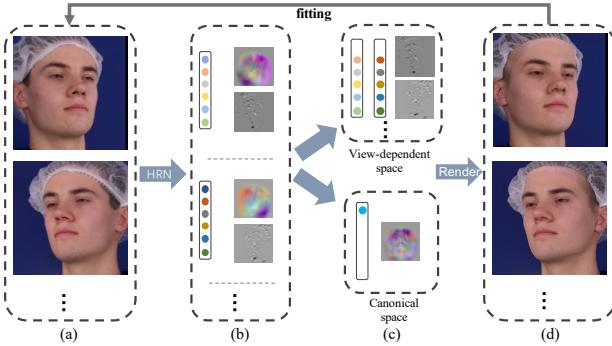


Figure 6. The pipeline of the proposed MV-HRN.

aspects. On the one hand, we develop two discriminators and use the adversarial loss [28] $L_{adv,mid}$ and $L_{adv,high}$ to supervise the domain distribution of the deformation map and displacement map. On the other hand, we acquire the paired data as mentioned above from 3D scans to conduct supervised learning to guide the self-supervised learning in Sec. 3.2. Specifically, we supplement the L1 loss L_{mid} and L_{high} for the predicted deformation map and displacement map respectively. Note that a mask is used in training to remove the distraction of eyes and hair area from scans.

3.4. De-Retouching Module

A face image is the result of a combination of geometry, lighting, and face albedo. Prior works assume that the face albedo is smooth and model it with the low-frequency albedo from 3DMM. However, the actual skin texture is full of high-frequency details such as moles, scars, freckles, and other blemishes, which bring ambiguities to the geometry details learning especially in the single view FR task. Inspired by the [33], we propose a de-retouching module (DRM), which aims to generate the face albedo with high-frequency details and facilitate more precise decoupling of geometry and appearance.

We collect 10, 000 face images from FFHQ [30], and hire a team of professional image editors to process the images, with the goal of removing the skin blemishes and other texture details while maintaining the shape-related content such as wrinkles, bumps, etc. Then we transform the paired

images into the texture maps in UV space by applying the process specified in Sec. 3.2 and train an image translation network G to fulfill skin retouching. Given the re-aligned texture T' , we firstly employ G to remove its texture details and get T_0 , as shown in Fig. 5. We aim to bake the texture details into the coarse albedo A_0 to obtain the improved albedo A' for rendering. We make an assumption that the shading from A_0 to T_0 should be consistent with the one from A' to T' , as:

$$T_0 = A_0 \odot S, \quad (3)$$

$$T' = A' \odot S, \quad (4)$$

where S denotes the shading map, \odot denotes element-wise matrix multiplication. Then we can solve the equations and obtain A' as:

$$A' = A_0 \odot B \approx A_0 \odot \frac{T' + \phi(T_0)}{T_0 + \phi(T_0)}, \quad (5)$$

$$\phi(T_0) = \frac{1}{T_0 \odot T_0 \odot T_0 + \varepsilon}, \quad (6)$$

where $\phi(T_0)$ avoids the value explosion near 0 and $\varepsilon = 1e-6$ as default. Compared to A_0 , the de-retouched albedo A' contains more high-frequency texture details, which alleviate the ambiguities between geometry and appearance, especially in single view FR task.

3.5. MV-HRN

Thanks to the hierarchical modeling and the 3D priors guidance, we can easily adapt the HRN to a multi-view fashion to fulfill precise modeling of the global facial geometry with only a few views by adding the geometry consistency between different views. Fig. 6 shows the pipeline of MV-HRN. We assume that the low-frequency identity part and the mid-frequency details are consistent between different views, while the lighting, expression, and HF details should be view-dependent to overcome the disturbance. Therefore, we develop a canonical space, which contains the shared identity coefficient and deformation map that are initialized by averaging all the single-view results, to represent the shared intrinsic face shape. Then other BFM coefficients and the displacement map of each view are utilized to dependently model the pose, lighting, expression and high-frequency details. Then we apply the loss functions mentioned in Sec. 3.2 and Sec. 3.3 to iteratively optimize all the coefficients and detail maps. Through the fitting process, the face shape is gradually restricted to a smaller and more accurate space under the supervision of different views. Extensive experiments show that MV-HRN achieves accurate reconstruction given only a few ($2 \sim 5$) views of images in a short time (less than one minute).

