# Physically-guided Disentangled Implicit Rendering for 3D Face Modeling

Zhenyu Zhang[1], Yanhao Ge[1], Ying Tai[1], Weijian Cao[1], Renwang Chen[1], Kunlin Liu[2],
Hao Tang[3], Xiaoming Huang[1], Chengjie Wang[1]*, Zhifeng Xie[4], Dongjin Huang[4]*
Tencent Youtu Lab, Shanghai, China[1]
University of Science and Technology of China[2]†
CVL, ETH Zurich, Switzerland[3]
Shanghai Film Academy of Shanghai University[4]
zhangjesse@foxmail.com lkl6949@mail.ustc.edu.cn hao.tang@vision.ee.ethz.ch
halege, yingtai, weijiancao, renwangchen, skyhuang, jasoncjwang@tencent.com
djhuang, zhifeng_xie@shu.edu.cn

## Abstract

*This paper presents a novel Physically-guided Disentangled Implicit Rendering (PhyDIR) framework for high-fidelity 3D face modeling. The motivation comes from two observations: Widely-used graphics renderers yield excessive approximations against photo-realistic imaging, while neural rendering methods produce superior appearances but are highly entangled to perceive 3D-aware operations. Hence, we learn to disentangle the implicit rendering via explicit physical guidance, while guaranteeing the properties of: (1) 3D-aware comprehension and (2) high-reality image formation. For the former one, PhyDIR explicitly adopts 3D shading and rasterizing modules to control the renderer, which disentangles the light, facial shape, and viewpoint from neural reasoning. Specifically, PhyDIR proposes a novel multi-image shading strategy to compensate for the monocular limitation, so that the lighting variations are accessible to the neural renderer. For the latter, PhyDIR learns the face-collection implicit texture to avoid ill-posed intrinsic factorization, then leverages a series of consistency losses to constrain the rendering robustness. With the disentangled method, we make 3D face modeling benefit from both kinds of rendering strategies. Extensive experiments on benchmarks show that PhyDIR obtains superior performance than state-of-the-art explicit/implicit methods on geometry/texture modeling.*

## 1. Introduction

3D face reconstruction gets increasingly attraction with applications such as digital human, games and mobile photography. The pioneering effort is 3DMM [6] which provides reliable facial priors. With this parametric model, the reconstruction can be achieved by optimization and fitting [47, 48, 79]. With the development of deep learning, recent methods [15, 18, 33, 45, 77] learn to regress 3DMM parameters from input images. Subsequent works are also proposed to contribute on non-linear modeling [17, 19, 56, 57, 59, 67, 76] and multi-view consistency [5, 9, 54, 64, 69]. Besides 3DMM based approaches, recent efforts [50, 65, 75] attempt to model 3D face without shape assumptions. These non-parametric methods have potential ability to improve the modeling quality over 3DMM limitations.

Actually, the aforementioned learning-based methods need differentiable renderers including OpenDR [36], neural mesh renderer [29], SoftRas [34] and Ray-tracing [32] for unsupervised learning. These renderers perform image formation under graphics pipelines which are well explainable. With the explicit 3D operations, the fine-grained 3D controls are naturally achieved. However, these graphics renderers yield hand-crafted approximation or ill-posed decomposition on reflectence, illumination or other 3D clues. In Fig. 1-(a), we observe the graphics-renderer-based methods [13, 32, 75] struggle to produce photo-realistic texture, which also limits their geometry reconstruction.

Against these limitations, another approach is employing a neural renderer such as StyleGAN [27, 28] to avoid approximation or ill-posed decomposition. Existing methods [7, 12, 42, 43, 66] mainly learn to embed 3DMM coefficients into StyleGAN's manifold, and constrain the generative network with 3DMM consistency. In this way, 3D controls are achieved implicitly by tuning the parameters. With StyleGAN's effectiveness, these methods show high-reality texture modeling performance. However, in Fig. 1-

*Chengjie Wang and Dongjin Huang are corresponding authors

†CAS Key Laboratory of Electromagnetic Space Information of USTC.

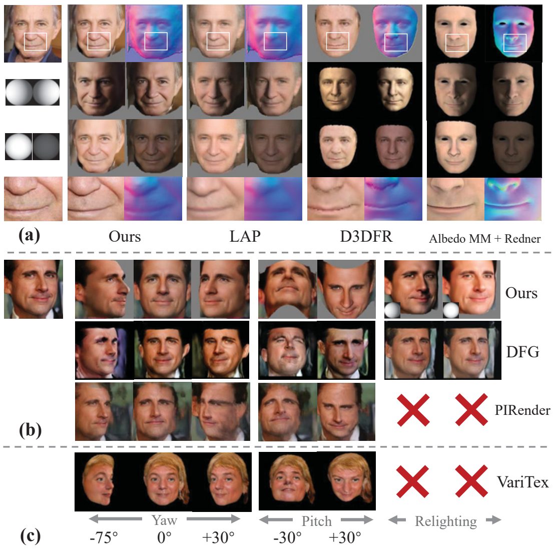

Figure 1. (a) Comparison with graphics-renderer based methods LAP [75], D3DFR [13] and Albedo MM [52] + Ray-tracing (redner) [32]. Our method models detailed facial shapes, photo-realistic texture and lighting effects. (b) Comparison with neural rendering methods DFG [12] and PIRender [43]. Our method produces more robust 3D controls and relighting results. (c) Results of 3D-aware generative method [7]. Our method well addresses real-world images and photo-realistic lighting effects.

(b), we observe that they cannot guarantee the identity, facial shape, lighting effect or texture consistency during 3D operating. The reason is due to the entangled image formation procedure. StyleGAN is trained as a 2D-aware 'black box' without 3D physical modeling. Hence, even with high-level 3D representations, the generator essentially needs to guess and simulate the exact 3D operations, which is highly indirect and complicated. Recent 3D-aware generative approaches [7, 8, 39, 42] are proposed against this problem and achieve better 3D controls. However, in Fig. 1-(c), we observe that this kinds of method cannot address real-world images nor lighting effects.

