ReEnFP: Detail-Preserving Face Reconstruction by Encoding Facial Priors

Yasheng Sun
Tokyo Institute of Technology
Tokyo, Japan
sun.y.aj@m.titech.ac.jp

Jiangke Lin
Zhejiang University
Hangzhou, China
jiangke_lin@zju.edu.cn

Hang Zhou
Baidu Inc.
Shanghai, China
zhouhang09@baidu.com

Zhiliang Xu
Baidu Inc.
Shenzhen, China
xuzhiliang@baidu.com

Dongliang He
Baidu Inc.
Beijing, China
hedongliang01@baidu.com

Hideki Koike
Tokyo Institute of Technology
Tokyo, Japan
koike@c.titech.ac.jp

Figure 1: **Qualitative Results of Reconstruction by Encoding Facial Priors (ReEnFP).** We focus on the task of detail-preserving face reconstruction, which recovers texture and geometry with vivid details manifestation from input image. The bottom-left shows the refined mesh obtained by geometry prior encoding. The bottom-right demonstrates the reconstructed appearance through texture prior encoding.

Abstract

We address the problem of face modeling, which is still challenging in achieving high-quality reconstruction results efficiently. Neither previous regression-based nor optimization-based frameworks could well balance between the facial reconstruction fidelity and efficiency. We notice that the large amount of in-the-wild facial images contain diverse appearance information, however, their underlying knowledge is not fully exploited for face modeling. To this end, we propose our ***Reconstruction by Encoding Facial Priors (ReEnFP)*** *pipeline to exploit the potential of unconstrained facial images for further improvement. Our key is to* encode generative priors learned by a style-based texture generator on unconstrained data for fast and detail-preserving face reconstruction. *With our texture generator pre-trained using a differentiable renderer, faces could be encoded to its latent space as opposed to the time-consuming optimization-based inversion. Our generative prior encoding is further enhanced with a pyramid fusion block for adaptive integration of input spatial information. Extensive experiments show that our method reconstructs photo-realistic facial textures and geometric details with precise identity recovery.*

1. Introduction

Reconstructing 3D facial geometry and texture from a single image is an important task in the computer vision and graphics field, leading to countless applications such as

face editing [54, 16], virtual reality [11, 8] and face recognition [43, 72]. It is very challenging to efficiently reconstruct realistic and identity-preserving appearance from a single input image. 1) Studies relying on parametric model such as 3D Morphable Model (3DMM) [7, 9, 23, 56, 55] assume linear property of facial appearance space. It sacrifices model expressiveness to describe detailed and realistic texture, thereby causing non-realism or blurry artifacts. 2) Another slew of studies [21, 53] that employ Generative Adversarial Networks (GANs) [24] could generate realistic texture, but they require laborious data collection procedure. This usually demands expensive 3D-scanning equipment and long-time monitoring from experienced specialists. The data diversity and quality will further constrain upper limit of model performance.

On the other hand, large corpus of in-the-wild images contain diverse appearance information. Intuitively, they are potentially beneficial for face reconstruction. Several recent attempts have been made [54, 43, 20] to involve in-the-wild images. Nevertheless, some of them [54, 43] produce unsatisfactory results [54]. Particularly, [20] requires a large amount of time for optimization and a tedious data post-processing procedure, which limits their applications. The problem of how to exploit the underlying rich information within real-world images for high-fidelity face reconstruction is still worth exploring.

To tackle the above challenge, we propose the **Reconstruction by Encoding Facial Priors (ReEnFP)** pipeline, aiming to take full advantage of StyleGAN's expressive power with a sophisticated encoding strategy. The key is to *learn high-quality facial priors with style-based generators, and encoding faces into the learned latent spaces for fast and detail-preserving face reconstruction*. Specifically, for realistic and diverse appearance representation, we adopt a style-based generator with modifications on the dual representation inspired by [27]. Rather than using limited UV texture dataset [21], enormous in-the-wild images are employed to facilitate model performance. Aided by a differentiable renderer [22], modeling UV textures solely with real-world images can be achieved through 3DMM-guided rendering and adversarial training. To ease the training difficulty and encourage superior appearance modeling, this generator is trained through dual learning of a set of (pseudo) albedo and illumination. Particularly, we devise a low-dimensional implicit illumination code which functions in a similar way as the shading process in Spherical Harmonics (SH) Lighting [47]. It synthesize ratio images [27] progressively by weight modulation operation. Baking it into the pseudo albedo [27], the final UV texture is yielded. Similarly, a displacement map [49, 67] generator describing the structural details like winkles is also introduced as geometric prior.

While the pre-trained facial priors in the generators guarantee both the diversity of the identities and the quality for recovery, a strategy is required to comprehensively exploit its potential. A natural approach is to project the image feature to $\mathcal{W}$ or $\mathcal{W}+$ space [2, 57, 48] of the generator by optimization. However, optimization-based methods are too slow to be acceptable on many cases. Thus we naturally seek an efficient and effective way of encoding images into the learned generative priors. Specifically, we involve a new encoder structure, termed as *Adaptive Fusion Block*, to adaptively integrate extracted multi-resolution spatial features with forwarded features of the fixed generators for better identity preservation. Armed with this encoder, the high-fidelity predictions of both textures and geometric displacements can be achieved in an efficient manner.