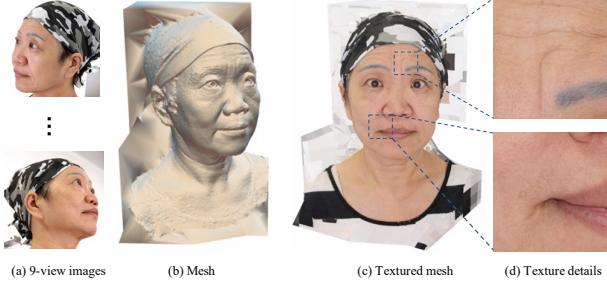


Figure 7. An example from the FaceHD-100 dataset.

3.6. FaceHD-100 Dataset

To boost the research of face reconstruction from sparse-view images, we introduce a high-quality 3D face dataset FaceHD-100, which consists of 2,000 high-definition 3D mesh and corresponding multi-view images from 100 subjects. The data is captured by a multi-view 3D reconstruction system, which is composed of 9 DSLR cameras and 8 LED lights. The 9 cameras are evenly distributed in front of and to the side of the face, and each provides 8K images for geometry and texture reconstruction. The capturing subjects include 50 males and 50 females, and mostly are from Asia. The ages of these subjects are normally distributed from 16 to 70 years old. For each person, we follow [1] and ask them to perform 20 expressions including the neutral expression for capturing. Fig. 7 gives an example of FaceHD-100, which shows the high quality of the reconstructed geometry and texture.

4. Experiments

4.1. Implementation Details

Training Data. The training data of the proposed model is composed of two parts: 2D in-the-wild images and 3D face scans with corresponding multi-view images. For the former, we collect in-the-wild face images from multiple sources following [15]. For the latter, the data is collected from FaceScape [50], ESRC [18] and FaceHD-100. To be specific, we split 359 samples of FaceScape into training (309 subjects) and testing sets (50 subjects), considering the balance of gender and age. The ESRC is also split in the same way as [5] and the whole FaceHD-100 dataset is used for training. In total, we collected $\sim 9K$ scans from nearly 500 subjects of different ethnicities. The majority of subjects have 3D scans for at least 8 different expressions. Then we process all the scans and corresponding multi-view images in the way shown in Fig. 4 to generate the ground-truth deformation maps and displacement maps for each image. In the end, we collected $\sim 260K$ in-the-wild images for self-supervised training and $\sim 150K$ lab images with corresponding ground-truth details map for supervised training. The input images are preprocessed following [15].

Training Strategy. Firstly, we employ the pretrained R-



Figure 8. The single-view qualitative comparison on FFHQ (first three rows), REALY (fourth row), and FaceScape (last row).

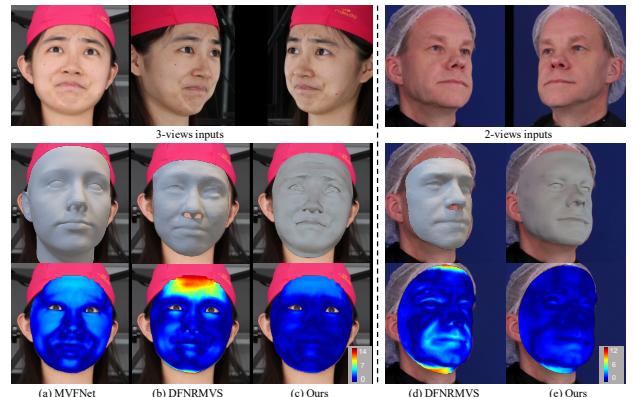


Figure 9. The multi-view qualitative comparison on FaceScape and ESRC datasets.

Net from [15] as our face analyzer to predict the 3DMM coefficients, position map and texture map as mentioned in Sec. 3.2 for the following training. We use the paired texture maps mentioned in Sec. 3.4 to train the de-retouching module. Finally, we fix the parameters of the face analyzer and de-retouching module and train the whole network with the input of face images, position maps and texture maps. We train our model using the Adam optimizer with a batch size of 4, an initial learning rate of 1e-4, and 800K iterations. Note that the model is trained alternately with in-the-wild images and lab images in a self-supervised and supervised manner respectively. More details about the parameters and training setting are specified in supplementary materials.

4.2. Qualitative Comparison

We evaluate our model with four SOTA methods: SADRNet [40], Deep3D [15], Pix2Vertex [42], and DECA [19] for single-view 3D FR task (we present comparisons with more methods [13, 49] in the supplementary material).