On top of these discussions, we argue that a proper rendering strategy should support **(1) explicit and fine-grained** 3D controls, **(2) a disentangled neural reasoning** for high-quality image formation and **(3) easily inverse rendering** to model faces from real images. In this paper, we propose a novel Physically-guided Disentangled Implicit Rendering (PhyDIR) framework for 3D face reconstruction. As shown in Fig. 1, by disentangling 3D physical pipelines from neural reasoning, PhyDIR achieves robust and photo-realistic 3D modeling/editing from input facial photos. The neural reasoning of PhyDIR contains a texture modeling network and a 2D-aware neural appearance renderer, while the 3D physical guidance bridges this two stages with explicit 3D pipelines. Concretely, the texture modeling network learns canonical implicit texture from input images, which avoids ill-posed intrinsic factorization. Then, PhyDIR employs facial shading and rasterization from a 3D proxy to warp the implicit texture into 2D space. Thus the fine-grained 3D controls, including facial shape, viewpoint and lighting, are explicitly modeled. Specifically, PhyDIR leverages a novel multi-image shading module to compensate for the monocular ambiguity, making the lighting variation well accessible in an unsupervised manner. After that, the neural appearance renderer takes the projected 2D texture for image formation, constrained by a series of 3D consistency losses. In this way, PhyDIR guarantees explainable 3D controls and photo-realistic image formation without hand-crafted rules. Finally, we demonstrate that with the disentangled paradigm, PhyDIR well acts as a reliable renderer to model detailed facial shapes.

| Methods | 3D Controls | Image Formation | Photo Collection |
|---|---|---|---|
| **Graphics-renderer-based** | **Explicit** | **Explainable** | |
| MOFA [57], DECA [17], Unsup3D [65]<br>LAP [75], FML [54], MVF [64] | Shape \| Pose \| Light | 3D Graphics Pipelines | ×<br>✓ |
| **Neural-rendering-based** | **Implicit** | **Entangled** | |
| DFG [12], StyleRig [55], PIRender [43] | 3DMM Parameters | 'Black Box' | × |
| **3D-aware Generative** | **Explicit** | **Disentangled** | |
| VariTex [7], Pi-GAN [8], GIRAFFE [39] | Shape \| Pose | 3D Operations +<br>2D Neural Reasoning | × |
| | **Explicit** | **Disentangled** | |
| **Ours** | Shape \| Pose \| Light | 3D Operations +<br>2D Neural Reasoning | ✓ |

Table 1. Discussion with selected existing methods.

In summary, our contributions are as follows:

**1)** A novel Physically-guided Disentangled Implicit Rendering (PhyDIR) framework is proposed to model high-fidelity 3D face. PhyDIR well integrates the advantages of graphics/neural renderers, and gets over the hand-crafted graphics rules or entangled neural image formation.

**2)** With the novel multi-image rasterizing, shading and texture mapping modules, PhyDIR guarantees fine-grained 3D controls of shape, viewpoint and lighting, as well as the photo-realistic imaging.

**3)** With a series of novel consistency losses, PhyDIR guarantees the rendering robustness under 3D operations.

## 2. Related Works

In Table 1, we make a discussion on existing face modeling methods. Compared with graphics-renderer-based methods, PhyDIR benefits from neural reasoning on photo-realistic image formation. Compared with neural-renderer-based approaches, PhyDIR tackles more explicit and explainable 3D controls. The most related works are the 3D-aware generative models. In contrast, our method addresses real-world images, multi-view consistency and light modeling which are crucial for 3D facial shape recovering.

**3D Face Reconstruction:** 3D face reconstruction is a long-standing problem [16] which can be divided into two mainstreams: *i.e.,* Parametric and non-parametric methods. The parametric methods are mainly developed from 3DMM [6]. Early works try to find suitable 3DMM parameters via optimization [47, 48, 79], while recent approaches [15, 18, 45, 77, 78] leverage deep neural networks to directly regress the parameters from input images. With the differentiable renderers proposed, efforts are made on aspects of unsupervised learning [20, 46, 57], improving the non-linear feasibility [14, 17, 19, 57, 59, 76] and multi-view consistency [5, 9, 54, 64, 69]. More recent works attempt to learn complete 3DMM basis [56] or implicit functions [67] which brings new possibilities to this topic.

For the non-parametric methods, part of recent works are developed by data-driven supervised training [3, 24, 60, 71]. Other efforts are also developed from shape-from-shading [72], including SFS-Net [50] and Unsup3D [65]. More recently, Zhang *et al.* propose LAP method [75] to leverage multi-image consistency in non-parametric paradigm. Gan2Shape [40] and LiftedGAN [51] try to distill knowledge from 2D GANs for 3D reconstruction. Different from these discussed methods, PhyDIR contributes 3D face modeling from a perspective of rendering process, in which it successfully integrates advantages from both graphics and neural rendering strategies.

**Differentiable Graphics Renderer:** Differentiable rendering is crucial for inverse graphics such as 3D face modeling, which is also a long-standing problem [23, 53]. Recent efforts such as OpenDR [36] and neural mesh renderer [29] are proposed as general pipelines, in which they approximate the primary visibility gradients for multi-triangle solution. Rezende *et al.* [26] leverages OpenGL renderer for 3D reconstruction. SoftRas [34] proposes differentiable functions upon the backward derivatives. Li *et al.* propose an edge-sampling solution for ray-tracing [32]. Cole *et al.* [10] propose an efficient surface rendering approach that supports different representations. In summary, these methods yields approximations or concessions on modeling realistic 3D faces. In contrast, PhyDIR gets rid of the limitations by integrating neural rendering, which confronts less ill-posed factorization or appearance degradation.

**Neural Rendering for Face Reconstruction:** Neural rendering methods on face modeling mainly depend on generative models such as GANs [27, 28]. General encoding methods [2, 44] introduce style vectors to control the facial attributes. By using 3D embeddings, StyleRig [55], DFG [12] and PIRender [43] implicitly control the GAN's prediction on pose, identity and lighting, but they cannot guarantee the robustness on physical perspective. Recently, approaches based on NeRF [37] are employed into GANs to accomplish 3D-aware operations [8, 39], but they cannot model high-quality geometry. More related works [7, 42] combines explicit 3D shapes with neural renderers. In contrast, PhyDIR has superiority on (1) addressing real-world images without per-image inversion; (2) explicitly modeling lighting and shadows which are also crucial for geometry learning and (3) leveraging multi-image mappings and consistency to better constrain reality and 3D robustness.