Our contributions are summarized as follows: **1)** We propose the **Reconstruction by Encoding Facial Priors (ReEnFP)** framework, which achieves efficient detail-preserving face reconstruction with high-quality texture and detail-enhanced geometry. **2)** For the purpose of mining the facial diversity underlying large corpus of in-the-wild images, a novel style-based architecture is proposed to learn appearance prior with dual representation. **3)** A pyramid fusion block is devised for generative prior encoding, which facilitates identity-consistent texture reconstruction and detail preservation.

2. Related Work

3D Morphable Models. Blanz and Vetter [6] first proposed the concept of 3DMM that represents face model with linear bases of shape, expression and texture by Principal Component Analysis (PCA) on collected 3D facial scans. The 3DMM model is composed of shape bases $\mathbf{A}_{id}$, expression bases $\mathbf{A}_{exp}$ and mean shape $\overline{\mathbf{S}}$. A template face mesh is constructed as $\mathbf{S} = \overline{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}$ after fitting the shape and expression coefficients α. Due to the inherently linear nature, approaches [6, 7, 9, 23, 29, 37] regressing 3DMM coefficients usually lack facial details and likeness of the subject. To extend the representation power of linear model, later approaches *et al.* [60, 61, 59] propose nonlinear 3DMM, encoding shape, texture and camera parameters as latent codes in deep neural network. While spanning wider space and outperforming previous work, their results still suffer from blurry and low-quality artifacts.

Non-Parametric Models. Some studies [30, 52, 3, 69, 71, 41] do not rely on the parametric model but directly regress 3D face model such as position map, depth or face norm. Despite higher flexibility, these methods [30, 3, 19] usually capture limited geometric details since many of them utilize synthetic data created by the statistical model as supervision. Other model-free approaches [71, 65] with only weakly symmetric constraints are able to recover more details but suffer from appearance ambiguity and incorrect geometric structure.

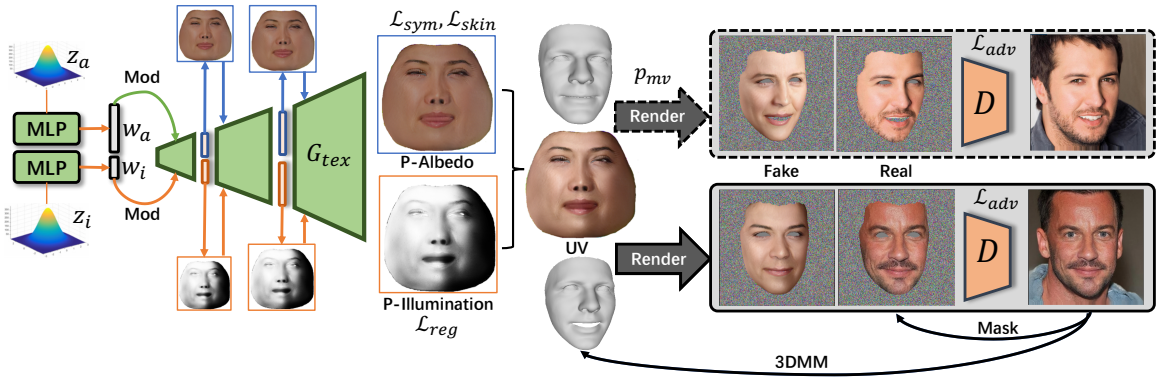

Figure 2: **The Architecture of Texture Prior Modeling.** The input latent codes z_a and z_i modulate common spatial features to generate p-albedo and p-illumination within two branches, respectively. After integrating and rendering them via differentiable renderer, we enforce rendered image as realistic as possible by adversarial loss $\mathcal{L}_{adv}$.

Detailed Reconstruction. A lot of work [10, 35, 38] has been conducted to yield vivid textures. For producing high-resolution photorealistic 3D faces, Lattas *et al.* [35] capture a large dataset of shape and reflectance, with which their methodology exhibits an unprecedented level of realism. Another slew of studies [51, 49, 26, 67] attempt to recover mesoscopic geometric facial details. Usually, they will refine an obtained coarse parametric model by adding facial details with displacement. Richardson *et al.* [49] and Guo *et al.* [26] regress displacement maps to reconstruct fine-grained structures in visible regions, resulting in unsatisfactory artifacts in the occlusion part. Yang *et al.* [67] make a high-resolution 3D scanned dataset and presents a pix2pix framework for displacement prediction, but still not robust to occlusions. Feng *et al.* [18] present an approach to learn animatable details without paired 3D training data, causing incorrect details and noise.

