Learning Neural Proto-face Field for Disentangled 3D Face Modeling In the Wild

Zhenyu Zhang[1], Renwang Chen[1], Weijian Cao[1], Ying Tai[1*], Chengjie Wang[1,2*]
Tencent Youtu Lab, Shanghai, China[1]
Shanghai Jiao Tong University, Shanghai, China[2]
zhangjesse@foxmail.com
renwangchen, weijiancao, yingtai, jasoncjwang@tencent.com

Abstract

Generative models show good potential for recovering 3D faces beyond limited shape assumptions. While plausible details and resolutions are achieved, these models easily fail under extreme conditions of pose, shadow or appearance, due to the entangled fitting or lack of multi-view priors. To address this problem, this paper presents a novel Neural Proto-face Field (NPF) for unsupervised robust 3D face modeling. Instead of using constrained images as Neural Radiance Field (NeRF), NPF disentangles the common/specific facial cues, i.e., ID, expression and scene-specific details from in-the-wild photo collections. Specifically, NPF learns a face prototype to aggregate 3D-consistent identity via uncertainty modeling, extracting multi-image priors from a photo collection. NPF then learns to deform the prototype with the appropriate facial expressions, constrained by a loss of expression consistency and personal idiosyncrasies. Finally, NPF is optimized to fit a target image in the collection, recovering specific details of appearance and geometry. In this way, the generative model benefits from multi-image priors and meaningful facial structures. Extensive experiments on benchmarks show that NPF recovers superior or competitive facial shapes and textures, compared to state-of-the-art methods.

1. Introduction

3D face reconstruction is a long-standing problem with applications including games, digital human and mobile photography. It is ill-posed in many cases requiring strong assumptions e.g., shape from shading [99]. With the 3D Morphable Model (3DMM) [10] proposed, such a problem can be solved by fitting parameters to the target faces [67, 68, 107]. Recently, deep-learning methods [22, 25, 43, 64, 105] are proposed to regress 3DMM parameters from input images. These approaches are then improved by non-linear modeling [24, 29, 79, 81, 84, 94] and multi-view consistency [7, 15, 76, 90]. Besides 3DMM methods, recent efforts [91, 100] attempt to model 3D face without shape assumptions. These non-parametric methods have potential ability to improve the modeling quality beyond 3DMM.

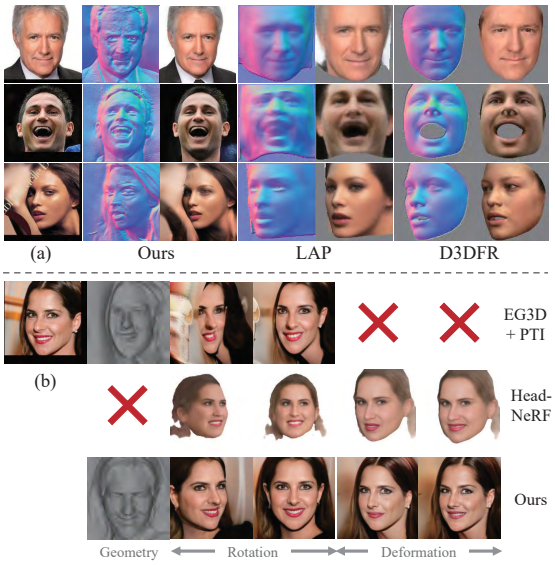

Figure 1. (a) Comparison with graphics-renderer-based methods LAP [100] and D3DFR [20]. Our method models geometry details and photo-realistic texture. (b) Results of neural rendering methods EG3D [13] + PTI [66], HeadNeRF [34] and our method. Our method produces high-quality geometry, robust texture modeling under rotation and deformation.

Although the aforementioned methods achieve impressive performance, they also have obvious drawbacks. On the one hand, as the parametric models are usually built from a small amount of subjects (e.g., BFM [58] with 200 subjects) and rigidly controlled conditions, they may be fragile to large variations of identity [106], and have limi-

*Chengjie Wang and Ying Tai are corresponding authors

tations on building teeth, skin details or anatomic grounded muscles [23]. On the other hand, all of these methods depend on graphics renderers [42, 44, 46] in the analysis-by-synthesis fitting procedure, and thus yields hand-crafted approximation or ill-posed decomposition on intrinsic clues. Hence, as illustrated in Fig. 1-(a), these methods struggle to produce photo-realistic texture or geometric details.

Against these limitations, efforts are made to use a neural renderer such as StyleGAN [38, 39] to model faces by inverting the corresponding images [1, 2] into $\mathcal{W}$ space. Existing methods [11, 18, 59, 62, 92] mainly learn to embed 3DMM coefficients to implicitly leverage 3D clues, but they have difficulty achieving precise 3D controls due to their entangled image formation. To disentangle neural rendering, recent works [13, 14, 34, 54] employ explicit 3D pipelines, e.g., Neural Radiance Field (NeRF) [52] into the StyleGANs' framework, so that face shapes and camera views can be extracted. In this way, precise 3D controls and detailed geometry can be obtained. However, these methods still show fragile performance under challenging conditions as shown in Fig. 1-(b). When confronting large poses, extreme appearance or lighting, the lack of facial priors disturbs the reconstruction and results in severe distortions. This is due to the essentially overfitting objective of inverting single target image, where the geometry ambiguity is unavoidable.

On top of this, one solution is to leverage reliable priors, e.g., multi-image consistency as a complement. While NeRF provides a natural paradigm to dig multi-view cues, it requires fully constrained images that are difficult to obtain. Even conditioned by style codes [13, 14, 54], there is no direct way to build 3D faces from unconstrained portrait collections in such a neural rendering mechanism. In this work, we present a novel Neural Proto-face Field (NPF) for unsupervised robust 3D face modeling, where ID, expression and scene-specific details can be disentangled from in-the-wild photo collections. To aggregate ID-aware cues, NPF leverages uncertainty modeling to extract multi-image priors and recovers a face prototype with ID-consistent face shape. To disentangle the expression, NPF then learns appropriate representations to deform the prototype, constrained by a expression consistency loss. In this way, the learned face shape is properly blended to avoid geometric ambiguity. Finally, to recover the scene-specific details, NPF is optimized to fit a target image in the collection. The robustness of fitting is guaranteed by a geometry and appearance regularization. As shown in Fig. 1-(b), NPF makes the generative method benefit from multi-image priors in unconstrained environments, and produces high-quality 3D faces under challenging conditions.

