

RAFaRe: Learning Robust and Accurate Non-parametric 3D Face Reconstruction from Pseudo 2D&3D Pairs

Longwei Guo, Hao Zhu[✉], Yuanxun Lu, Menghua Wu, Xun Cao

Nanjing University, Nanjing, China

{guolongwei, luyuanxun, menghuawu}@smail.nju.edu.cn, {zhuhaoese, caoxun}@nju.edu.cn

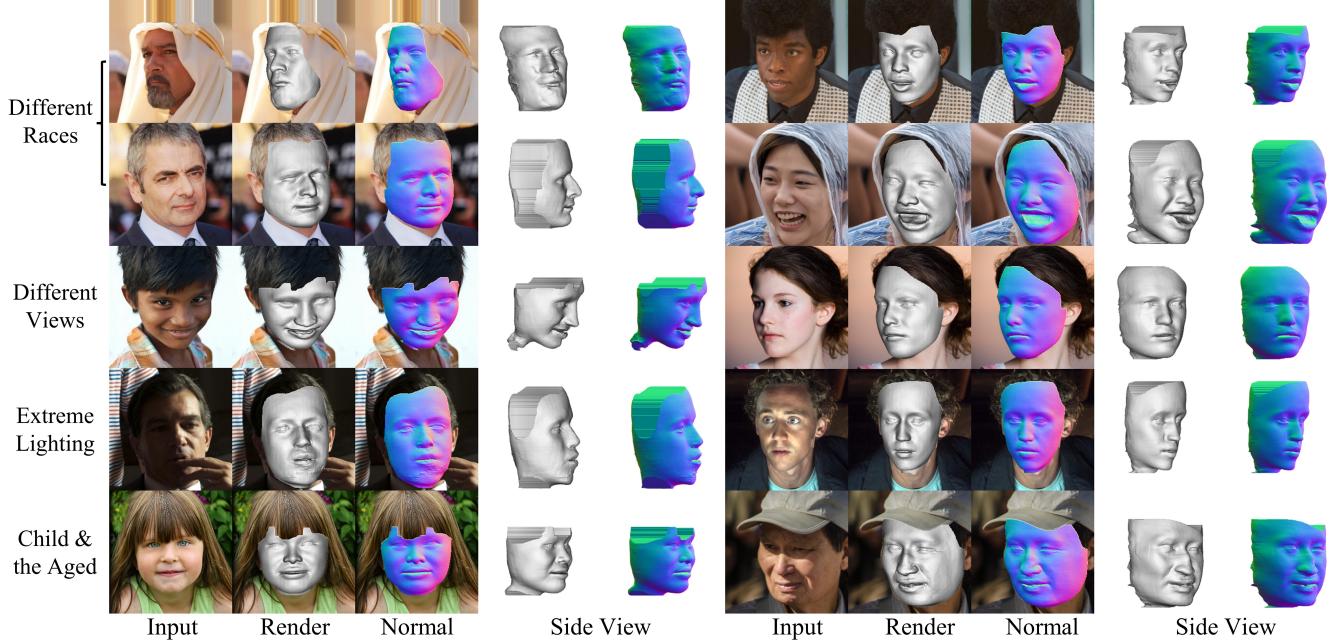


Figure 1: Our method reconstructs high-fidelity and accurate geometry that is generalized for different races, views, lighting and ages. We recommend watching the *supplementary video* for more results.

Abstract

We propose a robust and accurate non-parametric method for single-view 3D face reconstruction (SVFR). While tremendous efforts have been devoted to parametric SVFR, a visible gap still lies between the result 3D shape and the ground truth. We believe there are two major obstacles: 1) the representation of the parametric model is limited to a certain face database; 2) 2D images and 3D shapes in the fitted datasets are distinctly misaligned. To resolve these issues, a large-scale pseudo 2D&3D dataset is created by first rendering the detailed 3D faces, then swapping the face in the wild images with the rendered face. These pseudo 2D&3D pairs are created from publicly available datasets which eliminate the gaps between 2D and 3D data while covering diverse appearances, poses, scenes, and illumination. We further propose a non-parametric scheme to learn a well-generalized SVFR model from the created dataset, and the proposed hierarchical signed distance function turns out to be effective in predicting middle-scale and small-scale 3D facial geometry. Our model outperforms previous methods on FaceScape-wild/lab and MICC benchmarks and is well generalized to various appearances, poses, expressions, and in-the-wild environments. The code is released at <https://github.com/zuhao-nju/rafare>.

Introduction

Reconstructing a 3D face shape from an arbitrary single image is a long-standing problem in the computer vision and computer graphics community. It attracts much more attention in recent years for its fundamental impact on comprehensive downstream applications, e.g., facial editing, visual effects, facial animation, virtual make-up, and VR/AR character creation. About two decades ago, Blanz et al. (Blanz, Vetter et al. 1999) pioneered the 3D morphable model (3DMM) to tackle this problem, which further developed into a classic paradigm to alleviate the ambiguity of the problem. Many researchers employ a prior parametric model (Cao et al. 2013b; Li et al. 2017; Yang et al. 2020) as a coarse 3D model and optimize the parameters by minimizing semantic energy functions (Cao et al. 2013a; Thies et al. 2016). In recent years, the rapid development of deep learning pushes the field forward. Deep neural networks are proven to be more effective in regressing the 3DMM parameters, which can be trained on large-scale fitted or synthetic

[✉] Hao Zhu is the corresponding author.

datasets.