GAN inversion and Face Reconstruction. Recent years have witnessed unprecedented advances of GAN [33, 32] in achieving superior image quality with high realism. Tons of works [2, 48, 48, 1] have studied how to effectively and efficiently invert to the desired point of manifold. Researchers either directly optimize a latent code through gradient descent [13, 42] or design an encoder [4, 5, 57, 48, 2] to map to it. Gecer *et al.* [20, 21] achieves high-quality texture completion by minimizing the error of synthesized image, but consumes huge time on optimization. Furthermore, previous methods [4, 5, 57, 48, 2] mainly invert to the low-dimensional latent code in which case the perfect inversion may not lie in this space [57].

3. Methodology

We tackle the problem of single-image facial reconstruction, with the goal to recover realistic textures with fine-grained geometric details. To do so, we present **Reconstruction by Encoding Facial Priors (ReEnFP)** pipeline, where the facial reconstruction procedure is achieved by mapping input image to latent space of facial prior network. The whole architecture is depicted in Fig. 3. In this section, we first introduce the formulation of facial appearance prior by dual learning of pseudo-albedo and pseudo-illumination (Sec.3.1), then we briefly provide the dataset and design of geometry generator (Sec.3.2). Finally, we illustrate the facial prior encoding pipeline and its training strategy (Sec.3.3).

3.1. Facial Appearance Prior Learning

The key of facial appearance prior learning is to represent person-specific textures distribution by a generator G_{tex}. Since the appearance is supposed to be independent of facial poses and expressions, we devise the generator to synthesis facial textures that comply with predefined layout according to UV parameterization of parametric model [39]. The whole training architecture is illustrated in Fig. 2. Two random noise, z_a and z_i are separately mapped to style codes w_a and w_i, accounting for p-albedo and p-illumination generation respectively. After integrating them [27], we obtain UV texture. It will be rendered to image space controlled by the pose and shape of a randomly sampled real image. Masking out backgrounds, a discriminator D is introduced to constrain rendered image close to real one by adversarial loss L_{adv}.

Multi-View Rendering. One problem of current rendering strategy is that the generator might cheat by only synthesizing visible appearance for a given 3D mesh. To avoid incomplete UV texture map generation, we render texture UV under multiple views to ensure reasonable generation of different angles. Specifically, we enforce multiple occurrence of the same texture code within a batch with probability p_{mv}. The p_{mv} is empirically set to 0.5 in our experiments.

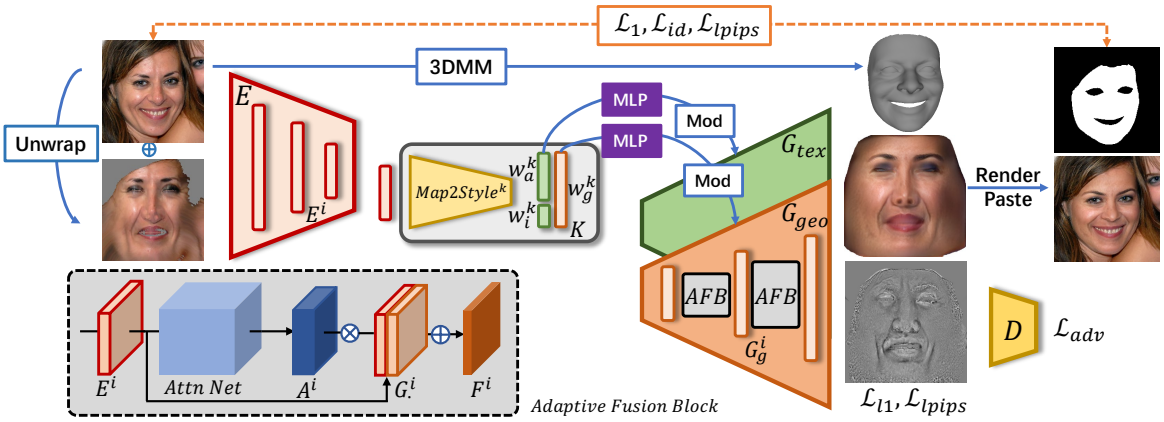

Figure 3: Overall architecture of **Reconstruction by Encoding Facial Priors (ReEncFP)** pipeline. We encode input image to latent space of well-trained texture and geometry generators. *Adaptive Fusion Block (AFB)* plugged in generators allows for comprehensive utilization of spatial features.

Dual Learning of Pseudo Albedo and Illumination. Although concentrating on synthesizing following fixed UV parameterization, our model still needs to handle infinite illumination conditions. In order to ease the texture learning difficulty, we propose to implicitly learn a set of *pseudo* albedo and illumination[1]. Notably, though they are not designed for perfect disentanglement, such formulation still benefits learning procedure. Inspired by the effectiveness of ratio-image based relighting [27], we represent p-illumination by ratio image, which will be multiplied to the Y channel of p-albedo and obtain the final UV texture.