## 3. 3D Proxy Building

Implementing neural networks as renderers for face modeling is not trivial, as the rendering procedure is highly entangled. All of the existing methods [7, 12, 42, 43, 55] encode 3D priors to the forms that are accessible to the networks. Following this perspective, we first physically guide the neural renderer with a 3D proxy for high-quality appearance modeling, then leverage the learned renderer to improve the geometry reconstruction. Theoretically, the proxy can be arbitrary. Here we choose Unsup3D [65] and LAP [75] to get 3D proxy, as they require no supervision and limited priors, meanwhile have good efficiency, non-linearity and source code.

Unsup3D and LAP share a similar framework and formulation. In summary, they disentangle a facial image $\mathbf{I}$ into intrinsic factors $(d, a, \omega, l)$ comprising a depth map $d \in \mathbb{R}_+$, an albedo image $a \in \mathbb{R}^3$, a directional light $l \in \mathbb{S}^2$ and a viewpoint $\omega \in \mathbb{R}^6$, where $d$, $a$, and $l$ are in canonical space. Each factor is predicted by a separate network which we denote as $\Phi^d$, $\Phi^a$, $\Phi^\omega$ and $\Phi^l$, respectively. Then, the 3D face can be reconstructed using these factors by lighting $\Lambda$ and rasterization $\Pi$ as follows:

$$\hat{\mathbf{I}} = \Pi(\Lambda(a, d, l), d, \omega), \tag{1}$$

where $\Pi$ is achieved by a differentiable renderer [29]. They also utilize a weakly symmetric canonical space by horizontally flipping: $\hat{\mathbf{I}}' = \Pi(\Lambda(a', d', l), d', \omega)$, where $a'$ and $d'$ are the flipped version of $a, d$. Learning encourages $\mathbf{I} \approx \hat{\mathbf{I}}, \hat{\mathbf{I}}'$. Confidence maps $\sigma, \sigma' \in \mathbb{R}_+$ are predicted by a network $\Phi^\sigma$ to calibrate the loss as follows:

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) = -\frac{1}{|\Omega|} \sum \ln \frac{1}{\sqrt{2}\sigma} \exp -\frac{\sqrt{2}|\hat{\mathbf{I}} - \mathbf{I}|}{\sigma}, \tag{2}$$

where $\Omega$ is the normalization factor. The flipped version $\mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}, \sigma')$ is also calculated. We train the 3D networks $\Phi^d, \Phi^\omega, \Phi^l$ following Unsup3D and LAP, then use them to provide 3D proxy as the physical guidance for neural rendering. The details are introduced in the following.

## 4. Methodology

In this section, we introduce the proposed Physically-guided Disentangled Implicit Rendering (PhyDIR) method. Our aim is to disentangle the neural rendering process via physical guidance, making 3D face modeling benefit from both explicit/implicit strategies. The overview is shown in Fig. 2, where PhyDIR contains compositions of Implicit Texture Modeling (Sec. 4.1), 3D Physical Guidance (Sec. 4.2) and Constrained Image Rendering (Sec. 4.3) to accomplish photo-realistic texture modeling. After the learning of texture reconstruction, we then introduce how to use PhyDIR for fine-detailed geometry modeling (Sec. 4.4).

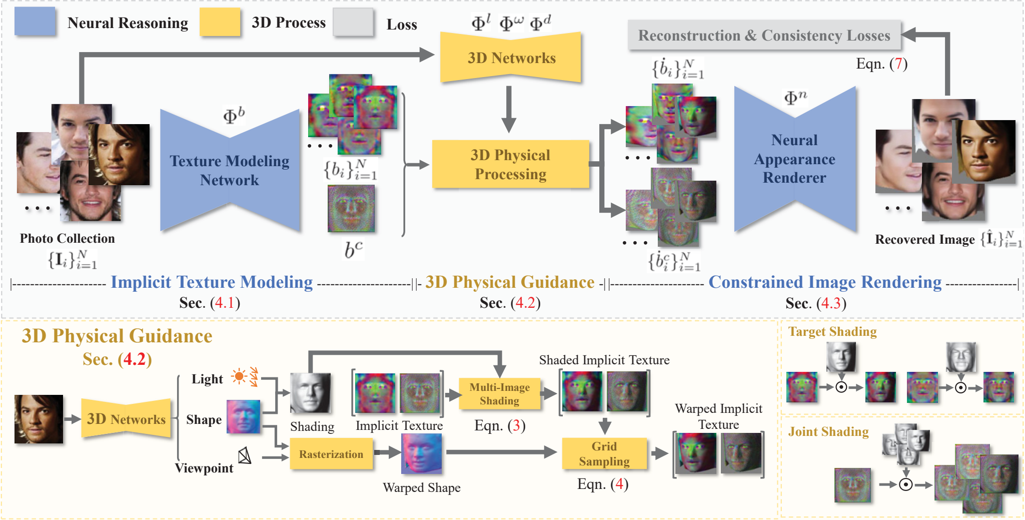

Figure 2. Overview of the proposed method. The 3D controls are explicitly disentangled from neural reasoning, making neural networks avoid tackling 3D processing and focus on 2D texture generation. For a photo collection $\{\mathbf{I}\}_{i=1}^N$ with a same identity, we first use texture modeling network $\Phi^b$ to get target/combined implicit texture $\{b_i\}_{i=1}^N, b^c$. Then, we apply explicit 3D physical processing including multi-image shading and rasterization modules to warp $\{b_i\}_{i=1}^N, b^c$ via graphics pipelines. Finally, the warped texture $\{\dot{b}_i\}_{i=1}^N$, $\{\dot{b}_i^c\}_{i=1}^N$ are fed into the neural appearance renderer $\Phi^n$ to recover $\{\hat{\mathbf{I}}\}_{i=1}^N$, constrained by different losses.