Despite the recent advances, achieving *robust* and *accurate* single-view 3D face reconstruction remains an open challenge. The difficulty is two-fold, lying in both face representations and training data. On the one hand, the representation ability of 3DMMs or other parametric models is limited to the fixed sub-linear space of the face database, which makes it hard to recover the fine-scale facial details and out-of-domain face attribute distributions. On the other hand, inaccurate and limited training databases restrict the reconstruction performance of both generalization and precision. Concretely, though accurate but limited studio-captured data is available (Yang et al. 2020; Dai et al. 2019), it is almost infeasible to obtain large-scale in-the-wild data at the same accuracy, which constrains the model generalization ability to a great extent. Previous approaches (Feng et al. 2018) tend to adopt in-the-wild image datasets with fitted 3DMM models as training data (Zhu et al. 2016), failing to reconstruct middle-scale 3D geometry.

In this paper, we create pseudo 2D&3D datasets from publicly available datasets to address these challenges. Specifically, we propose 1) a novel pipeline for large-scale in-the-wild 2D&3D facial data generation, and 2) a non-parametric method tailored to recover the facial geometry from a single image. Our reconstruction approach obviates the dependence on prior parametric models and allows for good generalization on poses, expressions, scenes, and illumination. The pipeline is designed to exploit the data augmentation insights and make the most effective use of existing large-scale in-the-wild facial image datasets and limited lab-environmental precise 3D facial shape datasets. Going a step further, we develop a non-parametric method to reconstruct facial geometry from monocular images directly, breaking the limitations of parametric models. The most notable improvement is the hierarchical signed distance function to recover detailed facial geometry at different levels. The non-parametric model is trained on the proposed large-scale faithful dataset and shows effective and superior performance over previous state-of-the-art approaches. In summary, we highlight the following contributions:

- We introduce a novel pipeline to generate a *large-scale*, *strictly-aligned*, and *in-the-wild* facial dataset with corresponding faithful geometry ground truth. This pipeline enables a higher ceiling for current non-parametric single-view face reconstruction models.
- We propose a hierarchical implicit function-based architecture to estimate the detailed 3D face geometry from a single image. The key point of the architecture is the non-parametric model, and unlike most previous 3DMM-based approaches, it is robust to variations of human races, poses, scenes, and illumination.
- We achieve the state-of-the-art accuracy on FaceScape-wild/lab (Yang et al. 2020) and MICC Florence (Bagdanov, Del Bimbo, and Masi 2011) benchmarks for single-view face reconstructions. Comprehensive qualitative and quantitative evaluations demonstrate the superiority of our method over competing approaches.

Related Work

Single-View 3D Face Reconstruction (SVFR) has been a hotspot for decades due to the enormous significance of potential applications and the intrinsic ambiguity and difficulty compared with other input modalities, e.g., video sequences, RGB-D data, and multi-view images. In this section, we briefly review the prior work in this field, along with related data augmentation methods. We recommend checking the survey (Zollhöfer et al. 2018) for a comprehensive overview. **Parametric SVFR Methods.** Parametric methods estimate the facial shape by regressing the parameters of 3DMM, which is a statistical model to transform the shape and texture of the faces into a vector space representation (Blanz, Vetter et al. 1999). For the formulation of 3DMM, please check the recent survey (Egger et al. 2020). Traditionally, these methods follow an analysis-by-synthesis schedule and build semantic correspondences between images and statistical models by optimization-based (Romdhani and Vetter 2005; Amberg, Romdhani, and Vetter 2007; Zhu et al. 2016; Thies et al. 2016; Dou, Shah, and Kakadiaris 2017) or learning-based algorithms (Sanyal et al. 2019b; Tran, Liu, and Liu 2019; Gecer et al. 2019; Tu et al. 2020; Koizumi and Smith 2020; Shang et al. 2020; Guo et al. 2020; Deng et al. 2019; Liu et al. 2018; Shang et al. 2020). Recently, researchers discover a self-supervised scheme to train the models by employing differentiable renderers (Tewari et al. 2018, 2017; Genova et al. 2018; Tan et al. 2020; Feng et al. 2021; Deng et al. 2019; Sanyal et al. 2019b). These methods inherit the limitations of the 3DMM, which lie in a fixed sub-linear space and generate only coarse shapes without facial details. Several works propose non-linear 3DMMs (Tran and Liu 2018; Tran, Liu, and Liu 2019; Yenamandra et al. 2021; Zhuang et al. 2022) to break the traditional limitations.

To add detailed facial shape to the coarse 3DMMs, some methods (Yang et al. 2020; Feng et al. 2021; Chen et al. 2019, 2020) predict displacement maps over the coarse models to represent details. Another attempts (Jiang et al. 2018; Sengupta et al. 2018; Richardson et al. 2017; Tran et al. 2018; Zhu et al. 2019, 2021b; Riviere et al. 2020) leverage the shape-from-shading method to reconstruct shape details. However, it is error-prone for outer occlusions, specular highlights, and strong cast shadows.

Non-parametric SVFR Methods. Non-parametric methods discard prior models and estimate facial geometry directly. These methods recover 3D faces in the form of volumes (Jackson et al. 2017), meshes (Feng et al. 2018; Ruan et al. 2021; Zeng, Peng, and Qiao 2019; Sela, Richardson, and Kimmel 2017; Zhou et al. 2019; Zhu et al. 2020), or depth maps (Zhang et al. 2021), and hence can capture finer shape details compared with parametric methods. However, the free-form representation leads to fewer constraints and makes it difficult for neural networks to predict the accurate shape. Unlike these approaches, we propose a hierarchical signed distance function-based model to reconstruct coarse- and fine-level geometry in different stages. This representation shows higher robustness, accuracy, and generalization over other non-parametric formulations and overcomes the flaws of 3DMMs.