Modulation as Shading. Recall that in Spherical harmonic lighting [47], irradiance is a quadratic polynomial of the coordinates of the (normalized) surface normal $\mathbf{n}$, which is formulated as

$$E(\mathbf{n}) = \mathbf{n}^t M \mathbf{n}. \tag{1}$$

The M is a symmetric 4x4 matrix, depending on 9 lighting coefficients. If we consider surface normal as a feature map, the rendering process is simply convolution by a kernel M determined by a few parameters. Naturally, this motivates us to simulate it with weight modulation [34], in which case we expect the network to automatically learn to synthesize p-illumination from extracted facial structure feature. Hence, we design an illumination latent code z_i with only 9 dimensions, responsible for modulating common features to generate p-illumination images.

P-Albedo Regularization. To constrain constant skin tone and eliminate highlights or shadows, we apply symmetric loss $\mathcal{L}_{sym}$ and standard deviation loss $\mathcal{L}_{skin}$ on blurred texture map following previous work [39].

[1]They are referred to as p-albedo and p-illumination for clearer presentation.

Figure 4: **Geometry Refinement with Displacement.** From left to right list input image, coarse shape by 3DMM, predicted displacement and refined shape with fine-scale details.

P-Illumination Regularization. We further enforce the smoothness of predicted ratio image by TV loss [50] to avoid learning facial structure information. Formally,

$$\mathcal{L}_{reg} = \sum_{i,j} |r_{i+1,j} - r_{i,j}| + |r_{i,j+1} - r_{i,j}|. \tag{2}$$

where r denotes the predicted ratio image.

3.2. Facial Geometry Prior Learning

Due to the limited representation power of parametric face models, they are incapable of recovering facial details such as wrinkles and dimples. Thus, the geometry prior attempts to offer reasonable detail information for refinement of coarse shape. Concretely, we describe facial geometric details with a displacement along the normal direction of each pixel in UV map. But unlike the texture generator with numerous in-the-wild facial images for training, there are very limited data with geometry details in real world. Thus, we train the geometry generator with collected displacement datasets [67].

Geometry Refinement with Displacement. Let d be our predicted displacement values of pixels in UV map. We first rasterize coarse face mesh to its UV space and represent its shape by position map. Then the updated position $\mathbf{p}'_{uv}$ in

UV map can be written as

$$\mathbf{p}'_{uv} = \mathbf{p}_{uv} + d\mathbf{n}_{uv}, \tag{3}$$

where $\mathbf{p}_{uv}$ and $\mathbf{n}_{uv}$ denote positions and normal directions. Projecting it back to the vertices of face mesh, we end up with a refined geometry with facial details as shown in Fig. 4.

3.3. Encoding Framework

The encoding framework targets to reverse an input image back to correct point of prior manifold, thereby recovering consistent facial appearance and geometric details. As illustrated in Fig. 3, input information will be injected into the fixed generator G_{tex} and G_{geo} in two flows. On one hand, we map image feature to K latent codes $w^k_{\cdot}$ with $1 \leq k \leq K$ in extended $\mathcal{W}+$ space following typical GAN inversion methods [2]. On the other hand, we directly fuse extracted spatial features to decoders by Adaptive Fusion Block (*AFB*) in a pyramid manner.

Shared Image Feature Extractor. The image encoder aims to extract image feature pyramids, capturing facial appearance information while being robust to various poses and expressions. Instead of solely feeding the input image, we also concatenated its unwrapped texture acquired by 3DMM [17]. Though it is inevitably contaminated by noisy illumination and occlusion, we believe its visible part offers aligned texture information and high-frequency geometric details. They are fed into an encoder E to extract spatial features and concatenated as E^i for later use.

Map2Style Block. With extracted spatial features pyramids, we use K different Map2Style blocks to obtain its corresponding style vector. Particularly, each Map2Style block gradually down-sample the 8×8 feature map in the lowest level of pyramids to a 1×1 latent code $w^k_{\cdot}$, which dominates the main direction of fixed facial priors.

Adaptive Fusion Block. Feature fusion [63] with prior network has proved effective on balancing prior and input information. Similarly, we introduce Adaptive Fusion Block (*AFB*) for integration of identity-specific details and facial prior regularization. Specifically, given the intermediate feature E^i, our network will learn an attention mask A^i to pay high attention to E^i for visible parts and resort to prior feature G_i when it comes to occlusion or blurry condition. After the weighted sum of E^i and G^i, the obtained feature F^i will be forwarded to the next level of pyramids.

Unsupervised Training of Texture Encoding. After encoding input image to latent space with Map2Style and Adaptive Fusion Block, our texture generator synthesize a UV texture. To constrain it containing consistent characteristics as input, we expect to minimize the difference between its rendered image and the input. We only consider differences on face regions M_{face} obtained with a pre-trained face parsing network [68] trained on CelebA-MaskHQ [36]. Firstly, we utilize masked pixel-wise $\mathcal{L}_1$ loss, which is formulated as

$$\mathcal{L}_1 = \frac{M_{face}M_{proj}\left\|I - I'\right\|_1}{M_{face}M_{proj}}, \tag{4}$$

where M_{proj} denotes the visible region projected by face mesh, and I and I' are the input and rendered image. Additionally, for perceptual similarities, we employ LPIPS [70] loss

$$\mathcal{L}_{lpips} = M_{face}M_{proj}\left\|F(I) - F(I')\right\|_2, \tag{5}$$

where $F(\cdot)$ represents perceptual feature extractor. Moreover, to encourage the reconstructed face to share same identity with input image, we incorporate a dedicated recognition network, ArcFace [15], to measure their cosine similarity. The identity-preserving loss $\mathcal{L}_{id}$ is defined as

$$\mathcal{L}_{id} = 1 - < F(I), F(I') >, \tag{6}$$

where $F(\cdot)$ indicates feature extractor of ArcFace [15].