In summary, our contributions are as follows:

1) A novel Neural Proto-face Field (NPF) is proposed to disentangle ID, expression and specific details from 3D face modeling, which uses in-the-wild photo collections to benefit the 3D generative model under challenging conditions.

2) With a novel face prototype aggregation method, NPF integrates multi-image face priors against the large variations in unconstrained environments.

3) With a series of novel consistency losses, NPF is well fit to specific scenes with personalized details, based on the guidance of face prototypes.

Methods	Rendering	Pipeline	Multi-view
EMOCA [81], DECA [24], Unsup3D [91]	Graphics	Disentangled	×
LAP [100], FML [76], MVF [90]	Graphics	Disentangled	✓
DFG [18], StyleRig [77], StyleFlow [3]	Neural	Entangled	×
Pi-GAN [14], StyleSDF [54], EG3D [13]	Neural	Disentangled	×
Ours	Neural	Disentangled	✓

Table 1. Discussion with selected existing methods.

2. Related Works

In Table 1, we make a discussion on existing methods. NPF benefits from neural rendering and avoids hand-crafted approximations of graphics renderers. In contrast to neural rendering methods, our approach has explicit 3D pipelines, and leverages multi-image consistency in the wild.

3D Face Reconstruction: As a long-standing problem, the studies mainly start from the pioneer work 3DMM [10]. With the parametric model, early works try to find suitable parameters via optimization [67, 68, 107], while recent approaches [25, 64, 105, 106] leverage deep neural networks to regress the results from input images. With the differentiable renderers proposed, efforts are made on aspects of unsupervised learning [30, 65, 81], improving the non-linear feasibility [16, 21, 24, 29, 81, 84, 104] and multi-view consistency [7, 15, 76, 90]. More recent works attempt to learn complete 3DMM basis [79] or implicit functions [94, 102] which brings new possibilities to this topic.

Beyond 3DMM, non-parametric models are also developed by data-driven supervised training [4, 35, 97]. With shape-from-shading algorithm [99], unsupervised methods are proposed including SFS-Net [70], Unsup3D [91] and LAP [100]. Recently, Gan2Shape [55] and LiftedGAN [72] try to distill knowledge from 2D GANs for 3D reconstruction. De3D [89] de-renders more intrinsic factors based on reliable priors. Different from the these discussed methods, NPF benefits 3D face modeling via neural rendering mechanisms, thus provides better performance on details, resolution and non-facial objects.

Neural Scene Representation: Neural scene representation is a novel way to parameterize signals, and attracts more attention on learning geometry [12, 49, 51, 74, 93]. Recent effort Neural Radiance Field (NeRF) [8, 9, 52] shows impressive performance on recovering appearance and shapes from multi-view images. Based on NeRF, 3D face can also be modeled [28, 61, 86, 87] from a few or single views. Recent efforts are developed to perform reconstruc-

tion against non-rigid deformation [56,57,60]. Conditioned on specific expressions, NeRF can also be used to realize editing [5, 36] or 4D facial avatars [27, 31, 103]. In summary, these methods depend on fully constrained environments, i.e., the used images have to be captured at a same time. In contrast, our method leverages face consistency from in-the-wild photo sets that are easy to collect, and addresses large variations of lighting, appearance or artifacts.

Neural Rendering on Face Modeling: Neural rendering aims to render images using neural networks [50,78,80] with much relaxed inputs. The style-based approaches, i.e., StyleGANs [37–39] obtain the state-of-the-art performance on face generation. With the pretrained StyleGANs, face modeling can be achieved by inverting the target images [1,2,63,66]. 3D embedding (e.g., 3DMM parameter) is then employed to implicitly control the StyleGAN's prediction on pose, identity and lighting [3,18,77]. However, they cannot guarantee the robustness on physical perspective, nor obtain explicit geometry. More recent works attempt to employ 3D pipelines, such as NeRF or deferred neural rendering [82], into StyleGANs [11, 13, 14, 19, 32, 53, 54, 75, 101]. While these methods achieve more precise controls and facial geometry prediction, they are fragile to large pose and ambiguous appearance due to the single-image fitting. In contrast, NPF well leverages multi-image consistency as complementary priors and shows more robust performance.

3. Preliminary

Exploiting the consistency of multiple images, especially from unconstrained environments, is non-trivial for neural rendering methods because there is no explicit topology that can be shared within the image set. While NeRF provides a natural paradigm to dig consistency from different views, it requires photos taken at the same time or constrained scene. Recent 3D-aware generative methods [13, 14, 32, 54] show that NeRF can represent different faces conditioned on style codes. Such an evidence makes it possible to learn common hybrid 3D representations from in-the-wild photo collections. Without loss of generality, we build our model on EG3D [13] as it achieves the state-of-the-art performance.

EG3D represents the 3D scene as a tri-plane denoted as:

$$\mathbf{F}_{xy}, \mathbf{F}_{xz}, \mathbf{F}_{yz} = \Phi_g(\mathbf{s}), \quad \mathbf{s} = \{s^j\}_{j=1}^{M}, \tag{1}$$

where $\mathbf{F} \in [H, W, C]$ is the orthogonal feature plane along different axis, and Φ_g is the StyleGAN generator. $\mathbf{s} \in \mathbb{R}^{14\times512}$ is the style code with stages $M = 14$, generated from a random noise $\mathbf{z}$ by a mapping network. Then any 3D position $x \in \mathbb{R}^3$ can be projected onto each of the three feature planes, retrieving its feature $f = f_{xy} + f_{xz} + f_{yz}$ via bilinear interpolation. The corresponding density σ and color $\mathbf{c}$ is predicted by a tri-plane decoder Φ_d from f. In this way, given a camera pose p, origin $\mathbf{o}$ and near/far bound t_n, t_f, the pixel color $\mathbf{C}$ for a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ can be obtained using volume rendering equation [48]:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt, \tag{2}$$

where $T(t) = \exp(\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds$, and t is the sampled point along each ray. Using Eqn. 2, EG3D renders a low-resolution image $\mathbf{I}_l$ as well as a feature image $\mathbf{I}_f$. $\mathbf{I}_f$ is then fed into a super-resolution module Φ_u to generate final face image $\mathbf{I}$. Adversarial loss is calculated between $\mathbf{I}, \mathbf{I}_l$ and the real image to update Φ_g, Φ_d and Φ_u. Once trained, given a target image $\mathbf{I}_t$, face modeling can be achieved by image inversion [1,63,66]. As discussed in Sec. 1, this kind of approaches suffer from ambiguous appearance and large poses. In contrast, NPF tackles the problem by further disentangling a face using multi-image priors, which is introduced in the following.