## 4.1. Implicit Texture Modeling

Illustrated in Fig. 2, we first model implicit texture from input images. Instead of learning RGB texture or albedo reflectance, our implicit texture modeling has advantages on: (1) Requiring no ill-posed factorization, (2) more abundant clues, and (3) fitness to neural rendering. Similar to neural texture [7, 58], for a photo collection $\{\mathbf{I}_i\}_{i=1}^N$ with a same identity, our texture modeling network $\Phi^b$ predicts implicit texture $\{b_i\}_{i=1}^N \in$ [h, w, c] ($c > 3$) in the canonical space. Note that, the implicit texture modeling is different from [7, 58]: First, we efficiently predict $b_i$ from $\mathbf{I}_i$ without per-image optimization. Then, our $\Phi^b$ can model multi-image consistent clues, which is introduced in Sec. 4.2. In contrast to the consistent face learning [75], our implicit texture models multi-image clues with less RGB conflicts.

## 4.2. 3D Physical Guidance

We employ explicit 3D guidance to warp the implicit texture for image formation. As discussed in Sec. 1, existing methods that embed 3DMM parameters [12, 43, 55] or style vector [2, 44] confront an entangled image formation procedure, losing robustness or fine-grained 3D controls. Further, without explicit 3D pipelines, these methods also struggle to reasonably recover 3D facial shapes, poses or lights from real images. Hence, we propose 3D physical modules to guide the neural renderer. Our 3D physical guidance contains multi-image shading and rasterization module.

**Multi-image Shading Module:** As discussed in Table 1, most neural rendering methods cannot tackle light clues. Actually, light is crucial for recovering facial details due to the shape-from-shading effects [65, 72]. Hence, we propose a novel algorithm to employ explicit shading operations on high-level neural features and in an unsupervised manner.

The multi-image shading module contains target and joint shading. For each target image $\mathbf{I}_i$, the 3D proxy networks provide canonical depth $d_i$ and light $l_i$. Then we get the shading map $\mathbf{S}_i$ by Lambertian function $f_{lam}(d_i, l_i)$. In the target shading, we directly apply shading clues by $\mathbf{S}_i \odot b_i$, which simulates the shadow condition on $\mathbf{I}_i$. However, only using the target shading cannot achieve suitable light controls. One reason is that the implicit texture cannot directly reveal the RGB lighting effects; another reason is that the neural renderer tends to overfit on $\mathbf{I}_i$, struggling to perceive lighting variations from single $b_i$ without seeing different lighting clues. As a result, we propose a joint shading module to compensate for the single-image limitation. We first adaptively combine $\{b_i\}_{i=1}^N$ by $b^c = f_{conv}([b_1, b_2, ..., b_N])$ ($f_{conv}$ is a conv-layer), then apply each $l_i$ to $b^c$. The total shading module is:

$$\hat{b}_i = \mathbf{S}_i \odot b_i, \ \ \hat{b}_i^c = \mathbf{S}_i \odot b^c, \qquad (3)$$

where $\hat{b}_i, \hat{b}_i^c$ are the shaded target/combined implicit textures for $\mathbf{I}_i$. Shaded by various lights in the photo collection, $b^c$ acts as 'roughness' and provides guidance on how to suitably effect the appearance, shadows and light intensities with different light conditions. Further, $b^c$ also enhances the texture consistency produced by a common facial shape.

**Rasterization Module:** We then use rasterization mod-

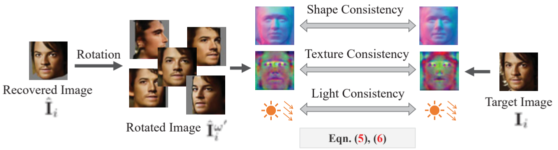

Figure 3. The proposed consistency losses to constrain the renderer's robustness on pose variations.

ule to warp and project the shaded canonical implicit texture $\hat{b}_i, \hat{b}_i^c$ to the 2D space. With the $d_i, \omega_i$ provided by 3D proxy networks, we leverage rasterization function $f_R$ (achieved by mesh-renderer [29]) to get warped depth by $f_R(d_i, \omega_i)$. Note that, although $f_R$ is approximated, the abundant clues of $b_i, b_i^c$ and neural reasoning well compensate it. $f_R$ provides a 3D grid transformation to sample $\hat{b}_i, \hat{b}_i^c$ as follows:

$$\dot{b}_i = f_{sam}(\hat{b}_i; d_i, \omega_i), \quad \dot{b}_i^c = f_{sam}(\hat{b}_i^c; d_i, \omega_i), \tag{4}$$

where $f_{sam}$ is the sampling function. In this way, the transformed $\dot{b}_i$ and $\dot{b}_i^c$ are 2D-spatially aligned to $\mathbf{I}_i$. Then, we apply a fusion module to combine $\dot{b}_i$ with $\dot{b}_i^c$ for enhancing multi-image clues by $\tilde{b}_i = f_{conv}([\dot{b}_i, \dot{b}_i^c])$. $f_{conv}$ is a conv-layer, and $\tilde{b}_i$ is the final fused implicit texture.

### 4.3. Constrained Image Rendering

To reconstruct $\mathbf{I}_i$ from the wrapped implicit texture $\tilde{b}_i$, we propose a neural appearance renderer $\Phi^n$ with various regularizations. Compared with the neural rendering methods, our input of the image formation network has been explicitly transformed into 2D space. Hence, $\Phi^n$ only needs to perform spatially-aligned texture recovering without guessing 3D operations. Defining the recovered image as $\hat{\mathbf{I}}_i$, we use $\mathcal{L}_{re} = \mathcal{L}(\hat{\mathbf{I}}_i, \mathbf{I}_i, \sigma_i)$ in Eqn. 2 as the reconstruction loss. To improve the reality, we also leverage an adversarial loss [4] $\mathcal{L}_{adv} = \min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}[\log(\mathcal{D}(\mathbf{I}_i)] + \mathbb{E}[\log(1 - \mathcal{D}(\mathcal{G}(\hat{\mathbf{I}}_i))]$, where $\mathcal{G}$ is $\Phi^n, \Phi^b$ and $\mathcal{D}$ is the discriminator. Further, under different poses, the $\Phi^n$ should robustly recover images with a consistent shape, texture and light. Defining $\hat{\mathbf{I}}_i^{\omega'}$ as the rendered image with randomly sampled pose $\omega'$, we leverage a series of consistency losses to constrain the robustness, illustrated in Fig. 3.