The overall learning objective for the texture inversion flow can be written as follows:

$$\mathcal{L}_{tex} = \mathcal{L}_{lpips} + \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_{id}, \tag{7}$$

where the λs are balancing coefficients.

Semi-Supervised Training of Geometry Encoding. We apply $\mathcal{L}_1$ and $\mathcal{L}_{lpips}$ loss on the labeled data [67]. For model generalization, adversarial loss is added, which formulates as

$$\mathcal{L}_{adv} = \min_{\mathrm{G}}\max_{\mathrm{D}}\mathbb{E}_I[\log \mathrm{D}(I)] + \mathbb{E}_{I'}[\log(1 - \mathrm{D}(\mathrm{G}(I')))]. \tag{8}$$

Thus, the overall training loss for geometry encoding flow is defined as:

$$\mathcal{L}_{geo} = \mathcal{L}_1 + \lambda_1\mathcal{L}_{lpips} + \lambda_2\mathcal{L}_{adv}. \tag{9}$$

We only use adversarial loss $\mathcal{L}_{adv}$ when there is no labeled displacement map.

4. Experiments

4.1. Experimental Settings

Datasets. For diverse and high-fidelity appearance embedding, the texture generator is trained with FFHQ [33] and CelebA-HQ [31], which contain 70,000 and 30,000 images with a resolution of $1{,}024^2$, respectively. The geometry generator is trained by Facescape [67]. It consists of 888 people with 17,760 displacement maps covering various expressions while 360 identities are not processed for privacy protection. We use a combination of these three datasets

Table 1: **Quantitative comparisons of texture on CelebA [40] test split.** Note that our method achieves higher similarity scores compared to previous methods. [17, 14]

Method	[17]	[14]	**Ours**
L_1 distance ↓	0.052	/	**0.025**
PSNR ↑	26.58	22.9∼26.5	**30.70**
SSIM ↑	0.826	0.887∼**0.898**	0.895
LightCNN ↑	0.724	/	**0.859**

Table 2: **Quantitative comparisons of texture on LFW [28] .** We achieve highest identity retrieval on a dataset of 13,000 photos containing over 5,000 identities than existing methods [62, 22].

Method	Rand	[62]	[22]	**Ours**
R@1 ↑	0.0002	0.001	0.16	**0.54**
R@5 ↑	0.001	0.002	0.51	**0.71**

Table 3: **Quantitative comparisons of texture on test split of FFHQ [33]**

Method	OSTeC	AvatarMe	**Ours**
LightCNN ↑	0.8093	0.6095	**0.8131**

to train our main architecture. We also conduct qualitative and quantitative comparison with existing approaches on Mofa [56], CelebA [40] and LFW [28].

Implementation Details To encourage clean facial appearance learning, we train both prior generators with manually screened images not involving heavy occlusions such as glasses and masks. Prior Networks are trained at resolution 512^2 following default hyper-parameters and losses of StyleGAN2 [34]. For texture generator, the pre-trained parameters of discriminator are loaded and lower layers are frozen for fast convergence. In the main pipeline, all input images are of size 224^2 while outputs are at resolution 512^2. For the shared image feature extractor E, we borrow blocks of style encoder in StarGAN v2 [12] with initialization by their pretrained weights due to the robustness of their setting in various poses and expressions. The architecture design of Map2Style block is quite similar to the encoder in Restyle [2]. The λs are empirically set to 1. Our models are implemented by PyTorch [46] with four 32 GB Tesla V100 GPUs.

Comparison Methods. We compare our method with state-of-art approaches in terms of texture and geometry. For texture, we employ Chen *et al.* [10], Deng *et al.* [17], Gecer *et al.* [21], Genova *et al.* [22] and Tran *et al.* [62] as our baselines. Specifically, Chen *et al.* [10] captures 366 high-quality scans of 122 people and exploits UNets for detailed facial synthesis. Deng *et al.* [17] leverages a robust, hybrid loss function for weakly-supervised learning to regress 3DMM coefficients. Gecer *et al.* [21] optimizes the latent code of a progressive GAN trained by 10,000 high-resolution textures. Genova *et al.* [22] proposes an unsupervised procedure in an end-to-end framework to predict 3DMM coefficients. Tran *et al.* [62] fits an expressionless model to many photographs with an iterative optimization. Besides, we also compare with very recent works, AvatarMe [35] and OSTeC [20].