4. Methodology

In this section, we introduce the proposed Neural Proto-face Field (NPF). Our aim is digging multi-image priors to improve the robustness of 3D face modeling. This is achieved by disentangling ID-consistent shape, expression and specific details from in-the-wild photo collection. The overview is shown in Fig. 2, where NPF contains identity-aware aggregation and deformation modeling to recover face prototypes (Sec. 4.1). After NPF is learned, the target face is reconstructed by a fitting procedure (Sec. 4.2).

4.1. Neural Proto-face Field Learning

Face appearance of an identity shares consistent cues due to the invariant face geometry, even under different in-the-wild environments. Such consistency has been widely used in previous methods by sharing the coefficients [20,69,79,90] or combining the UV space [100,101]. Inspired by these efforts, we propose NPF to aggregate a common hybrid 3D representation from each photo collection, which models a face prototype with ID-consistent shape from reliable priors.

Identity-aware Aggregation: Our method starts from the style code $\mathbf{s} = \{s^j\}_{j=1}^{M}$, which well conditions the radiance field of tri-planes to render a face image $\mathbf{I}$. Theoretically, $\mathbf{s}$ can be obtained from various methods [1, 2, 63, 83]. Without specially statement, we use the most direct scheme proposed in [38] to optimize $\mathbf{s}$ for real-image inversion. Given a photo collection $\{\mathbf{I}_i\}_{i=1}^{N}$ of a same ID, we obtain $\{\mathbf{s}_i\}_{i=1}^{N}$ in the $\mathcal{W}$ space. Although finding optimal solution is difficult, $\mathbf{s}_i$ still represents the face shape of $\mathbf{I}_i$, and such consistent cues should have much lower uncertainty than any other information within the image set. As a result, we propose an uncertainty-aware aggregation method inspired by [73] to extract the consistency. As illustrated in Fig. 2,

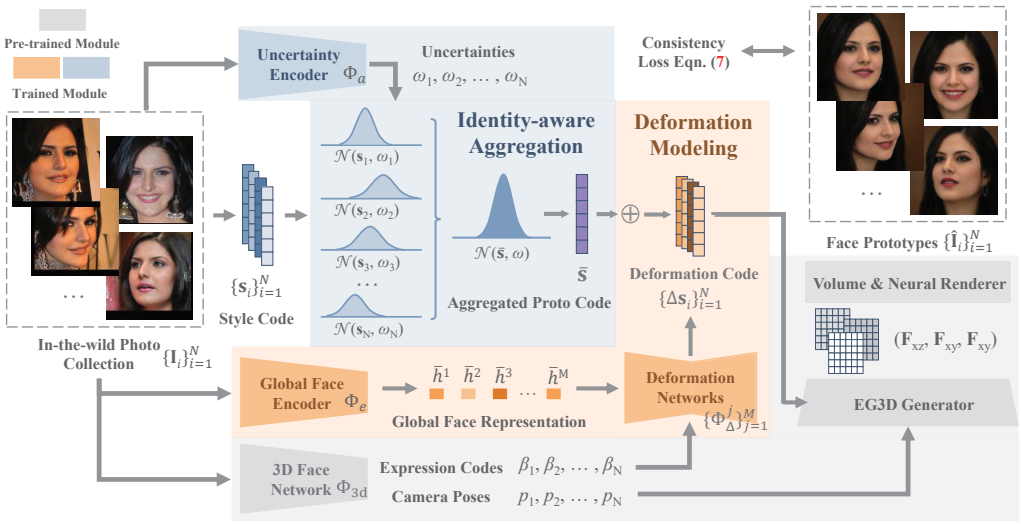

Figure 2. Overview of NPF learning method. For an in-the-wild photo collection $\{\mathbf{I}_i\}_{i=1}^N$, we first obtain its style codes $\{\mathbf{s}_i\}_{i=1}^N$ and uncertainties $\{\omega_i\}_{i=1}^N$, then perform identity-aware aggregation to get the pivotal proto code $\bar{\mathbf{s}}$. To model deformation, we use 3DMM expression parameters $\{\beta_i\}_{i=1}^N$ as conditions, and train deformation networks $\{\Phi_\Delta^j\}_{j=1}^M$ to predict deformation code $\Delta\mathbf{s}_i$. We also learn global face representation $\bar{h}^j$ at j-th stage to complement for personalized idiosyncrasies. In this way, the ID-consistent face is suitably blended on the manifold by Eqn. 5. Training by Eqn. 7, NPF well leverages the consistency within in-the-wild portrait sets.

we use an encoder Φ_a to get uncertainty $\omega_i \in \mathbb{R}^{512}$ from $\mathbf{I}_i$, and assume $\mathbf{I}_i$ to have a multivariate Gaussian representation $\mathcal{N}(\mathbf{s}_i, \omega_i)$ in the latent space. Then the aggregation is performed as:

$$\bar{\mathbf{s}} = \sum_{i=1}^{N} \frac{\omega^2}{\omega_i^2}\mathbf{s}_i, \quad \frac{1}{\omega^2} = \sum_{i=1}^{N} \frac{1}{\omega_i^2}, \tag{3}$$

where $\bar{\mathbf{s}}$ is the pivotal proto code similar to PTI [66] but from the photo collection instead of a single image. When forcing $\bar{\mathbf{s}}$ to approximate each image, the consistency which has lower uncertainty within $\{\mathbf{I}_i\}_{i=1}^N$ is remained, while other information is suppressed. In this way, the the common facial shape and appearance is well aggregated into $\bar{\mathbf{s}}$ as a face prototype. Further, compared with average pooling, the adaptively fusion algorithm of Eqn. 3 can be jointly learned with deformation to improve the modeling quality, which is introduced in the following.

Deformation Modeling: Although $\bar{\mathbf{s}}$ has the common prior of an identity, it considers no facial deformation which is suppressed after the aggregation, resulting in a mean expression. Actually, modeling deformation is crucial for face reconstruction [10]. On the one hand, if the face prototype has a different expression from the target image, then recovering the gap will be difficult in the fitting procedure. On the other hand, face images have different appearance and silhouettes due to expression variation, and thus discarding the expression brings ambiguity or conflicts to the consistent geometry. Hence, we propose deformation modeling method to disentangle expressions of face prototype.