Synthetic Data Augmentation. Datasets play a crucial role

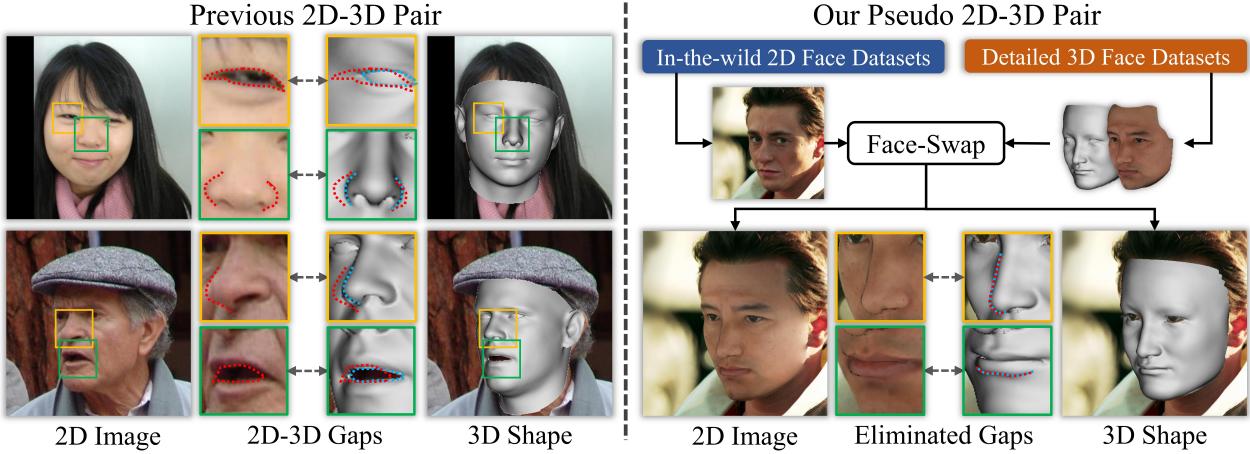


Figure 2: Left: In 300W-LP(Sagonas et al. 2016) dataset, the middle-scale features in the 2D images and 3D models are mismatched. The zoom-in figure shows that the outlines of the image (red dotted line) and that of the 3D shape (blue dotted line) are obviously misaligned. These 2D&3D gaps *commonly* exist in previous 2D&3D face datasets which limited the ceiling for previous reconstruction models. Right: our pseudo-2D&3D pairs (right) eliminate such gaps.

in the task of learning-based SVFR because their quality decides the performance of neural networks, while current public datasets cannot meet the demands. High-quality scans or multi-view reconstruction systems provide detailed and accurate shapes but lack generalization due to the complexity of the capture system (Xiao et al. 2022). On the other hand, in-the-wild datasets (Zhu et al. 2016) with optimization-based 3DMM fitting satisfy generality but introduce misalignment between 3D models and 2D images. Methods trained with these datasets (Feng et al. 2018; Jackson et al. 2017; Guo et al. 2020) therefore can only recover coarse geometry. Some methods (Dou, Shah, and Kakadiaris 2017; Richardson, Sela, and Kimmel 2016; Varol et al. 2017; Zhu et al. 2018) generate synthetic data for training, but their image is not photo-realistic and the 3D model is not precise enough. Instead, we utilize both accurate but limited lab-captured data and countless unconstrained images and then take the best of both sides. Our face-swapping-based pipeline is able to generate innumerable faithful pseudo-2D&3D pairs, meaning the network can be trained with diverse and reliable data to attain better performance over previous schemes.

Method

Revisit 2D&3D Gap

The quality of 2D&3D data for training has a crucial impact on the performance of the SVFR model but remains to be a weak point. To achieve optimal performance, the 2D&3D data for training should be 1) in large amounts; 2) covering diverse facial shapes, appearances, expressions, and environments; and 3) well aligned and accurately corresponded between 2D&3D. However, these requirements can hardly be met in practice, as capturing accurate 3D facial shapes in the wild is extremely difficult and expensive. A compromised solution that is commonly used is to build 2D&3D pairs by fitting a 3D face model to the in-the-wild

images(Sagonas et al. 2016), as shown in Figure 2. Though the requirements of amount and image diversity are met, the fitted 3D shapes and 2D images are visibly misaligned, leading to a poor fidelity of the predicted facial shape.

Pseudo 2D&3D Pairs

To eliminate the 2D&3D gaps, we propose to create pseudo 2D&3D pairs by swapping in-the-wild faces with accurately reconstructed faces. As shown in Figure 3, the pipeline consists of three stages.

In the first stage, we firstly select a face image from the large-scale in-the-wild datasets (CelebA(Lee et al. 2020), FFHQ(Karras, Laine, and Aila 2019), IMDB-WIKI(Rothe, Timofte, and Van Gool 2015)), then randomly fit a 3DMM created by FaceScape dataset(Yang et al. 2020) or HeadSpace dataset(Dai et al. 2019) to this image by optimizing 2D facial landmark locations(Yang et al. 2020). Then, a UV texture map is randomly selected from the two 3D face datasets, and a 2D face can be rendered from the fitted 3D mesh and the selected UV texture map.

In the second stage, the parsing mask is extracted from the in-the-wild image using BiSeNet(Yu et al. 2018). Then the complementing network synthesizes the missing eyes and teeth for the rendered face, and the inpainting network adjusts the edge of the rendered image to match the parsing mask. Specifically, the face region of the parsing mask and rendered face are concatenated as input of the inpainting network, which aims to obtain the complete face of the intersection region of the facial mask and rendered face. Meanwhile, we use the image warping algorithm to adjust the edge of the in-the-wild image to fit the intersection region. The adjustment aims to make the in-the-wild image and the inpainted image completely aligned.