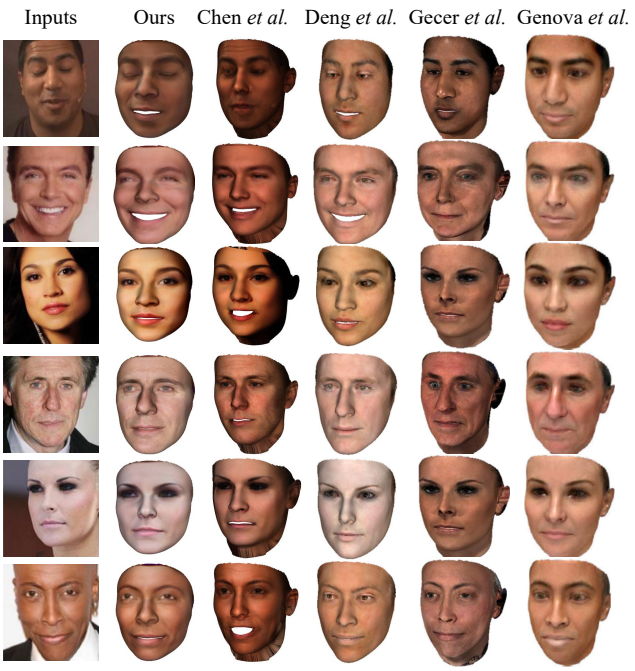

Figure 5: **Qualitative comparison of texture on Mofa [56].** Facial reconstruction results of our architecture and [10, 17, 21, 22] are listed from left to right respectively.

For the geometry, we exploit non-parametric models including PRNet [19] and LAP [71], 3DMM based approaches including Deng *et al.* [17] and 3DDFA *et al.* [25], and frameworks specifically designed for facial detail synthesis such as DECA [18] and Extreme-3D [58].

4.2. Quantitative Evaluation

Evaluation Metrics. We conduct quantitative evaluations on metrics that are commonly utilized in facial reconstruction field. $\mathbf{L_1}$ distance and **PSNR** are adopted to evaluate accuracy of reconstruction while **SSIM** [64] account for image quality. **LightCNN** [66] is employed to asses identity similarity. Retrieval rate **R@K** [22], indicating the ratio of successful retrievals in the top-K similar faces, is utilized to measure VGG-Face [45] feature distance.

Evaluation Results. The comparison results measuring the similarity between reprojection and input image are presented in Table 1. Lower L_1 distance and higher PSNR score indicate our reconstructed texture achieves better similarity in pixel level. Closer feature distance calculated by

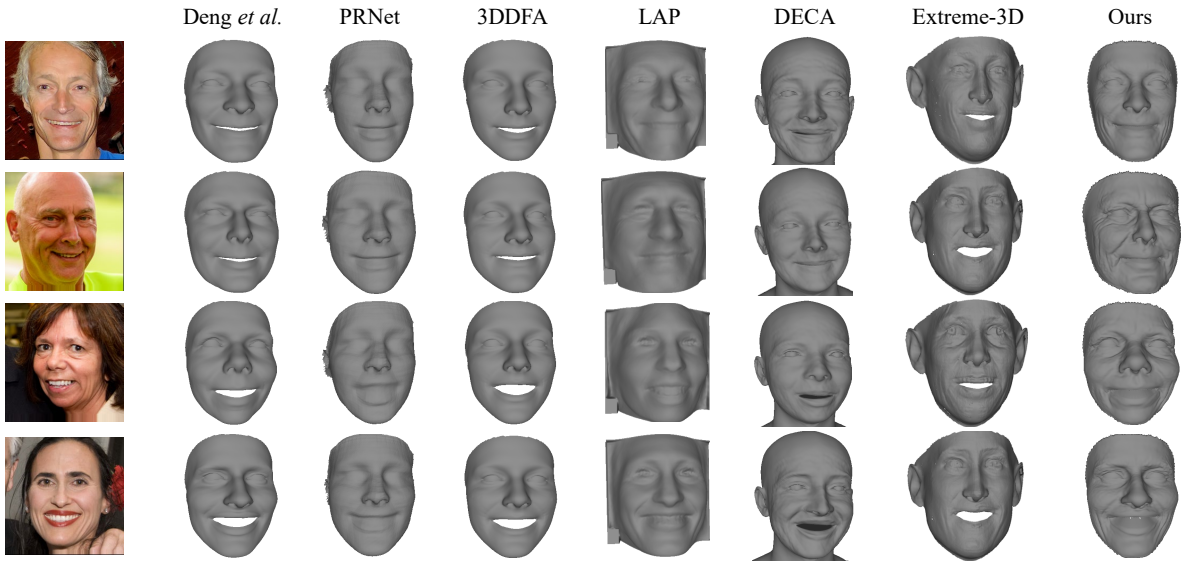

Figure 6: **Qualitative comparison of geometry with [71, 19, 25, 17, 18, 58].**

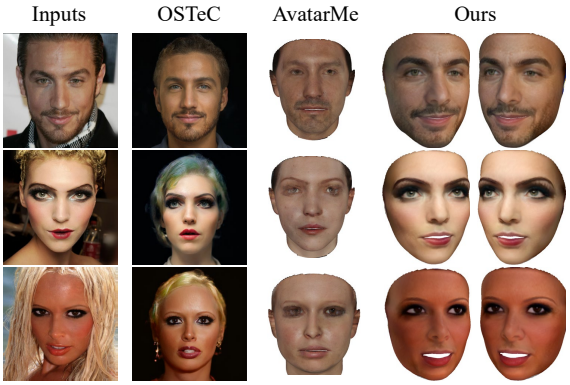

Figure 7: **Qualitative comparison of texture on test split of FFHQ with OSTeC and AvatarMe.**