Previous works [56, 57] learn deformation during NeRF sampling, which depends on plentiful images captured in a constrained environment. As the data we use is in-the-wild with limited numbers, using such a deformation strategy is difficult. In contrast, we model expression by modifying style codes. As illustrated in Fig. 2, we first extracts the 3DMM expression parameters $\{\beta_i\}_{i=1}^N, \beta_i \in \mathbb{R}^{64}$ and camera poses $\{p_i\}_{i=1}^N$ from a pretrained 3D face network Φ_{3d} [20], which provides guidance on tuning $\bar{\mathbf{s}}$. Note that, $\{\beta_i\}_{i=1}^N$ cannot well represent the personalized expression with unique idiosyncrasies, as it is unrelated to the identity. Hence, we extracts global image representations from $\{\mathbf{I}_i\}_{i=1}^N$ to complement for the characteristics. We use a similar encoder Φ_e as PSP [63] to get the representation $h_i^j \in \mathbb{R}^{512}, j = 1, 2, ..., M$ and the corresponding uncertainty μ_i^j from $\mathbf{I}_i$. Then following Eqn. 3, we fuse the $\{h_i^j\}_{i=1}^N$ using $\{\mu_i^j\}_{i=1}^N$ to get the global representation $\bar{h}^j$ at j-th stage. With the global $\bar{h}^j$ of the collection, we propose deformation networks $\{\Phi_\Delta^j\}_{j=1}^M$ to predict the deformation code on each stage, which can be defined as:

$$\Delta s_i^j = \Phi_\Delta^j(\bar{h}^j, \beta_i). \tag{4}$$

Denoting $\{\Delta s_i^j\}_{j=1}^M$ as $\Delta\mathbf{s}_i$ for simplification, the personalized deformation of $\mathbf{I}_i$ can be achieved via modifying the pivotal proto code:

$$\hat{\mathbf{s}}_i = \bar{\mathbf{s}} + \Delta\mathbf{s}_i. \tag{5}$$

Finally, the face prototype $\hat{\mathbf{I}}_i$ can be obtained from $\hat{\mathbf{s}}_i$ and p_i using pretrained EG3D networks. In this way, $\hat{\mathbf{I}}_i$ suffers from less appearance conflicts to $\mathbf{I}_i$. Further, by jointly

learning of $\Delta\mathbf{s}_i$ and $\bar{\mathbf{s}}$, the consistent face shape within a collection is adaptively blended against non-rigid ambiguity, and have more precise geometry and texture.

Training Loss: Note that we do not require $\hat{\mathbf{I}}_i$ to be exactly the same as $\mathbf{I}_i$, but have a proper expression and a consistent blended shape among $\{\mathbf{I}_i\}_{i=1}^N$. To achieve such an aim, we propose a series of consistency losses to learn Φ_a, Φ_e and Φ_Δ. The reconstruction loss is defined as $\mathcal{L}_{re} = \mathcal{L}_{lpips}(\hat{\mathbf{I}}_i - \mathbf{I}_i) + \mathcal{L}_{L2}(\hat{\mathbf{I}}_i - \mathbf{I}_i)$, where $\mathcal{L}_{lpips}$ is the perceptual loss [98] and $\mathcal{L}_{L2}$ is the L2 distance. To constrain the expression of $\hat{\mathbf{I}}_i$, we propose an expression consistency loss using pretrained Φ_{3d}:

$$\mathcal{L}_c(\hat{\mathbf{I}}_i, \beta_i) = \frac{1}{\Omega}|\Phi_{3d}(\hat{\mathbf{I}}_i) - \beta_i|, \tag{6}$$

where Ω is the normalization factor. To prevent distortion, we limit the deformation code via a regularization as $\mathcal{L}_d = \|\Delta\mathbf{s}_i\|_2^2$. The final loss function is denoted as:

$$\mathcal{L}_{proto} = \sum_{i=1}^{N} \mathcal{L}_{re}(\hat{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_c \mathcal{L}_c(\hat{\mathbf{I}}_i, \beta_i) + \lambda_d \mathcal{L}_d(\mathbf{s}_i), \tag{7}$$

where λ_c, λ_d are the weights. When using $\mathcal{L}_{proto}$ to update Φ_a, Φ_e and Φ_Δ, the networks of EG3D are all frozen. In this way, we learn neural proto-face field from in-the-wild photo collections, and disentangle identities and expressions of the face prototypes.

4.2. Neural Proto-face Field Fitting

Once NPF is learned, we get a 3D hybrid initialization containing consistent cues of the photo collection, constrained by reliable multi-image priors. To model the precise 3D faces, scene-specific details need to be recovered. Such an aim can be achieved by tuning the generator using pivotal tuning (PTI) algorithm [66]. However, PTI has obvious drawbacks of two aspects: 1) its pivotal code has no multi-view prior; 2) its tuning procedure easily overfits to a single image. The former one has been addressed by NPF. For the latter, we propose a consistent fitting method to overcome the monocular ambiguity.

Multi-image Warm-Up: Fitting is achieved by tuning the generator Φ_g of EG3D while freezing other networks. In the early stage of fitting, we need to prevent the generator converging to a local optimum that provides distortions due to challenging conditions. In contrast to fitting a single image, we propose a multi-image warm-up method to fit Φ_g to the photo collection $\{\mathbf{I}_i\}_{i=1}^N$. We use $\mathcal{L}_{re}$ to approach the target image set. To guarantee the geometry consistency at unseen views, we randomly sample a camera pose p' at every step, and use the volume renderer of EG3D to predict the corresponding normal maps $\hat{\mathbf{n}}_i$ and $\dot{\mathbf{n}}_i$ from the original/new weights of Φ_g, respectively. Then we calculate the

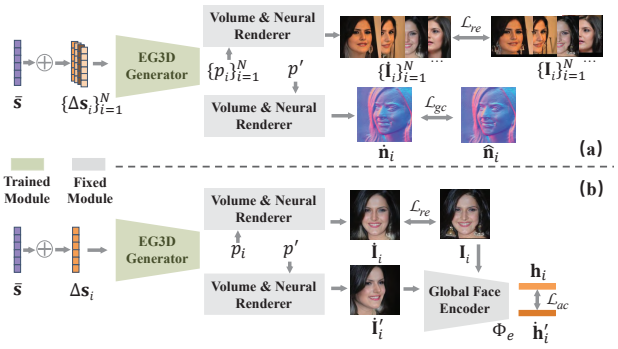

Figure 3. The proposed consistency losses of NPF fitting. (a) Multi-image Warm-up; (b) Robust Target-image Fitting.

geometry consistency loss $\mathcal{L}_{gc}(\dot{\mathbf{n}}_i, \hat{\mathbf{n}}_i) = \|\dot{\mathbf{n}}_i - \hat{\mathbf{n}}_i\|_2^2$ to constrain the robustness of facial geometry. The final objective of multi-image warm-up is denoted as:

$$\mathcal{L}_{warm} = \sum_{i=1}^{N} \mathcal{L}_{re}(\dot{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_g \mathcal{L}_{gc}(\dot{\mathbf{n}}_i, \hat{\mathbf{n}}_i), \tag{8}$$

where $\dot{\mathbf{I}}_i$ is the reconstructed image. Using $\mathcal{L}_{warm}$, we make Φ_g initialize to a solution near the optimum of each image, and suppress distortions of unseen poses.