In the final stage, Poisson blending is used to merge the in-the-wild face and the inpainted face, generating the blended face. The blended face and the fitted 3D model constitute the

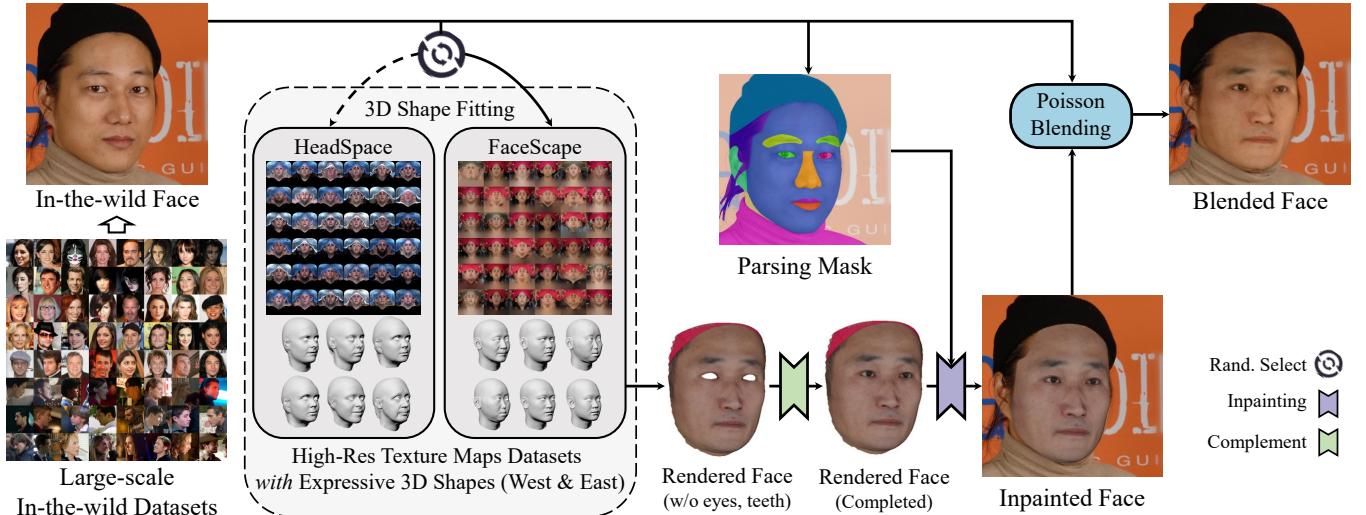


Figure 3: Overview of the proposed pseudo 2D&3D pair synthesis pipeline.

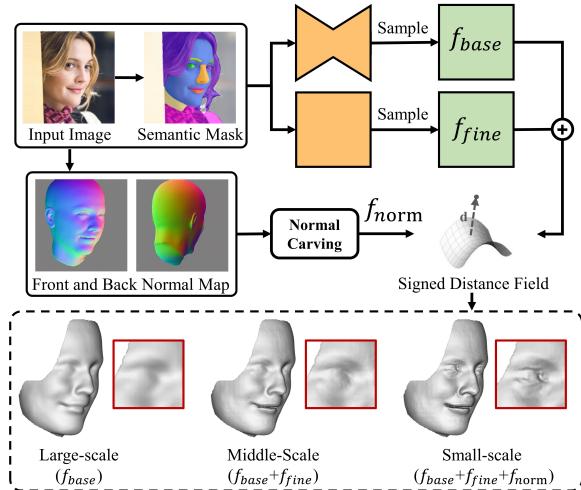


Figure 4: Overview of the hierarchical SDF-based network.

pseudo 2D&3D pairs, which are accurately corresponded and cover diverse facial shapes, expressions, poses, scenes and illuminations.

In practice, we synthesize 100k pseudo 2D&3D pairs, of which 18k images contain large-pose faces. Both our complementing network and inpainting network adopt pix2pixHD (Wang et al. 2018) as the backbone. Please refer to the supplementary material for implementation details.

Hierarchical Signed Distance Function

The creation of pseudo 2D&3D pairs makes it possible to train a non-parametric SVFR model and break through the constraints of 3DMM solution space. We propose a hierarchical signed distance function to serve as the base of our non-parametric SVFR network, as shown in Figure 4. Inspired by PIFu (Saito et al. 2019), our network represents 3D facial shapes by learning a pixel-aligned signed distance function. The source image and the parsed semantic mask (Yu et al. 2018) are fed into two feature extractors,

then two MLPs take pixel-aligned features and depth value z as input, predicting a signed distance for each sample point. The final predicted signed distance field is the sum of the predicted results from f_{base} , f_{fine} , and f_{norm} . Both f_{base} and f_{fine} are modeled by two MLPs, and the difference is that they adopt a stacked hourglass network and a shallow convolutional neural network as feature extractors respectively. This design aims to extract global features for large-scale shape recovery and to extract local features for middle-scale shape recovery. The two implicit functions are formulated as:

$$f_{base} : F(x, y), z \rightarrow d \quad (1)$$

$$f_{fine} : F(x, y), z \rightarrow \nabla d = d - (d \otimes g_k) \quad (2)$$

where d is the signed distance sampled in the 3D volume; $F(x, y)$ is the pixel-aligned features extracted from the images; \otimes is the 3D convolution operation with and the k^3 mean kernel. k is set to 5 in all our experiments.

We observed that the SDF and other implicit functions fail to model the geometric details when representing a large-scale database, so an additional normal carving operation f_{norm} is presented to model small-scale geometry in an explicit manner. The pix2pixHD (Wang et al. 2018) is used to regress frontal and back normal maps by training on ground-truth normal maps. The normal carving operation f_{norm} transforms the normal map to a 3D displacement field, which is formulated as:

$$f_{norm} : d \rightarrow d + \mathcal{N}(x, y) \otimes G_s \quad (3)$$

where $\mathcal{N}(x, y)$ is the regressed frontal or back normal maps, and G_s is the 7×7 Sobel operator. f_{norm} apply the front and back normal maps to the front and back half of the signed distance field respectively.

Empirically, the three signed distance fields predicted by f_{base} , f_{fine} , and f_{norm} are designed to focus on large-scale, middle-scale, and small-scale 3D geometry respectively. Experiments show that the combination leads to both visually plausible and quantitatively accurate 3D face reconstruction.