face-recognition network suggests that our results preserve more identity characteristics. To further measure identity similarity, we also render our texture to neutral pose and retrieve its nearest neighbor by VGG-Face similarity. Table 2 shows the retrieval results. Higher retrieval rate implies capability of our model in retaining person-specific information. Table 3 demonstrates that our approach could even achieve slightly better identity similarity than very recent optimization-based method OSTeC [20].

4.3. Qualitative Evaluation

Qualitative Comparison of Texture. Similar to previous methods [10, 17, 21, 22], we demonstrate our facial reconstruction results[2] on Mofa [56] test dataset in Fig. 5. 3DMM-based methods [22, 17] lose detailed facial information due to limited representation ability. Gecer *et al.* [21] and Chen *et al.* [10] predict inconsistent skin colors while our results demonstrate better performance under challenging scenarios such as heavy makeup, complicated illumination and extreme expressions. We further evaluate our approaches by comparison with more recent works, OSTeC [20] and AvatarMe [35] as Fig. 7. Both their methods exhibit extreme photo-realism. However, AvatarMe omits many identity details especially their makeup. OSTeC demonstrate slight inconsistency in whole skin color. But we achieve higher identity similarity with fast inference speed.

[2]For more comparison results, readers are highly recommended to read supplementary.

Table 4: **User study on geometry recovering by Mean Opinion Scores.** Larger is higher, with the maximum value to be 5.

MOS	Detail Consistency	Identity Similarity
Deng *et al.*	2.49	4.35
LAP	3.35	3.49
DECA	3.09	4.05
Extreme-3D	3.64	3.21
EncFP(Ours)	**3.87**	**4.39**

Qualitative Comparison of Geometry. As can be seen in Fig. 6, our methods exhibit fine facial details compared to pure 3DMM-based methods [17, 25]. Without relying on synthetic data by 3DMM, non-parametric method LAP [71] reconstructs rough local structure with higher precision than PRNet [19], but still struggling at tiny details. Attempting to increase facial details, DECA [18] and Extreme-3D [58] refine coarse shape with displacement. However, their models suffer from noisy geometric artifacts while our results show correct detailed reconstruction with less ambiguity even on extreme expressions.

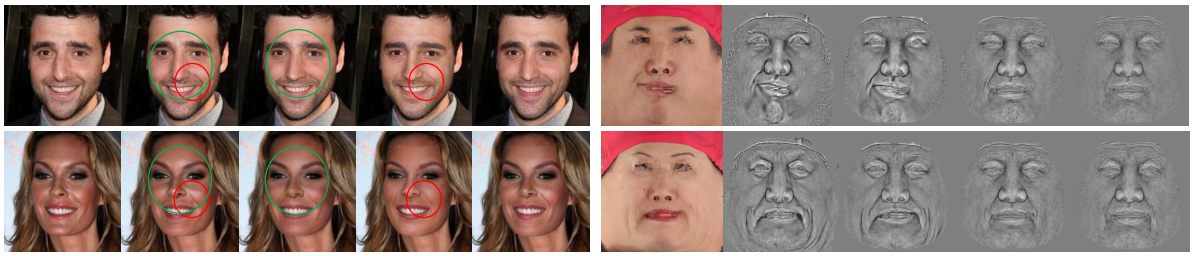

Figure 8: **Ablation study with visual results.** The facial highlights without dual learning (green circle) are weaker. The dimple without *AFB* (red circle) are shallow. For geometry, the model struggles at capturing mouth shape without *AFB*.

Table 5: **Ablation study with quantitative comparisons on CelebA [40].** The results are shown when we vary encoder design, loss function and training paradigm.

Metric	L_1 distance ↓	PSNR ↑	SSIM ↑	LightCNN ↑	CPBD ↑
psp encoder	0.032	28.14	0.858	0.812	0.324
w/o AFB	0.034	28.29	0.862	0.806	**0.327**
w/o $\mathcal{L}_{id}$	**0.025**	**30.74**	**0.895**	0.826	0.316
w/o light	0.028	30.06	0.886	0.828	0.317
w/o m-view	0.031	29.12	0.882	0.821	0.312
Full model	**0.025**	30.70	**0.895**	**0.859**	0.321

4.4. Further Analysis

User Study. We conduct a user study of 15 participants for their opinions on 25 geometric reconstruction results generated by our methods and the competing ones as Fig. 6. We adopt the widely used Mean Opinion Scores (MOS) rating protocol. The users are required to give their ratings (1-5) on the following two aspects for each video: (1) detail consistency (2) identity similarity. The results are listed in Table 4. As 3DMM based methods [17, 18] either produce over-smoothed results or cause noisy displacements, their scores on detail consistency are reasonably low. But participants believe they acquire higher identity similarity since they look natural and close to input. Our pipeline refines geometry on top of accurate proxy [17] with dedicated architecture, thereby performing favorably in terms of both aspects.