The rotated rendered image $\hat{\mathbf{I}}_i^{\omega'}$ should contain a same facial shape as $\mathbf{I}_i$. To encourage this, we propose a shape-consistency loss using 3D proxy, which is formulated as:

$$\mathcal{L}_{shape} = \frac{1}{\Omega} |(\Phi^d(\hat{\mathbf{I}}_i^{\omega'}) - \Phi^d(\mathbf{I}_i))|. \tag{5}$$

$\Phi^d$ is the 3D proxy network to predict canonical facial depth. In this way, we constrain the renderer to keep the shape consistency. Similarly, we encourage $\hat{\mathbf{I}}_i^{\omega'}$ to contain a same texture and light as $\mathbf{I}_i$, and the loss is formulated as:

$$\mathcal{L}_{tex} = \frac{1}{\Omega} |(\Phi^b(\hat{\mathbf{I}}_i^{\omega'}) - \Phi^b(\mathbf{I}_i))|, \; \mathcal{L}_l = \frac{1}{\Omega} |(\Phi^l(\hat{\mathbf{I}}_i^{\omega'}) - \Phi^l(\mathbf{I}_i))|. \tag{6}$$

$\Phi^b$ and $\Phi^l$ are our texture modeling network and light proxy network, respectively. Finally, the total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{re} + u_1 \mathcal{L}_{adv} + u_2 \mathcal{L}_{shape} + u_3 \mathcal{L}_{tex} + u_4 \mathcal{L}_l, \tag{7}$$

where $u_{1-4}$ are the weighted constants. We optimize $\sum_i^N \mathcal{L}_{total}$ for a collection $\{\mathbf{I}_i\}_{i=1}^N$ in practice. In this way, we constrain the robustness of the renderer under pose variations, and suppress the overfitting on the target image.

### 4.4. Geometry Learning

Once PhyDIR is trained, we can use it as a differentiable renderer for geometry modeling. In contrast to implicit methods [7, 12, 42, 43, 55], PhyDIR disentangles the 3D operations from the neural reasoning procedure. In this way, the lighting, shape and viewpoint clues, which are crucial for geometry learning, can be explicitly back-propagated to $\Phi^l, \Phi^d, \Phi^\omega$. To learn geometry, we use a new $\Phi^d$ with several upsampling-conv layers and a 256 × 256 output size to tack place the proxy. We first freeze the neural reasoning networks $\Phi^b, \Phi^n$, and only optimize the 3D networks $\Phi^l, \Phi^d, \Phi^\omega$. This procedure can be conducted from scratch or started from the 3D proxy. In practice, we find only tiny difference between this two settings. Then, we jointly fine-tune $\Phi^b, \Phi^n$ with the geometry networks using Eqn. 7. Compared with the 3D proxy and other methods, our approach benefits from neural texture modeling and multi-image consistency. These advantages lead to high-fidelity facial shape modeling performance.

## 5. Experiment

### 5.1. Setup

**Dataset:** We train our method mainly on CelebA [35] and CASIA-WebFace [68], then fine-tune it on a high-resolution dataset CelebAMask-HQ [31]. Following [75], 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. For evaluation on facial geometry, following [3, 65, 75], we perform testing on 3DFAW [21, 25, 73, 74], BFM [41] and Photoface [70] dataset. 3DFAW contains 23K images with 66 3D keypoint annotations, and we use the same protocol as [65] to perform testing. For BFM dataset, we use the same generated data released by [65] to evaluate depth maps. Photoface dataset contains 12K images of 453 people with face/normal image pairs, and we follow the protocol of [3, 50] for testing.

**Implementation Details:** We keep the 3D networks $\Phi^d, \Phi^\omega, \Phi^l$ with the same architectures as Unsup3D [65] and LAP [75]. For neural reasoning networks $\Phi^b, \Phi^n$, we

| No. | method | SIDE ($\times 10^{-2}$) ↓ | MAD (deg.) ↓ | SSIM ↑ |
|---|---|---|---|---|
| (1) | Ours-LAP | **0.683**$\pm$0.102 | **15.01**$\pm$1.06 | **87.95** |
| (2) | Ours-Unsup3D | 0.695$\pm$0.110 | 15.12$\pm$1.14 | 86.89 |
| (3) | Implicit texture (c=3) as RGB | 0.724$\pm$0.141 | 15.37$\pm$1.54 | 77.67 |
| (4) | w/o shading | 0.793$\pm$0.202 | 16.03$\pm$1.74 | 78.38 |
| (5) | Target shading only | 0.719$\pm$0.183 | 15.24$\pm$1.72 | 80.56 |
| (6) | Joint shading only | 0.725$\pm$0.118 | 15.40$\pm$1.31 | 79.92 |
| (7) | w/o $\mathcal{L}_{shape}$ | 0.728$\pm$0.115 | 15.81$\pm$1.88 | 83.25 |
| (8) | w/o $\mathcal{L}_{tex}$ | 0.715$\pm$0.109 | 15.46$\pm$1.50 | 80.28 |
| (9) | w/o $\mathcal{L}_{l}$ | 0.701$\pm$0.112 | 15.23$\pm$1.26 | 85.41 |
| (10) | w/o joint learning | 0.708$\pm$0.121 | 15.21$\pm$1.38 | 82.26 |
| (11) | LAP [75] (proxy) | 0.703$\pm$0.137 | 15.30$\pm$1.26 | 62.30 |

Table 2. Comparison with Different Baselines and Settings.

use U-net [49] with a size of 256×256. This leads to 256×256 $b_i, b_i^c, d_i$ and $\hat{\mathbf{I}}_i$. Theoretically, larger modeling sizes are feasible, but we use a similar setting as [7, 51, 75] due to the time and memory cost. We upsample the depth proxy to 256 × 256 to match our prediction for rasterization. A same discriminator as StyleGAN2 [28] is leveraged with the objective of [22]. For implicit texture $b_i, b_i^c$, we set their channel size $c = 32$. We further set $u_1 = 0.5$, $u_{2,3,4} = 0.3$ in Eqn. 7. During training, the size $N$ of photo collection $\{\mathbf{I}\}_{i=1}^{N}$ is randomly selected for the robustness. We train $\Phi^b, \Phi^n$ for 40 epochs on CelebA and CASIA-WebFace, then freeze them to train $\Phi^d, \Phi^\omega, \Phi^l$ for 20 epochs. Finally, we jointly fine-tune all the networks on CelebAMask-HQ for 60 epochs. $\Phi^\sigma$ keeps updating at each stage. We use Adam [30] as the optimizer, and set the learning rate as 0.0001 with a batch size of 8 on a V-100 GPU.