SADRNet tackles 3D dense face alignment and face reconstruction simultaneously with a self-aligned dual regression framework, and Deep3D leverages hybrid loss function to train CNN for 3DMM coefficients regression in a weakly-supervised manner. With the goal of adding more expressiveness and details, Pix2Vertex utilizes the Image-to-Image translation network to provide high-quality reconstructions under extreme expressions, while DECA presents an animatable displacement model to generate dynamic expression-depended face details. To make a fair comparison, we use their publicly released models and codes, and conduct experiments on FFHQ, REALY, and Facescape datasets.

The comparison results of the single-view reconstruction scenario are shown in Fig. 8. Since Deep3D only reconstructs faces using predicted 3DMM coefficients, the results are in low-dimensional space and lack high-frequency details. Although SADRNet tries to regress face shape deformation separately, the high-frequency details are still ignored in learning because of their minority. Pix2Vertex and DECA provide more fine details, such as blemishes and wrinkles. However, Pix2Vertex brings artifacts at the same time due to its unrestricted manner, and DECA cannot accurately recover face identities and expressions, resulting in similar wrinkles on the forehead among various faces. By contrast, our proposed method can produce high-fidelity 3D faces with expressive details, which are extremely consistent with the original input images.

In the multi-view scenario, we test the performance of DFNRMVS [5], MVFNet [46] and the proposed model given 3-view or 2-view images from FaceScape and ESRC datasets. As illustrated in Fig. 9, our model outperforms the other two methods in terms of fidelity, details and geometry accuracy, proving that our framework can be well generalized to both single-view and multi-view tasks.

4.3. Quantitative Comparison

Three public datasets are employed to quantitatively evaluate our method with several state-of-the-art approaches. Specifically, we choose FaceScape [50] dataset for both single-view and multi-view evaluation, and additionally use REALY [12] and ESRC [18] datasets for single-view and multi-view respectively. To evaluate the geometry accuracy of single-view face reconstruction, we use Chamfer Distance (CD) and Mean Normal Error (MNE) on the FaceScape dataset following FaceScape benchmark [50], while leveraging average Normalized Mean Square Error (NMSE) of different face regions on REALY dataset, which applies region-wise alignment and is more accurate for shape error computing.

Table 1 shows the quantitative comparison of single-view reconstruction. Our approach outperforms other methods on FaceScape-wild and REALY datasets, and achieves

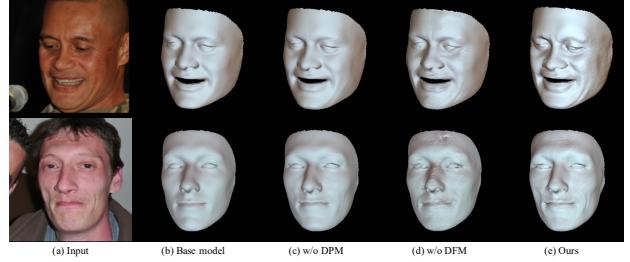


Figure 10. Ablation study toward hierarchical modeling on FFHQ.

SOTA performance on the FaceScape-lab dataset. We find that our side-view metric is even better than the frontal-view on REALY dataset, we speculate that the slightly side-view images provide more information about mouth/node heights, which is beneficial to geometry prediction.

For multi-view reconstruction evaluation, as guided by NoW benchmark [41], we firstly choose 7 face landmarks for each predicted mesh and then apply rigid alignment to ground truth mesh, and report their scan-to-mesh distances. Due to that MVFNet public model cannot handle 2-views images, we only use it in FaceScape dataset testing. The results in Table 2 show that our approach performs better against MVFNet and DFNRMVS with the lowest reconstruction errors. We also test the performance of our method on the MICC dataset [4], and the results are presented in supplementary materials.

Table 1. Single-view quantitative comparison. REALY-F and REALY-S denote frontal-view and side-view reconstruction on REALY benchmark respectively.

Methods	FaceScape-wild		FaceScape-lab		REALY-F NMSE (mm)	REALY-S NMSE (mm)
	CD (mm)	MNE (rad)	CD (mm)	MNE (rad)		
Deep3D	3.8	0.092	5.28	0.118	1.657	1.691
MCGNet	3.22	0.077	4.00	0.093	1.774	1.787
PRNet	3.47	0.123	3.56	0.126	2.013	2.032
SADRNet	7.12	0.123	6.75	0.133	1.913	1.958
DECA	3.31	0.089	4.69	0.108	2.210	2.261
3DDFA-V2	3.00	0.080	3.60	0.096	1.926	1.943
Ours	2.91	0.065	3.67	0.087	1.537	1.468

Table 2. Multi-view quantitative comparison. We only report MVFNet performance on FaceScape because its released model cannot process two-view inputs.