Robust Target-image Fitting: After the multi-image warm-up, we fit Φ_g to the target image $\mathbf{I}_i$ and model specific details. Besides the reconstruction loss $\mathcal{L}_{re}$, we constrain the recovered appearance of unseen poses. Similar to $\mathcal{L}_{gc}$, we randomly sample a camera pose p' at each step and render an image $\dot{\mathbf{I}}'_i$, then encourage $\dot{\mathbf{I}}'_i$ and the target $\mathbf{I}_i$ to have similar representations via $\mathcal{L}_{ac} = \|\Phi_e(\dot{\mathbf{I}}'_i) - \Phi_e(\mathbf{I}_i)\|_2^2$. The final loss of robust target-image fitting is:

$$\mathcal{L}_{target} = \mathcal{L}_{re}(\dot{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_a \mathcal{L}_{ac}. \tag{9}$$

In this way, we improve the robustness of the scene-specific details recovering.

5. Experiment

Dataset: We first train our Φ_a, Φ_e and Φ_Δ on CelebA [45] and CASIA-WebFace [95], then fine-tune them on a high-resolution dataset CelebAMask-HQ [41]. Following [100], we organize CelebA and CASIA-WebFace using ID-labels and keep each identity with at least 6 photos. This provides 600K images with 16K identities. We select images of 12K/2K/2K identities as train/val/test set. For CelebAMask-HQ, we organize it into 24K different identities using ground truth ID-labels, and randomly select 20K/1K/3K identities as train/val/test set. Following [4, 91, 100], we perform testing on MICC [6], Photoface [96] and FG3D [26] dataset. Photoface dataset contains 12K images of 453 people with face/normal image

No.	method	p2p (mm) ↓	MAD (deg.) ↓	IDE ↓
(1)	Ours	**1.77**	**11.87**	**0.257**
(2)	Ours face prototype	1.96	12.39	0.307
(3)	w/o uncertainty = avg pooling	1.87	12.13	0.298
(4)	w/o deformation modeling	2.25	12.90	0.401
(5)	w/o Φ_e	2.01	12.46	0.350
(6)	w/o multi-image warm-up	1.84	12.30	0.301
(7)	w/o $\mathcal{L}_{gc}$	1.80	12.15	0.274
(8)	w/o $\mathcal{L}_{ac}$	1.79	11.87	0.262
(9)	single-image PTI [66]	2.35	12.10	0.482

Table 2. Comparison with Different Baselines and Settings.

pairs, and we follow the protocol of [4, 70] for testing. More details are introduced in Appendix.

Implementation Details: We build Φ_a using a similar architecture to the encoder of [91, 100]. The deformation network Φ_Δ^j of j-th stage contains 3 MLP layers to encode β_i, and another 3 MLP layers to predict Δs_i^j from the combination of $\beta_i, \bar{\mathbf{h}}^j$. We set $\lambda_c = 1, \lambda_d = 1e-3, \lambda_g = \lambda_a = 0.2$ for the losses. Following [13], the output images of the model are of 512×512, and the meshes are extracted using Marching Cubes [47] with a same size. During training, we randomly set N from 1 to 6 to adapt different sizes of a photo collection. We train the model with batch size of 4 on CelebA and CASIA-WebFace for 15 epochs, and finetune it on CelebAMaskHQ for 30 epochs. During fitting, the multi-image warm-up lasts 100 steps and robust target-image fitting lasts another 100 steps. We use Adam [40] as the optimizer, and set the learning rate as 0.0001 on a V-100 GPU. More details are in the Appendix.

Evaluation Protocol: Without special statements, we use 4-image results to compare with other methods. To evaluate the modeled texture, we calculate Structural Similarity Index (SSIM) [88], Cosine-similarity of Arcface [17] representation and RMSE of identity parameters (IDE) of 3D face network [20] between the original high-quality images and rendered ones. We render images under different poses to measure the robustness. To evaluate the geometry, we rigid-align predictions to the ground truth via ICP using hand-crafted key points. Mean Angle Deviation (MAD) of normal maps and point-to-plane (p2p) distance between the aligned prediction and ground truth are utilized as metrics. Please see Appendix for more details.

5.1. Ablation Study

Comparison with Baselines: We first analyse different settings of NPF in Table 2. We calculate NME, MAD and IDE metrics on MICC, Photoface and CelebaMaskHQ dataset, respectively. In row (1), our full method obtains the best performance in all metrics, approving the effectiveness of each proposed component. In row (2), we observe that the face prototype gets satisfactory, or even better results than several baselines, which reveals that NPF well integrates consistency from photo collection. In row (3), we replace the adaptive uncertainty with average pooling to get the proto code $\bar{\mathbf{s}}$, and the performance is reduced to some extent. This reveals our identity-aware aggregation well contributes on digging consistent cues. In row (4), the significant accuracy drop demonstrates the effectiveness of deformation modeling on avoiding expression-aware shape ambiguity. Row (5) also proves that the global face representation from Φ_e improves the quality of deformation modeling. Rows (6)-(8) reveal that the multi-image warm-up and consistency losses also improve the robustness. Finally, row (9) is the state-of-the-art single-image baseline, where it achieves even worse results than face prototype in IDE. This is due to the distortion caused by overfitting or ambiguity under challenging conditions of pose, lighting and occlusion.