Ablation Studies. We conduct ablation studies given several key aspects such as the encoder design, the optimization loss of encoding framework and the training paradigm of texture generator. Hence, we do experiments on 1) psp encoder 2) w/o *Adaptive Fusion Block*; 3) w/o identity-preserving loss; 4) w/o illumination representation; 5) w/o multi-view rendering. The overall settings are similar to previous quantitative comparison on CelebA, but introducing an extra **CPBD** [44] metric to evaluate the sharpness of recovered texture. The quantitative results are demonstrated in Table. 5. Directly employing an existing psp encoder [48] lead to inferior results. The LightCNN score dramatically reduces without *Adaptive Fusion Block*, validate its effectiveness in preserving identity-related information. Dropping identity loss $\mathcal{L}_{id}$ also causes lower LightCNN score, implying its crucial role in identity retaining. Furthermore, degradation results without light identifying and multi-view rendering when training prior suggest their necessity in obtaining sufficient diverse appearance embedding. One interesting fact to notice is that our model reaches best CPBD score without *AFB*. We speculate that modifying spatial feature with *AFB* may damage the prior distribution and lead to slightly worse image quality. Particularly, we demonstrate examples of visual results in Fig. 8. The reconstruction results suffer from imprecise illumination without disentangling illumination while inaccurate facial details without *AFB*.

Discussion of Limitation. The disentanglement of p-albedo and p-illumination is imperfect due to no special design of light elimination mechanism. In our work, they are solely proposed for complicated lighting conditioned texture modeling to ease the training difficulty of our generator. Further exploration on ideal albedo and illumination disentanglement require sophisticated design of joint optimization of geometric and light environment.

Discussion of Ethics. The inappropriate use of face reconstruction such as synthesizing others' portraits for commercial profit will cause violation of portrait rights.

5. Conclusion

In this paper, we propose an unified pipeline, **Reconstruction by Encoding Facial Priors (ReEnFP)**, which reconstructs photorealistic facial textures with precise identity recovery and manifests expressive geometric details. We emphasize several appealing properties of our framework: 1) We achieve high-quality texture reconstruction and fine-grained facial detail expression without requiring large-scale 3D scans or long-time optimization. 2) For superior facial appearance prior modeling, we exploit a large corpus of diverse in-the-wild images to train a style-based generator through dual learning of p-illumination and p-albedo. 3) The adaptive fusion strategy is proposed to encourage identity-consistent texture reconstruction and detail preservation.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

[3] Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2017.

[4] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Inverting layers of a large generator. In *ICLR Workshop*, volume 2, page 4, 2019.

[5] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019.

[6] 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.

[7] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018.

[8] Chen Cao, Vasu Agrawal, Fernando De La Torre, Lele Chen, Jason Saragih, Tomas Simon, and Yaser Sheikh. Real-time 3d neural facial animation from binocular video. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021.

[9] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[10] 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.

[11] Lele Chen, Chen Cao, Fernando De la Torre, Jason Saragih, Chenliang Xu, and Yaser Sheikh. High-fidelity face tracking for ar/vr via deep lighting adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13059–13069, 2021.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[13] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018.

[14] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[16] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.

[17] 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.

[18] 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.

[19] 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.

[20] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7628–7638, 2021.

[21] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[22] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018.

[23] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[25] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 152–168. Springer, 2020.

[26] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018.

[27] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14719–14728, 2021.

[28] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[29] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, Willem P Koppen, William Christmas, Matthias Rätsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th joint conference on computer vision, imaging and computer graphics theory and applications*, pages 79–86. SciTePress, 2016.

[30] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039, 2017.

[31] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[32] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

[33] 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.

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[35] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2020.

[36] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[37] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[38] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020.

[39] Jiangke Lin, Yi Yuan, and Zhengxia Zou. Meingame: Create a game character face from a single portrait. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[40] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[41] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11672, 2021.

[42] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. 2019.

[43] Richard T Marriott, Sami Romdhani, and Liming Chen. A 3d gan for improved large-pose facial recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13445–13455, 2021.

[44] Niranjan D Narvekar and Lina J Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, pages 87–91. IEEE, 2009.

[45] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[47] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001.

[48] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[49] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017.

[50] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[51] 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.

[52] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance

and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018.

[53] Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–24, 2019.

[54] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[55] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3361–3371, 2021.

[56] 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.

[57] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[58] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018.

[59] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019.

[60] 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.

[61] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):157–171, 2019.

[62] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017.

[63] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021.

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[65] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020.

[66] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[67] 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.

[68] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[69] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019.

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[71] 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.

[72] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5911–5920, 2020.