**Evaluation Protocol:** Without special statements, we use single-image results to fairly compare with other methods. Following [3, 65], we use Scale-Invariant Depth Error (SIDE) and Mean Angle Deviation (MAD) to evaluate depth and normal. For evaluating the modeled texture, we calculate Structural Similarity Index (SSIM) [63] and cosine-similarity of encoded representation of Arcface [11] between the original high-quality images and rendered ones, denoted as Cosine-O. Further, we relight/rotate the images with different lights/poses, and compare them with original images using cosine-similarity, denoted as Cosine-L and Cosine-P, respectively. This paradigm can analyse if the image formation method robustly keeps the identity under different light/pose conditions. Please see appendix for more details.

## 5.2. Ablation Study

**Comparison with Baselines:** We first analyse different settings of PhyDIR in Table 2. To analyse the geometry and texture, we fine-tune and test our model on BFM dataset and our CelebAMask-HQ dataset, respectively. Note that, as BFM dataset has no identity labels, we only use single input for fine-tuning. In rows (1) and (2), we observe that PhyDIR has a robust performance between LAP [75] and Unsup3D [65] as proxies. In row (3), we set the channel number of implicit texture $b_i$ and $b_i^c$ as 3, which makes it degrade to the RGB space. This significantly reduces the texture modeling performance, as the representation ability is limited. In rows (4-6), we analyse the effect of shading operations. First, we find that removing shading operation provides an obvious reduction on geometry accuracy. Then, only using target shading or joint shading module cannot obtain satisfactory results. In rows (7-9), we compare the effects of different regularizations. According to the results, we find that $\mathcal{L}_{shape}$ contributes more to geometry reconstruction, while $\mathcal{L}_{tex}$ guarantees the texture modeling performance. $\mathcal{L}_l$ also constrains the rendering stability. In rows (10-11), we find that without joint learning, the geometry modeling performance cannot significantly outperform the proxy. This reveals that joint learning indeed makes the shape modeling benefit from neural rendering.

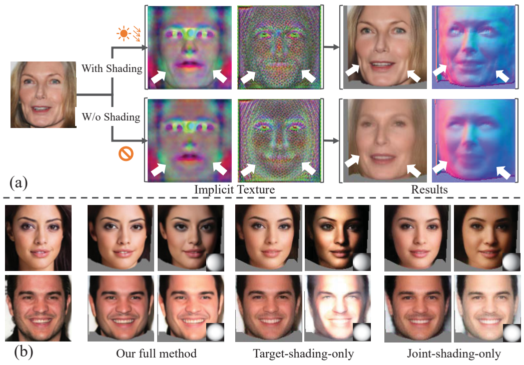

Figure 4. Analysis on the multi-image shading module. (a) How the light modeling improves the details. (b) How the two kinds of shading modules influence the results.

**Analysis on the Shading Module:** We analyse how our shading modules influence the reconstruction performance. In Fig. 4-(a), we compare the model with or without shading modules, and highlight the differences between this two settings. We observe that shading procedure enhances the 'wrinkle' effect on implicit texture maps. In the final reconstruction results, the model with shading modules successfully recovers the wrinkles, while the one without shading fails on predicting such details. Intuitively, as the details are usually produced from the lighting effect of geometry, our shading module is able to mutually improve the joint learning of facial shape and texture. In Fig. 4-(b), we compare the different shading operations. In this comparison, we set the photo collection with 4 images. We observe that only using one of the shading modules produces similar reconstruction results but different relighting effects. The target-shading-only model cannot perceive suitable light intensity, showing heavy overexposure. In contrast, the joint-shading-only model predicts suitable relit effects. This phenomenon indicates that the target-shading-only model tends to overfit on the input image, thus it cannot simulate the effect with unseen lights. On the contrary, as shaded with different lights, the joint-shading-only model adapts to unseen conditions. However, joint shading mixes the implicit texture of photo collection thus losses the specific feature of the target

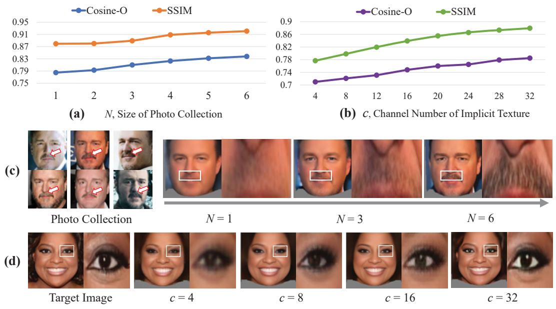

Figure 5. The influence of different sizes $N$ of photo collection and channel numbers $c$ of implicit texture. (a), (b): Quantitative results. (c), (d): Qualitative analyses.

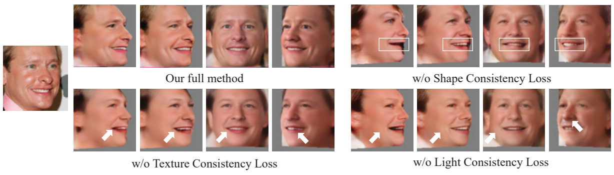
Our full method
w/o Shape Consistency Loss
w/o Texture Consistency Loss
w/o Light Consistency Loss

Figure 6. Analysis on the consistency losses. Without the losses, the model produces artifacts and abrupt changes during rotating.

image. Our full method successfully integrates the advantages of both shading modules, providing consistent results.

**Analysis on Multi-image Consistency:** In Fig. 5, we analyse the influence of photo collection $\{\mathbf{I}\}_{i=1}^{N}$ and implicit texture $b_i, b^c$. The quantitative results are obtained from our CelebAMask-HQ test set to evaluate the quality of modeled texture. Fig. 5-(a) reveals that the accuracy of reconstruction increases with more input images. The accuracy gets an obvious improvement from $N = 1$ to $N = 4$, while performs stable after that. In Fig. 5-(c), the input photo collection contains a common feature of mustache. With the increasing of $N$, the mustache of modeled texture gets clearer and more significant. For the channel number of implicit texture, Fig. 5-(b) indicates that larger $c$ produces superior texture quality, which is also approved with the increasing clarity of the modeled texture in Fig. 5-(d). These analyses well demonstrate that PhyDIR addresses multi-image consistency to improve the performance.