Methods	FaceScape (3 views)			ESRC (2 views)		
	Median (mm)	Mean (mm)	Std (mm)	Median (mm)	Mean (mm)	Std (mm)
MVFNet	1.76	2.12	1.66	N.A.	N.A.	N.A.
DFNRMVS	1.79	2.41	2.61	1.59	2.13	2.29
Ours	1.13	1.51	1.79	1.29	1.69	1.72

4.4. Ablation Study

In order to verify the rationality and effectiveness of our network design, we conduct extensive ablation experiments on FFHQ and FaceScape. Table 3 shows the quan-

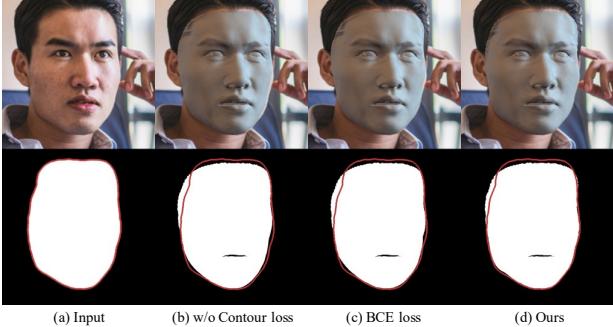


Figure 11. Ablation study toward contour loss on FFHQ.

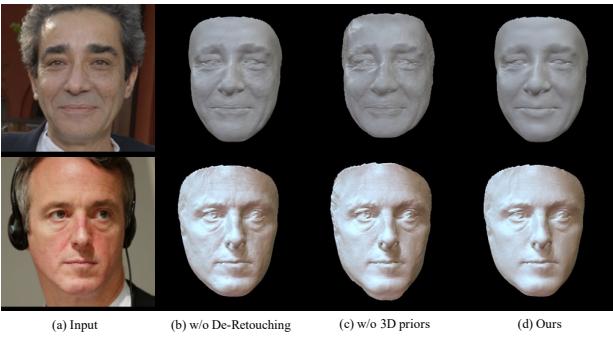


Figure 12. Ablation study toward de-retouching module and 3D detail priors on FFHQ.

Base model	HRM	L_{con}	detail priors	DRM	CD(mm)	MNE(rad)
✓					3.8	0.092
✓	✓				3.34	0.081
✓	✓	✓			3.18	0.079
✓	✓	✓	✓		2.9	0.067
✓	✓	✓	✓	✓	2.91	0.065

Table 3. Quantitative ablation experiments on FaceScape-wild. HRM denotes the hierarchical representation modeling.

Methods	2 views	3 views	4 views	5 views	naive version (5 views)
Ours	1.17	1.13	1.11	1.10	1.23

Table 4. Quantitative ablation study toward sparse-view reconstruction on FaceScape. Only median distance(mm) is reported.

titative results on FaceScape-wild benchmark. As revealed in the table, compared to the base 3DMM, the hierarchical modeling strategy brings a huge improvement ($\sim 0.5\text{mm}$). The contour loss produces 0.16mm improvement. 3D priors of details play a key role in our framework, achieving $\sim 0.3\text{mm}$ improvement. The quantitative contribution of the de-retouching module is minor, while the following qualitative results on FFHQ prove its effectiveness.

On hierarchical modeling. To demonstrate the necessity of the hierarchical modeling, we employ the deformation map (DFM) and displacement map (DPM) respectively (columns c, d in Fig. 10) to solely learn the overall facial details and compare the results. Without DPM, the DFM exhibit the capability of capturing high-frequency details

to a certain extent. However, the performance is limited by the mesh density and the trade-off between the MF details and HF details. In contrast, the DPM is more effective in learning some micro details, but it fails to handle some larger-scale deformation. Apparently, by introducing the hierarchical modeling strategy, the proposed method achieves more accurate and detailed reconstruction.

On contour-aware loss. The contour-aware loss L_{con} aims to enhance the reconstruction accuracy of the facial contours. As shown in Fig. 11, the network with L_{con} exhibit superior performance on learning the face contour compared with the one without L_{con} or with BCE loss.

On 3D priors of facial details. The introduced 3D priors provide the geometry distribution of the real facial details, which guide the model to achieve accurate reconstructions. As shown in Fig. 12, the network without 3D priors produces some unrealistic details. In contrast, the reconstruction results with 3D priors are smoother and more faithful.

On de-retouching module. The de-retouching module is proposed to achieve better decoupling of the facial appearance and geometry. As shown in Fig. 12, without DRM, the model is more susceptible to the distraction of various skin textures and yields some nonexistent details.