Table 1: Quantitative evaluation on FaceScape in-the-wild benchmark

Pose Angle → Method ↓	0° – 5°			5° – 30°			30° – 60°			60° – 90°			
	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	Succ.
Ext3dFace (Tran et al. 2018)	5.03	0.158	61.5	5.52	0.176	55.7	7.92	0.208	40.4	25.39	0.266	27.1	85.5
PRNet (Feng et al. 2018)	2.61	0.119	83.0	3.11	0.114	82.7	4.26	0.119	78.2	3.88	0.140	75.3	100.0
D3DFR (Deng et al. 2019)	2.30	0.070	83.1	2.50	0.072	83.0	3.57	0.082	77.8	6.81	0.143	62.4	100.0
RingNet (Sanyal et al. 2019b)	2.40	0.085	99.8	2.99	0.085	99.7	4.78	0.100	98.4	10.71	0.190	97.1	100.0
DFDN (Chen et al. 2019)	3.66	0.090	86.6	3.27	0.091	86.5	7.29	0.128	84.3	27.48	0.302	57.2	88.2
DF2Net (Zeng, Peng, and Qiao 2019)	2.92	0.121	57.1	4.21	0.128	55.3	6.55	0.159	46.3	19.76	0.309	30.8	68.8
UDL (Chen et al. 2020)	2.27	0.091	69.0	2.59	0.092	68.3	3.46	0.106	65.0	6.32	0.176	49.0	86.2
FaceScape (Yang et al. 2020)	2.81	0.086	83.7	3.17	0.092	82.0	4.09	0.108	79.0	6.57	0.162	67.9	96.0
MGCNet (Shang et al. 2020)	2.97	0.073	84.4	2.94	0.073	84.5	2.78	0.070	81.6	4.21	0.091	74.3	100.0
3DDFA-v2 (Guo et al. 2020)	2.49	0.074	86.5	2.66	0.074	86.0	3.18	0.078	83.1	3.67	0.093	79.9	100.0
SADRNet (Ruan et al. 2021)	6.60	0.113	90.2	6.87	0.113	89.4	6.40	0.103	84.4	8.63	0.163	82.7	100.0
LAP (Zhang et al. 2021)	4.19	0.111	93.5	4.47	0.116	92.8	6.16	0.148	87.3	13.71	0.205	68.1	100.0
DECA (Feng et al. 2021)	2.88	0.080	99.9	2.64	0.079	99.9	2.88	0.082	99.8	4.83	0.116	99.7	100.0
Ours	1.79	0.058	99.5	1.92	0.062	99.3	2.12	0.071	98.4	5.26	0.123	96.7	100.0

The highest 5 scores are colored: ‘ green → yellow ’ means ‘high→low’, similarly hereinafter.

Table 2: Quantitative evaluation on FaceScape in-the-lab benchmark

Pose Angle → Method ↓	0°			30°			60°			
	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	
Ext3dFace (Tran et al. 2018)	4.59	0.131	86.2	7.42	0.170	69.1	8.51	0.175	55.2	85.9
PRNet (Feng et al. 2018)	2.94	0.133	92.5	3.40	0.125	90.1	3.74	0.122	85.2	100.0
D3DFR (Deng et al. 2019)	3.99	0.106	87.6	5.90	0.120	81.3	5.55	0.137	75.3	98.9
RingNet (Sanyal et al. 2019b)	3.62	0.102	99.9	5.03	0.111	99.7	6.82	0.151	94.5	100.0
DFDN (Chen et al. 2019)	4.28	0.111	98.4	6.71	0.132	95.2	23.63	0.280	81.0	94.7
DF2Net (Zeng, Peng, and Qiao 2019)	4.48	0.152	64.1	7.64	0.200	52.2	—*	—*	—*	73.6
UDL (Chen et al. 2020)	2.21	0.092	79.5	5.34	0.123	71.3	5.63	0.167	61.9	87.0
FaceScape (Yang et al. 2020)	3.21	0.090	94.2	4.87	0.119	86.2	4.68	0.146	81.7	92.0
MGCNet (Shang et al. 2020)	3.45	0.085	92.7	3.91	0.093	90.1	3.65	0.090	83.2	100.0
3DDFA-v2 (Guo et al. 2020)	3.05	0.093	95.2	3.41	0.096	93.8	3.82	0.097	88.2	100.0
SADRNet (Ruan et al. 2021)	4.25	0.109	95.8	7.07	0.137	94.9	7.09	0.148	87.6	100.0
LAP(Zhang et al. 2021)	4.27	0.112	96.4	7.33	0.149	93.2	8.70	0.195	85.6	99.2
DECA (Feng et al. 2021)	3.30	0.093	99.8	4.14	0.100	99.4	4.20	0.107	97.1	100.0
Ours	2.95	0.093	98.9	3.69	0.102	98.7	5.28	0.111	98.3	100.0

* — means no valid results are generated in this category.

Experiment

Dataset and Metric

The qualitative evaluation is on FFHQ (Karras, Laine, and Aila 2019), IMDB-WIKI (Rothe, Timofte, and Van Gool 2015), and AFLW2000 (Zhu et al. 2016) datasets, which are in-the-wild 2D image datasets covering different poses, races, ages, environments, and lighting. The quantitative evaluation is on FaceScape (Yang et al. 2020; Zhu et al. 2021a), and MICC Florence (Bagdanov, Del Bimbo, and Masi 2011), which are face datasets containing accurate 3D shapes and 2D images.

We follow the evaluation methodology explained in FaceScape benchmark(Yang et al. 2020; Zhu et al. 2021a) to evaluate the accuracy of the reconstructed shape at the time of the photo, which means poses and expressions are also factored into the calculation of error. Generally, CD measures the overall shape accuracy; MNE measures the local shape accuracy; CR indicates if the result face is complete. Different from NoW benchmarks (Sanyal et al. 2019a),

which evaluate the ‘expression-neutralized’ and ‘unposed’ facial shape, the benchmark we used evaluates the accuracy of the predicted 3D shape at the ‘time of the photo’ and takes pose estimation into consideration.