**Analysis on the Losses:** In Fig. 6, we illustrate the results with different consistency losses. In summary, lacking each of the loss leads to artifacts and inconsistent rendering performance. We observe that without the shape consistency loss, the shape of mouth cannot be maintained during rotating. Without texture consistency loss, the results contains obvious texture corruption and artifacts. The model without light consistency loss cannot predict details such as wrinkles or nostrils that are highly related to lighting effect. In contrast, our full method predicts stable result on the consistency of rendering.

| Method | Depth Corr. ↑ | Time (ms) |
|---|---|---|
| Ground Truth | 66 | - |
| AIGN [62] (supervised) | 50.81 | - |
| DepthNetGAN [38] (supervised) | **58.68** | - |
| MOFA [57] (3DMM based) | 15.97 | - |
| DepthNet [38] | 35.77 | - |
| D3DFR [13] | 50.14 | - |
| DECA [17] | 52.23 | - |
| Unsup3D [65] | 54.64 | 0.6 |
| LAP [75] | 57.92 | 2.0 |
| Ours (Unsup3d-proxy) | 58.26 | 1.7 |
| Ours (LAP-proxy) | **59.03** | 2.8 |

Table 3. 3DFAW keypoint depth evaluation of different methods.

| | MAD ↓ | < 20° ↑ | < 25° ↑ | < 30° ↑ |
|---|---|---|---|---|
| Extreme [61] | 27.0$\pm$6.4 | 37.8% | 51.9% | 47.6% |
| SfSNet [50] | 25.5$\pm$9.3 | 43.6% | 57.5% | 68.7% |
| PRN [18] | 24.8$\pm$6.8 | 43.1% | 62.9% | 74.1% |
| DF2Net [71] (GT) | 24.3$\pm$5.7 | 42.2% | 62.7% | 74.5% |
| D3DFR [13] | 23.5$\pm$6.1 | 46.1% | 61.8% | 73.3% |
| Cross-Modal [3] (GT) | 22.8$\pm$6.5 | 49.0% | 62.9% | 74.1% |
| DECA [17] | **22.5**$\pm$5.3 | 48.7% | 62.3% | 73.7% |
| LAP [75] | 23.0$\pm$5.1 | 48.2% | 63.1% | 74.9% |
| Ours | 22.7$\pm$**4.3** | **49.2**% | **63.4**% | **75.3**% |
| SfSNet-ft [50] | 12.8$\pm$5.4 | 83.7% | 90.8% | 94.5% |
| Cross-Modal-ft [3] (GT) | 12.0$\pm$5.3 | 85.2% | 92.0% | 95.6% |
| LAP-ft | 12.3$\pm$**4.5** | 84.9% | 92.4% | 96.3% |
| Ours-ft | **12.0**$\pm$4.9 | **85.3**% | **92.7**% | **96.9**% |

Table 4. Facial normal evaluation on Photoface dataset.

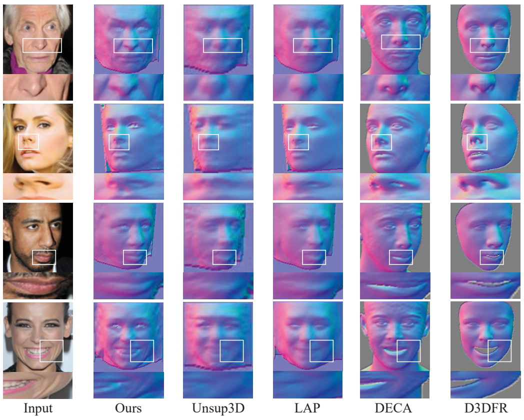

Figure 7. Quantitative comparison on geometry against Unsup3D [65], LAP [75], DECA [17] and D3DFR [13].

## 5.3. Comparison with State-of-the-Art Methods

**Evaluation on Geometry:** We first evaluate the modeled geometry of our method on 3DFAW dataset. Following [65, 75], we use the 2D keypoint locations to sample our predicted depth and calculate the depth correlation score [38] on frontal faces. For a fair comparison, we use our CelebA-pretrained model which is aligned to the setting of Unsup3D, LAP, D3DFR [13] and DECA [17]. We illustrate the results in Table 3, where our method obviously outperforms AIGN, DepthNet, MOFA and 3DMM-based methods. For the proxy methods Unsup3D and LAP, our method successfully outperforms them. Although the inference times are slightly longer, the significant improvement of accuracy brings a satisfactory trade-off.

We then evaluate predicted facial geometry on Photoface

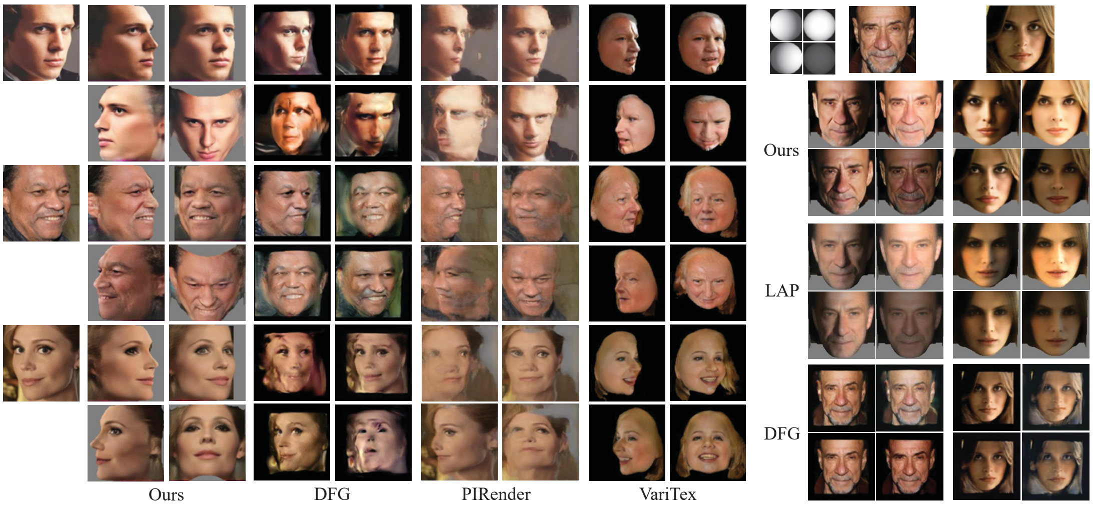

Figure 8. Qualitative comparison with LAP [75], DFG [12], PIRender [43] and VariTex [7] on the robustness of rotation and Relighting.