On sparse-view reconstruction. We test the performance of MV-HRN on FaceScape given different numbers of input views. As Shown in Table 4, MV-HRN achieves comparable results given only two views for input. With the increase of the number of views, the performance is gradually better, showing the ability of MV-HRN to aggregate multi-view information. Besides, we compare MV-HRN with its naive version (simply average the results of all views) and the results demonstrate the effectiveness of the network design.

4.5. Extention work and Limitation

Due to the limited paper space, more extension work (such as high-quality head reconstruction) and limitation of our method are provided in the supplementary materials.

5. Conclusion

In this paper, we propose a novel hierarchical representation network(HRN) for accurate and detailed face reconstruction from in-the-wild images. Specifically, we achieve facial geometry disentanglement and modeling by hierarchical representation learning. The 3D priors of details are further incorporated to improve the reconstruction results in accuracy and visual effects. Besides, we propose a de-retouching network, which alleviates the ambiguities between geometry and appearance. Moreover, we extend HRN to a multi-view fashion and introduce a high-quality 3D face dataset FaceHD-100 to boost the research of sparse-view FR. Extensive experiments reveal that our method achieves superior performance to the existing methods in terms of accuracy and visual effects.

References

- [1] triplegangers. <https://triplegangers.com/>. 6
- [2] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. The digital emily project: photoreal facial modeling and animation. *ACM*, 2009. 1
- [3] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 4
- [4] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80, 2011. 7
- [5] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5850–5860, 2020. 2, 6, 7
- [6] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9, 2010. 2
- [7] Volker Blanz, Albert Mehl, Thomas Vetter, and H-P Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 293–300. IEEE, 2004. 2
- [8] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 1, 2
- [9] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 papers*, pages 1–10. 2010. 2
- [10] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, 34(4):46:1–46:9, 2015. 1
- [11] Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiyong Li, and Jingyi Yu. Sparse photometric 3d face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4635–4644, 2018. 1
- [12] Zhenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 7
- [13] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019. 1, 6
- [14] Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020. 2
- [15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4, 6
- [16] A. Dib, C. Thebault, J. Ahn, P. H. Gosselin, and L. Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. 2021. 1
- [17] Pengfei Dou and Ioannis A Kakadiaris. Multi-view 3d face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80:80–91, 2018. 2
- [18] ESRC. Esrc. <http://pics.stir.ac.uk/ESRC/>, 2022. 6, 7
- [19] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 1, 2, 3, 6
- [20] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018. 1
- [21] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2
- [22] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1
- [23] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (9):44, 2022. 1
- [24] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. 1
- [25] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017. 2
- [26] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 2
- [27] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3
- [28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5
- [29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

- [30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5
- [31] Ira Kemelmacher-Shlizerman and Steven M Seitz. Face reconstruction in the wild. In *2011 international conference on computer vision*, pages 1746–1753. IEEE, 2011. 2
- [32] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction “in-the-wild”. In *IEEE*, 2020. 1
- [33] Biwen Lei, Xiefan Guo, Hongyu Yang, Miaomiao Cui, Xuansong Xie, and Di Huang. Abpn: Adaptive blend pyramid network for real-time local retouching of ultra high-resolution photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2108–2117, 2022. 2, 5
- [34] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 4
- [35] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*, 2007(9):10, 2007. 1
- [36] Marcel Piotraschke and Volker Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3418–3427, 2016. 2
- [37] Eduard Ramon, Janna Escur, and Xavier Giro-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2
- [38] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1
- [39] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 2, pages 986–993. IEEE, 2005. 2
- [40] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Transactions on Image Processing*, 30:5793–5806, 2021. 2, 6
- [41] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7
- [42] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 2, 6
- [43] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 2
- [44] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model, 2019. 1
- [45] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2
- [46] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngai Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019. 2, 7
- [47] Yunze Xiao, Hao Zhu, Haotian Yang, Zhengyu Diao, Xiangju Lu, and Xun Cao. Detailed facial geometry recovery from multi-view images by learning an implicit function. *arXiv preprint arXiv:2201.01016*, 2022. 2
- [48] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 2
- [49] Zhenyu Zhang, Yanhao Ge, Renwang Chen, Ying Tai, Yan Yan, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning to aggregate and personalize 3d face from in-the-wild photo collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14224, 2021. 6
- [50] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *arXiv preprint arXiv:2111.01082*, 2021. 6, 7
- [51] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015. 2
- [52] Wojciech Zielenka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. *arXiv preprint arXiv:2204.06607*, 2022. 2