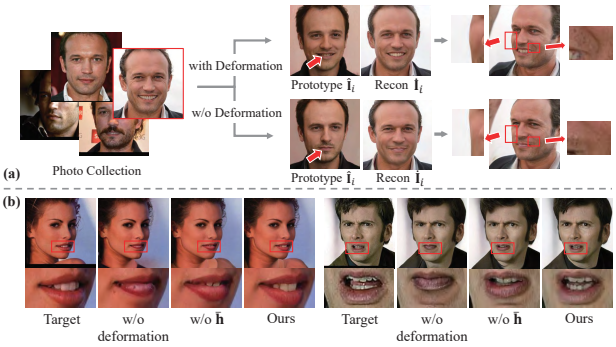

Figure 4. Analysis on the deformation modeling. (a) How the deformation modeling improves the details and shape accuracy. (b) How the settings of deformation modules influence the results.

Analysis on Deformation Modeling: We analyse how the proposed deformation modeling method improves the performance. In Fig. 4-(a), we highlight the differences between this two settings. Without deformation modeling, fusing the style codes yields $\bar{\mathbf{s}}$ to represent a mean expression on the face prototype, which is different from the target face. Hence, it brings difficulty to recover the target's expression during NPF fitting. Further, as the face shapes with various expressions show different appearance and silhouettes on the images, ignoring the deformation brings ambiguity when digging the multi-image consistency. Hence, such a model produces improper shapes and details. In contrast, modeling the deformation recovers the target's expression, the corresponding shapes and appearance effects. In Fig. 4-(b), we show the influence of different settings. Without deformation modeling, the target expression cannot be recovered. Without the global representation $\bar{\mathbf{h}}$, the expression effect is produced but not accurate enough. Finally, our full method recovers precise target deformation.

Analysis on Multi-image Consistency: We analyse how our method leverages the multi-image priors from in-the-wild photo collections. In Fig. 5-(a), we show the influence of the size N of photo collections. Compared with the single-image setting (same as PTI [66]), $N = 2$ sig-

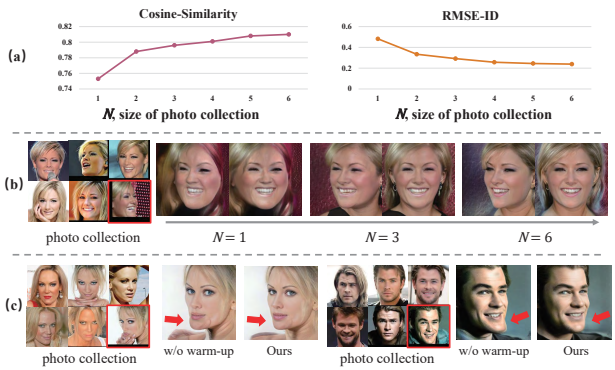

Figure 5. Analysis on the multi-image consistency. (a) (b) Performance under different photo collection size N; (c) The effectiveness of multi-image warm-up. Red box signs the target image.

		MICC		FG3D	
Method	3DMM	indoor	outdoor	LQ	HQ
Extreme3D [85]	✓	3.66	3.70	3.49	3.58
PRN [25]	✓	2.32	2.47	2.38	2.06
RingNet [69]	✓	1.93	2.02	2.08	2.02
3DDFA v2 [33]	✓	1.88	1.96	2.10	1.91
D3DFR [20] (mv)	✓	**1.66**	**1.69**	**1.90**	1.95
MGCnet [71]	✓	1.72	1.78	1.95	1.90
DECA [24]	✓	1.69	1.70	1.91	**1.89**
Cross-modal [4]	×	2.39	2.33	2.30	2.04
Unsup3D [91]	×	2.48	2.52	2.41	2.29
LAP [100] (mv)	×	2.32	2.25	2.18	2.16
PhyDIR [101] (mv)	×	2.29	2.20	2.25	2.09
Ours	×	**1.75**	**1.81**	**2.15**	**1.98**

Table 3. Point-to-plane distance (mm) on benchmarks.

nificantly improves the accuracy, revealing that even two images still provide reliable priors. With N increasing, the accuracy gets better. Fig. 5-(b) illustrates the qualitative performance. The single-image result has obvious distortion of facial shape due to the target's large pose and challenging appearance. Increasing $N = 3$ obviously suppresses the degradation, and the setting $N = 6$ obtains best performance with proper geometry. In Fig. 5-(c), we observe that during fitting, the proposed multi-image warm-up suppresses the overfitting on occlusion or shadows, boosted by the consistency from other images.

6. Comparison with State-of-the-art

Evaluation on Geometry: We first analyse the modeled geometry on MICC and FG3D dataset. We use the provided indoor/outdoor videos of MICC, low-quality (LQ) and high-quality (HQ) image sets to optimize the trained NPF. The geometry for each identity is obtained from 4 images, and calculated the average results. The comparisons are shown in Table 3, where we divide the methods into 3DMM and non-3DMM group. The tag '(mv)' means the method uses multi-view input. We observe that our method obtains the best performance among non-3DMM approaches, but higher errors than D3DFR, MGCnet and

	MAD ↓	$< 20°$ ↑	$< 25°$ ↑	$< 30°$ ↑
SfSNet [70]	$25.5_{\pm 9.3}$	43.6%	57.5%	68.7%
PRN [25]	$24.8_{\pm 6.8}$	43.1%	62.9%	74.1%
DF2Net [97] (GT)	$24.3_{\pm 5.7}$	42.2%	62.7%	74.5%
D3DFR [20] (mv)	$23.5_{\pm 6.1}$	46.1%	61.8%	73.3%
Cross-Modal [4] (GT)	$22.8_{\pm 6.5}$	49.0%	62.9%	74.1%
DECA [24]	$\mathbf{22.5}_{\pm 5.3}$	48.7%	62.3%	73.7%
LAP [100] (mv)	$23.0_{\pm 5.1}$	48.2%	63.1%	74.9%
PhyDIR [101] (mv)	$22.7_{\pm 4.3}$	**49.2%**	**63.4%**	**75.3%**
Ours prototype	$23.7_{\pm 4.7}$	45.6%	61.4%	73.0%
SfSNet-ft [70]	$12.8_{\pm 5.4}$	83.7%	90.8%	94.5%
Cross-Modal-ft [4] (GT)	$12.0_{\pm 5.3}$	85.2%	92.0%	95.6%
LAP-ft [100] (mv)	$12.3_{\pm 4.5}$	84.9%	92.4%	96.3%
PhyDIR-ft [101] (mv)	$12.0_{\pm 4.9}$	85.3%	92.7%	**96.9%**
Ours	$\mathbf{11.87}_{\pm 5.2}$	**85.9%**	**93.0%**	**96.9%**

Table 4. Facial normal evaluation on Photoface dataset.