Visual Comparison

We show our results on extreme conditions in Figure 1 and compare our method with previous methods in Figure 5. We can see that our results are visually more faithful and well-aligned with the source image. We believe it is because our created datasets eliminate the gaps between 2D images and 3D shapes and boost the performance of the non-parametric SVFR model. Please watch our video to validate the temporal continuity.

Quantitative Comparison

We conduct quantitative evaluations on three datasets. The evaluation on FaceScape-wild (Zhu et al. 2021a) dataset is reported in Table 1. FaceScape-wild dataset contains 400 synthetic in-the-wild images categorized by the pose angle.

Table 3: Quantitative evaluation on MICC Florence dataset.

Pose Angle → Method ↓	0°			30°			60°			
	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	Succ.
Ext3dFace (Tran et al. 2018)	3.33	0.114	96.0	3.74	0.129	73.3	5.71	0.150	53.7	88.8
PRNet (Feng et al. 2018)	2.53	0.119	98.1	2.34	0.114	97.8	2.19	0.124	98.1	100.0
D3DFR (Deng et al. 2019)	3.07	0.119	91.2	4.09	0.122	89.5	7.06	0.154	84.4	100.0
RingNet (Sanyal et al. 2019b)	2.12	0.100	99.7	3.27	0.102	99.7	6.98	0.176	98.7	100.0
DFDN (Chen et al. 2019)	4.28	0.107	99.3	5.70	0.118	99.3	23.34	0.245	82.4	91.2
DF2Net (Zeng, Peng, and Qiao 2019)	3.60	0.130	79.5	6.09	0.190	64.1	7.81	0.210	46.1	58.8
UDL (Chen et al. 2020)	2.70	0.110	94.4	2.96	0.111	93.7	5.42	0.169	84.0	80.0
FaceScape (Yang et al. 2020)	3.90	0.124	97.2	3.66	0.129	95.0	5.76	0.187	86.9	99.2
MGCNet (Shang et al. 2020)	3.00	0.086	95.7	2.83	0.090	95.4	2.82	0.096	94.5	100.0
3DDFA_v2 (Guo et al. 2020)	2.56	0.088	97.9	2.19	0.086	97.7	2.27	0.091	98.0	100.0
SADRNet (Ruan et al. 2021)	6.10	0.134	99.3	5.66	0.129	98.6	6.73	0.141	96.9	100.0
LAP (Zhang et al. 2021)	3.74	0.121	97.6	6.02	0.152	96.8	10.78	0.214	88.4	99.2
DECA (Feng et al. 2021)	2.55	0.107	100.0	2.83	0.108	100.0	4.27	0.110	100.0	100.0
Ours	2.25	0.102	100.0	2.53	0.104	100.0	4.66	0.114	99.9	100.0

Table 4: Ablation Study on FaceScape-wild dataset

Pose Angle → Method ↓	0° – 5°			5° – 30°			30° – 60°			60° – 90°			
	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	
PRNet (Feng et al. 2018)	2.61	0.119	83.0	3.11	0.114	82.7	4.26	0.119	78.2	3.88	0.140	75.3	100.0
PRNet-pseudo data	1.97	0.096	78.5	2.01	0.102	78.3	2.61	0.122	75.5	6.93	0.212	51.2	100.0
PIFu (Saito et al. 2019)	2.14	0.066	99.4	2.25	0.070	99.2	3.10	0.084	98.6	4.92	0.118	97.1	100.0
f_{base}	1.81	0.062	99.0	1.95	0.066	98.8	2.16	0.072	98.4	5.04	0.124	95.7	100.0
$f_{base} + f_{fine}$	1.79	0.062	98.9	1.90	0.066	98.8	2.12	0.073	98.4	5.16	0.125	95.7	100.0
$f_{base} + f_{fine} + f_{norm}$	1.79	0.058	99.5	1.92	0.062	99.3	2.12	0.071	98.4	5.26	0.123	96.7	100.0

We can see that our method leads in CD and MNE when pose angles are in $0^\circ - 60^\circ$ angles. Our performance in $60^\circ - 90^\circ$ angles are slightly worse but still ranks top 5 in all metrics. Our method is the only one that ranks the top 5 in all metrics and leads in the $0^\circ - 60^\circ$ pose by a large margin.

The quantitative evaluation on FaceScape-lab (Zhu et al. 2021a) dataset is reported in Table 2. FaceScape-lab dataset contains 660 images rendered from 20 studio-captured detailed 3D facial raw scans using the perspective projection camera model. Similarly, our method is the only one that is in the top five across all metrics.

The quantitative evaluation on MICC Florence (Bagdanov, Del Bimbo, and Masi 2011) dataset is reported in Table 2. MICC Florence dataset contains 52 studio-captured detailed 3D face models, which are rendered to 260 images using the perspective projection camera model. The major difference compared with FaceScape-lab is that MICC Florence mainly contains western faces while FaceScape-lab contains oriental faces. Our method leads on MICC Florence benchmark in the mean of average ranking of all metrics, which shows that our models work robustly and accurately for both western and oriental faces.

Ablation Study

The results of the ablation study are shown in Table 4. ‘PRNet’ is the model that is trained on the original fitted dataset; ‘PRNet-pseudo data’ means the model trained on our pseudo 3D&3D data; ‘PIFu’ is a non-parametric single-view re-

construction method trained on our 2D&3D data; ‘ f_{base} ’, ‘ $f_{base} + f_{fine}$ ’ and ‘ $f_{base} + f_{fine} + f_{norm}$ ’ means part of our hierarchical SDF as explained in Figure 4.