Figure 6. Qualitative comparison on predicted facial geometry.

DECA. Note that, 3DMM is a highly reliable face shape assumption made from real 3D ground truth, and thus it fundamentally boosts the models' accuracy. Even though, our method still outperforms Extreme3D, PRN, RingNet and competes with 3DDFA v2. We further analyse the modeled facial normal on Photoface dataset in Table 4, where '-ft' means finetuning. Our prototype has already outperforms several methods. After finetuning, our method obtains best performance. Finally, we illustrate qualitative results in Fig. 6, where our method provides finer details. Compared with the EG3D [13] + PTI [66] baseline, our method is much more robust against distortion from large pose, extreme lighting and appearance.

Evaluation on Texture: We then evaluate the modeled texture on CelebAMaskHQ dataset. Besides the reconstructed image, we render images along the yaw angle in 3 ranges: $[0°, \pm 20°]$, $[\pm 20°, \pm 40°]$ and $[\pm 40°, \pm 60°]$,

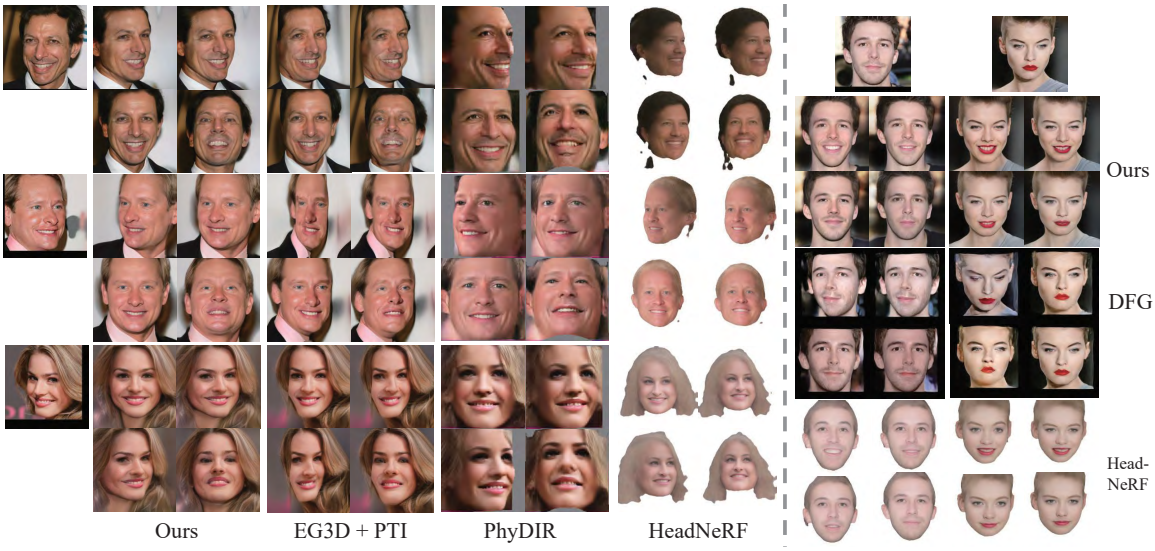

Figure 7. Qualitative comparison with EG3D [13] + PTI [66] baseline, PhyDIR [101], HeadNeRF [34] and DFG [18] on the robustness of rotation and expression editting. Our method produces robust and photo-realistic rendering results. Zooming in to see the details.

	[0°, ±20°]		[±20°, ±40°]		[±40°, ±60°]		
Method	Cos-sim	IDE	Cos-sim	IDE	Cos-sim	IDE	SSIM
DFG [18]	0.732	0.342	0.644	0.385	0.581	0.430	0.751
Unsup3D [91]	0.701	0.318	0.677	0.334	0.632	0.387	0.514
LAP [100] (mv)	0.740	0.293	0.685	0.301	0.658	0.359	0.623
PhyDIR [101] (mv)	0.826	0.267	0.792	0.294	0.745	0.328	0.880
MoFANeRF [108]	0.693	0.337	0.666	0.360	0.640	0.395	0.523
HeadNeRF [34]	0.830	0.240	0.801	0.268	0.671	0.364	0.823
EG3D + PTI	0.835	0.248	0.783	0.279	0.709	0.347	0.919
Ours	**0.842**	**0.233**	**0.806**	**0.251**	**0.750**	**0.287**	**0.921**

Table 5. Quality of rendered images on CelebAMask-HQ.

and calculate the cosine similarity and ID parameter error between them and target images. The performance is illustrated in Table 5, where DFG cannot provide satisfactory results due to its entangled image formation. For Unsup3D and LAP, the quality of texture is limited by the graphics renderer. MoFANeRF suffers from limited data diversity of constrained environments. PhyDIR and HeadNeRF suffer from large-pose degradation. The single-image baseline EG3D + PTI obtains high reconstruction performance, but fragile results under rotation due to the degraded shapes. In contrast, our method obtains the best performance in all the metrics, and shows robust and high-quality modeling results. Finally, we show qualitative results in Fig. 7 under challenging conditions of extreme lighting and pose. EG3D + PTI suffers from overfitting and produces flatten shapes, while our method obtains best performance on the rendering quality and robustness. Our method also show satisfactory ability on editing the expressions via deformation codes. More results and comparisons can be seen in the Appendix.

Limitation: Our method requires multiple images of a same person as input, which may limit the application under limitative conditions. This problem can be suppressed by searching images with similar identity and building a 'relaxed' photo collection. Further, the performance also depends on the pretrained EG3D models. As a result, extreme expressions and poses may bring degradation on accuracy. Such issues can be addressed by re-training EG3D via targeted dataset. We make the discussions in the Appendix.

7. Conclusion & Future Work

In this paper, we propose a Neural Proto-face Field (NPF) method for unsupervised robust 3D face modeling. NPF digs multi-image priors from in-the-wild photo collections to boost the 3D generative models, and well complements the single-image overfitting inversion procedure. To aggregate consistency against large variations in unconstrained environments, a novel identity-aware aggregation method is proposed to adaptively combine the style codes, and build ID-consistent face prototypes. To suppress the non-rigid ambiguity, NPF blends the consistent hybrid representation via a novel deformation modeling method. In this way, NPF obtains face prototypes containing common facial cues within the collection, and disentangles specific expressions. The final recovered 3D face is obtained via fitting NPF with consistency losses, and thus scene-specific details can be recovered. Extensive experiments demonstrate that our method models robust and detailed face shapes under challenging conditions, and recovers photo-realistic texture with pose/expression controlling. In the future, the works could be proposed to tackle the limitation such as supporting single-image input. Further, decomposing the NeRF pipeline, the texture could be disentangled into different intrinsic factors.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *CVPR*, pages 4432–4441, 2019. 2, 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020. 2, 3

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 2, 3

[4] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *CVPR*, pages 4979–4989, 2020. 2, 5, 6, 7

[5] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *CVPR*, pages 20364–20373, 2022. 3

[6] 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. 5

[7] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *CVPR*, pages 5850–5860, 2020. 1, 2

[8] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *CVPR*, pages 5855–5864, 2021. 2

[9] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 2

[10] 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, 4

[11] Marcel C Bühler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *ICCV*, pages 13890–13899, 2021. 2, 3

[12] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *ECCV*, pages 608–625. Springer, 2020. 2

[13] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 1, 2, 3, 6, 7, 8

[14] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 2, 3

[15] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. In *ECCV*, 2020. 1, 2

[16] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, pages 20311–20322, 2022. 2

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 6

[18] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, pages 5154–5163, 2020. 2, 3, 8

[19] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, pages 10673–10683, 2022. 3

[20] 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 *CVPRW*, 2019. 1, 3, 4, 6, 7

[21] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. *arXiv preprint arXiv:2103.15432*, 2021. 2

[22] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *CVPR*, pages 5908–5917, 2017. 1

[23] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2

[24] 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, 7

[25] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 1, 2, 7

[26] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786, 2018. 5

[27] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, pages 8649–8658, 2021. 3

[28] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903*, 2020. 2

[29] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for

high fidelity 3d face reconstruction. In *CVPR*, pages 1155–1164, 2019. 1, 2

[30] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, pages 8377–8386, 2018. 2

[31] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *CVPR*, pages 18653–18664, 2022. 3

[32] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[33] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 7

[34] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, pages 20374–20384, 2022. 1, 2, 8

[35] 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 *ICCV*, pages 1031–1039, 2017. 2

[36] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *CVPR*, pages 18623–18632, 2022. 3

[37] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 34:852–863, 2021. 3

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3

[39] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 3

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[41] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5

[42] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2

[43] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, pages 545–560. Springer, 2016. 1

[44] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, pages 7708–7717, 2019. 2

[45] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 5

[46] Matthew M Loper and Michael J Black. Opendr: An approximate differentiable renderer. In *ECCV*, pages 154–169, 2014. 2

[47] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6

[48] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3

[49] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 2

[50] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *CVPR*, pages 6878–6887, 2019. 3

[51] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *CVPR*, pages 4743–4752, 2019. 2

[52] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 2

[53] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 3

[54] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13503–13513, 2022. 2, 3

[55] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844*, 2020. 2

[56] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5865–5874, 2021. 3, 4

[57] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3, 4

[58] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 1

[59] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting generative adversarial renderer for face reconstruction. In *CVPR*, pages 15619–15628, 2021. 2

[60] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields

for dynamic scenes. In *CVPR*, pages 10318–10327, 2021. 3

[61] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *ICCV*, pages 5620–5629, 2021. 2

[62] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, pages 13759–13768, 2021. 2

[63] 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 *CVPR*, pages 2287–2296, 2021. 3, 4

[64] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, pages 460–469, 2016. 1, 2

[65] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pages 1259–1268, 2017. 2

[66] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 1, 3, 4, 5, 6, 7, 8

[67] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *Computer Vision*, page 59. IEEE, 2003. 1, 2

[68] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, volume 2, pages 986–993, 2005. 1, 2

[69] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, pages 7763–7772, 2019. 3, 7

[70] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, pages 6296–6305, 2018. 2, 6, 7

[71] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 7

[72] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *CVPR*, pages 6258–6266, 2021. 2

[73] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, pages 6902–6911, 2019. 3

[74] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 32, 2019. 2

[75] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and Hongsheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. In *Advances in Neural Information Processing Systems*, 2021. 3

[76] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, pages 10812–10822, 2019. 1, 2

[77] 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 *CVPR*, pages 6142–6151, 2020. 2, 3

[78] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 3

[79] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *CVPR*, pages 3361–3371, 2021. 1, 2, 3

[80] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 3

[81] 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 *ICCVW*, pages 1274–1283, 2017. 1, 2

[82] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[83] 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. 3

[84] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018. 1, 2

[85] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018. 7

[86] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[87] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2021. 2

[88] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 6

[89] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3d objects in the wild. In *CVPR*, pages 18490–18499, 2022. 2

[90] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, pages 959–968, 2019. 1, 2, 3

[91] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, pages 1–10, 2020. 1, 2, 5, 6, 7, 8

[92] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *CVPR*, pages 7710–7720, 2020. 2

[93] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34:4805–4815, 2021. 2

[94] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*, pages 12803–12813, 2021. 1, 2

[95] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[96] Stefanos Zafeiriou, Mark Hansen, Gary Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn Smith, and Lyndon Smith. The photoface database. In *CVPRW*, pages 132–139, 2011. 5

[97] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *ICCV*, pages 2315–2324, 2019. 2, 7

[98] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[99] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE TPAMI*, 21(8):690–706, 1999. 1, 2

[100] 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 *CVPR*, pages 14214–14224, 2021. 1, 2, 3, 5, 6, 7, 8

[101] Zhenyu Zhang, Yanhao Ge, Ying Tai, Weijian Cao, Renwang Chen, Kunlin Liu, Hao Tang, Xiaoming Huang, Chengjie Wang, Zhifeng Xie, et al. Physically-guided disentangled implicit rendering for 3d face modeling. In *CVPR*, pages 20353–20363, 2022. 3, 7, 8

[102] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *CVPR*, pages 20343–20352, 2022. 2

[103] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *CVPR*, pages 13545–13555, 2022. 3

[104] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, pages 1097–1106, 2019. 2

[105] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 1, 2

[106] Xiangyu Zhu, Fan Yang, Chang Yu Di Huang, Hao Wang, Jianzhu Guo, Zhen Lei, and Stan Z Li. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *ECCV*, 2020. 1, 2

[107] Xiangyu Zhu, Dong Yi, Zhen Lei, and Stan Z Li. Robust 3d morphable model fitting by sparse sift flow. In *ICCV*, pages 4044–4049. IEEE, 2014. 1, 2

[108] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. *arXiv preprint arXiv:2112.02308*, 2021. 8