By comparing ‘PRNet’ and ‘PRNet-pseudo data’, we can conclude that our pseudo 2D&3D data can be used in improving other non-parametric methods. By comparing ‘PIFu’ to ‘ f_{base} ’, we verify that introducing SDF improves the performance of SVFR. By comparing ‘ f_{base} ’, ‘ $f_{base} + f_{fine}$ ’ and ‘ $f_{base} + f_{fine} + f_{norm}$ ’, we verify that the hierarchical SDF further improves performance in each level. ‘ f_{fine} ’ focuses on improving the middle-scale shape and improving the accuracy in terms of CD. Though the additional ‘ f_{norm} ’ doesn’t affect much to CD, it enhances the MNE which means the detailed shape is recovered more accurately. These improvements are also confirmed by visual comparison in Figure 4. For the viewing angle $> 60^\circ$ where almost half of the facial regions are occluded, ‘ f_{fine} ’ and ‘ f_{norm} ’ struggled in hallucinating uncertain occluded shape, thus the score of CD reduces slightly.

Discussion of Limitations.

Non-uniform mesh topology. Due to the non-parametric scheme, the topology of our result meshes is not uniform, so an additional registration phase is required before being used in downstream applications like rigging and animation. The non-uniform topology also increases the possibility to produce broken mesh in extreme poses and lighting conditions, which will be shown in the supplementary.



Figure 5: Qualitative comparison. Images with blue bars are from in-the-wild datasets(FFHQ, IMDB-WIKI, AFLW2000); images with green bars are from synthetic or studio-captured datasets (FaceScape-wild/lab, MICC-Florence).

Inveracious lightings. In some of our pseudo images, the synthesized lighting is not photo-realistic enough with no self-shading and specular highlight. Besides, the illuminations on the face and the background may differ much in some cases. Whether these factors will affect performance remains an open question.

Inferior performance under large poses. From the quantitative evaluation (Table 1/2/3), we can see that the reconstruction accuracy of minor-pose faces in our method is significantly higher than that of large-pose faces. We believe that the key reason is the insufficient large-pose wild images used in generating our pseudo 2D&3D data. Specifically, only 0.18% of the photos in CelebA dataset and 0.21% of the photos in the FFHQ dataset are taken from views at $60^\circ - 90^\circ$. We manually add more large-pose images from other datasets, increasing the ratio of large-pose images in

our training set to 12.14%, but they are still too few. This problem is not unique to our method but to most previous methods, as illustrated in Table 1/2/3. We consider that how to augment the data with large-pose faces remains a challenge for the SVFR task.

Conclusion

We propose a novel approach for the single-view 3D face reconstruction task in a non-parametric scheme. Our method gets rid of the heavy dependence on the statistic model and, therefore, its limitations and achieves state-of-the-art performance by learning from our created pseudo 2D&3D datasets. A novel solution to build a large-scale and accurate in-the-wild 3D face dataset is presented, filling the gap of image-shape alignment in previous datasets. We hope our work could inspire future researchers in this field.

Acknowledgements

This work was supported by the NSFC grant 62001213, 62025108, the National Key R&D Program of China grant 2022YFF0902401, and gift funding from Huawei Research and Tencent Rhino-Bird Research Program.

References

- Amberg, B.; Romdhani, S.; and Vetter, T. 2007. Optimal step nonrigid ICP algorithms for surface registration. In *CVPR*, 1–8.
- Bagdanov, A. D.; Del Bimbo, A.; and Masi, I. 2011. The Florence 2D/3D Hybrid Face Dataset. In *Joint ACM Workshop on Human Gesture and Behavior Understanding*, 79–80.
- Blanz, V.; Vetter, T.; et al. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, volume 99, 187–194.
- Cao, C.; Weng, Y.; Lin, S.; and Zhou, K. 2013a. 3D shape regression for real-time facial animation. *TOG*, 32(4): 1–10.
- Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013b. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20(3): 413–425.
- Chen, A.; Chen, Z.; Zhang, G.; Mitchell, K.; and Yu, J. 2019. Photo-realistic facial details synthesis from single image. In *ICCV*, 9429–9439.
- Chen, Y.; Wu, F.; Wang, Z.; Song, Y.; Ling, Y.; and Bao, L. 2020. Self-supervised learning of detailed 3d face reconstruction. *TIP*, 29: 8696–8705.
- Dai, H.; Pears, N.; Smith, W.; and Duncan, C. 2019. Statistical Modeling of Craniofacial Shape and Texture. *IJCV*, 128(2): 547–571.
- Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 0–0.
- Dou, P.; Shah, S. K.; and Kakadiaris, I. A. 2017. End-to-end 3D face reconstruction with deep neural networks. In *CVPR*, 5908–5917.
- Egger, B.; Smith, W. A.; Tewari, A.; Wuhrer, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 2020. 3d morphable face models—past, present, and future. *TOG*, 39(5): 1–38.
- Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *TOG*, 40(4): 1–13.
- Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 534–551.
- Gecer, B.; Ploumpis, S.; Kotsia, I.; and Zafeiriou, S. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 1155–1164.
- Genova, K.; Cole, F.; Maschinot, A.; Sarna, A.; Vlasic, D.; and Freeman, W. T. 2018. Unsupervised training for 3d morphable model regression. In *CVPR*, 8377–8386.
- Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; and Li, S. Z. 2020. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 152–168.
- Jackson, A. S.; Bulat, A.; Argyriou, V.; and Tzimiropoulos, G. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *ICCV*, 1031–1039.
- Jiang, L.; Zhang, J.; Deng, B.; Li, H.; and Liu, L. 2018. 3D face reconstruction with geometry details from a single image. *TIP*, 27(10): 4756–4770.
- Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.
- Koizumi, T.; and Smith, W. A. 2020. “Look Ma, no landmarks!”—Unsupervised, model-based dense face alignment. In *ECCV*, 690–706. Springer.
- Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 5549–5558.
- Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *TOG*, 36(6): 194–1.
- Liu, F.; Zhu, R.; Zeng, D.; Zhao, Q.; and Liu, X. 2018. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *CVPR*, 5216–5225.
- Richardson, E.; Sela, M.; and Kimmel, R. 2016. 3D face reconstruction by learning from synthetic data. In *3DV*, 460–469. IEEE.
- Richardson, E.; Sela, M.; Orel, R.; and Kimmel, R. 2017. Learning Detailed Face Reconstruction from a Single Image. In *CVPR*, 5553–5562.
- Riviere, J.; Gotardo, P.; Bradley, D.; Ghosh, A.; and Beeler, T. 2020. Single-shot high-quality facial geometry and skin appearance capture. *TOG*, 39(4): 81–1.
- Romdhani, S.; and Vetter, T. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, volume 2, 986–993.
- Rothe, R.; Timofte, R.; and Van Gool, L. 2015. Dex: Deep expectation of apparent age from a single image. In *ICCV Workshops*, 10–15.
- Ruan, Z.; Zou, C.; Wu, L.; Wu, G.; and Wang, L. 2021. SADRNet: Self-Aligned Dual Face Regression Networks for Robust 3D Dense Face Alignment and Reconstruction. *TIP*.
- Sagonas, C.; Antonakos, E.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2016. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47: 3–18.
- Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2304–2314.
- Sanyal, S.; Bolkart, T.; Feng, H.; and Black, M. 2019a. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *CVPR*.
- Sanyal, S.; Bolkart, T.; Feng, H.; and Black, M. J. 2019b. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, 7763–7772.

- Sela, M.; Richardson, E.; and Kimmel, R. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 1576–1585.
- Sengupta, S.; Kanazawa, A.; Castillo, C. D.; and Jacobs, D. W. 2018. SfSNet: Learning Shape, Reflectance and Illuminance of Faces in the Wild. In *CVPR*, 6296–6305.
- Shang, J.; Shen, T.; Li, S.; Zhou, L.; Zhen, M.; Fang, T.; and Quan, L. 2020. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *ECCV*, 53–70.
- Tan, F.; Zhu, H.; Cui, Z.; Zhu, S.; Pollefeys, M.; and Tan, P. 2020. Self-supervised human depth estimation from monocular videos. In *CVPR*, 650–659.
- Tewari, A.; Zollhöfer, M.; Garrido, P.; Bernard, F.; Kim, H.; Pérez, P.; and Theobalt, C. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2549–2559.
- Tewari, A.; Zollhofer, M.; Kim, H.; Garrido, P.; Bernard, F.; Perez, P.; and Theobalt, C. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *CVPRW*, 1274–1283.
- Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2387–2395.
- Tran, A. T.; Hassner, T.; Masi, I.; Paz, E.; Nirkin, Y.; and Medioni, G. G. 2018. Extreme 3D Face Reconstruction: Seeing Through Occlusions. In *CVPR*, 3935–3944.
- Tran, L.; Liu, F.; and Liu, X. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *CVPR*, 1126–1135.
- Tran, L.; and Liu, X. 2018. Nonlinear 3D face morphable model. In *CVPR*, 7346–7355.
- Tu, X.; Zhao, J.; Xie, M.; Jiang, Z.; Balamurugan, A.; Luo, Y.; Zhao, Y.; He, L.; Ma, Z.; and Feng, J. 2020. 3d face reconstruction from a single image assisted by 2d face images in the wild. *TMM*, 23: 1160–1172.
- Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *CVPR*, 109–117.
- Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*.
- Xiao, Y.; Zhu, H.; Yang, H.; Diao, Z.; Lu, X.; and Cao, X. 2022. Detailed Facial Geometry Recovery from Multi-view Images by Learning an Implicit Function. In *AAAI*.
- Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *CVPR*.
- Yenamandra, T.; Tewari, A.; Bernard, F.; Seidel, H.-P.; Elgharib, M.; Cremers, D.; and Theobalt, C. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*, 12803–12813.
- Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*.
- Zeng, X.; Peng, X.; and Qiao, Y. 2019. DF2Net: A dense-fine-finer network for detailed 3D face reconstruction. In *ICCV*, 2315–2324.
- Zhang, Z.; Ge, Y.; Chen, R.; Tai, Y.; Yan, Y.; Yang, J.; Wang, C.; Li, J.; and Huang, F. 2021. Learning to Aggregate and Personalize 3D Face from In-the-Wild Photo Collection. In *CVPR*, 14214–14224.
- Zhou, Y.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2019. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, 1097–1106.
- Zhu, H.; Su, H.; Wang, P.; Cao, X.; and Yang, R. 2018. View extrapolation of human body from a single image. In *CVPR*, 4450–4459.
- Zhu, H.; Yang, H.; Guo, L.; Zhang, Y.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2021a. FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction. *arXiv preprint arXiv:2111.01082*.
- Zhu, H.; Zuo, X.; Wang, S.; Cao, X.; and Yang, R. 2019. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 4491–4500.
- Zhu, H.; Zuo, X.; Yang, H.; Wang, S.; Cao, X.; and Yang, R. 2021b. Detailed avatar recovery from single image. *PAMI*.
- Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *CVPR*, 146–155.
- Zhu, X.; Yang, F.; Huang, D.; Yu, C.; Wang, H.; Guo, J.; Lei, Z.; and Li, S. Z. 2020. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *ECCV*, 343–358.
- Zhuang, Y.; Zhu, H.; Sun, X.; and Cao, X. 2022. Mofanerf: Morphable facial neural radiance field. In *ECCV*, 268–285. Springer.
- Zollhöfer, M.; Thies, J.; Garrido, P.; Bradley, D.; Beeler, T.; Pérez, P.; Stamminger, M.; Nießner, M.; and Theobalt, C. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. In *CGF*, volume 37, 523–550.