| Method | Cosine-O ↑ | Cosine-L ↑ | Cosine-P ↑ | SSIM ↑ |
|---|---|---|---|---|
| Unsup3D [65] | 0.622 | 0.593 | 0.568 | 0.514 |
| D3DFR [13] | 0.398 | 0.384 | 0.380 | 0.335 |
| LAP [75] | 0.692 | 0.670 | 0.631 | 0.623 |
| DFG [12] | 0.730 | 0.359 | 0.623 | 0.751 |
| PIRender [43] | 0.702 | 0.417 | - | 0.733 |
| Ours (Unsup3D-proxy) | 0.776 | 0.768 | 0.742 | 0.869 |
| Ours (LAP-proxy) | **0.785** | **0.773** | **0.750** | **0.880** |

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

dataset. Following [3], we transform our predicted facial depth to normal map in order to compute MAD with ground truth. Results are illustrated in Table 4, where '-ft' means fine-tuning on Photoface. We observe that our method obtains competitive results to DECA in the 'no-fine-tuning' condition. Note that, DECA utilizes 3DMM as reliable shape assumption, while our method needs no such prior. For the fine-tuned condition, our method obtains the best performance. Compared with Cross-Modal [3] approach, our method obtains slightly better accuracy but without using ground truth in the training stage. Finally, we perform qualitative evaluation in Fig. 7, where our method produces detailed and realistic facial shapes.

**Evaluation on Texture:** We perform quantitative evaluation in Table 5 on our CelebAMask-HQ test set. As introduced in the evaluation protocol, the Cosine-O is the cosine similarity between the rendered image and the target one on the original pose. The Cosine-L means we add different lights to relight the rendered images, while the Cosine-P means we rotate the rendered images with different yaw and pitch angles. To make DFG [12] address real images, we use a StyleGAN inversion algorithm [1] to optimize the corresponding latent codes. We observe that our method obtains the best performance. While DFG and PIRender [43] produce satisfactory reconstructed results, they suffer from obvious accuracy reduction with rotation and relighting. In contrast, our method is robust to these 3D operations. Then we illustrate qualitative results in Fig. 8. For the neural-rendering-based methods DFG [12] and PIRender [43], we observe that they cannot guarantee precise viewpoint controls or appearance reality during rotating. Although the 3D-aware generative method [7] produces better 3D operations, it cannot well tackle the texture consistency of real images. The graphics-renderer-based method LAP [75] produces unreal relighting performances, while DFG cannot correctly control the lighting effect. Our method shows significantly superior performance and reality on 3D consistency. Besides, we also show **more results** and make discussions on potential **limitation** in the appendix.

## 6. Conclusion

In this paper, we propose a novel Physically-guided Disentangled Implicit Rendering (PhyDIR) framework for 3D face reconstruction. PhyDIR leverages the effectiveness of neural image formation, and disentangles explicit 3D physical operations from this process. To avoid the ill-posed intrinsic factorization, PhyDIR learns implicit texture which helps to integrate photo-collection facial clues. To transform the implicit texture into 2D space, PhyDIR then employs physical graphics pipelines on neural features with explicit controls. A novel multi-image shading module is also proposed to make lighting effect perceivable against single-image limitation. PhyDIR outperforms SOTA rendering methods on texture modeling, and also achieves the best accuracy on 3D facial shape prediction.

**Broader Impact:** The statistics of the training data may bring biases with negative societal impacts. Besides, while the model keeps the input identity, it may generate inexistent contents. These issues warrant further research when building upon this work to model 3D faces.

## References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020. 8

[2] 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. 3, 4

[3] 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. 3, 5, 6, 7, 8

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv: 1701.07875*, 2017. 5

[5] 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, 3

[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. 1, 2

[7] Marcel C Bühler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. *arXiv preprint arXiv:2104.05988*, 2021. 1, 2, 3, 4, 5, 6, 8

[8] 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

[9] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. In *ECCV*, 2020. 1, 3

[10] Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, and Zhoutong Zhang. Differentiable surface rendering via a non-differentiable sampling. In *ICCV*, pages 6088–6097, 2021. 3

[11] 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

[12] 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. 1, 2, 3, 4, 5, 8

[13] 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, 2, 7, 8

[14] 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. 3

[15] 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, 3

[16] 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 modelspast, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 2, 3, 7

[18] 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, 3, 7

[19] 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, 3

[20] 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. 3

[21] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017. 6

[23] Homan Igehy. Tracing ray differentials. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 179–186, 1999. 3

[24] 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. 3

[25] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *IEEE international conference and workshops on automatic face and gesture recognition*, volume 1, pages 1–8, 2015. 5

[26] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. *Advances in neural information processing systems*, 29:4996–5004, 2016. 3

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 3

[28] 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. 1, 3, 6

[29] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 1, 3, 5

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

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

[32] 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. 1, 2, 3

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

[34] 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. 1, 3

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

[36] Matthew M Loper and Michael J Black. Opendr: An approximate differentiable renderer. In *ECCV*, pages 154–169, 2014. 1, 3

[37] 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. 3

[38] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised depth estimation, 3d face rotation and replacement. In *NeurIPS*, pages 9736–9746, 2018. 7

[39] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 2, 3

[40] 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. 3

[41] 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. 5

[42] 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. 1, 2, 3, 5

[43] 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. 1, 2, 3, 4, 5, 8

[44] 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

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

[46] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pages 1259–1268, 2017. 3

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

[48] 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

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 6

[50] 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. 1, 3, 5, 7

[51] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *CVPR*, pages 6258–6266, 2021. 3, 6

[52] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *CVPR*, pages 5011–5020, 2020. 2

[53] Frank Suykens and Yves D Willems. Path differentials and applications. In *Eurographics Workshop on Rendering Techniques*, pages 257–268, 2001. 3

[54] 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, 3

[55] 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, 4, 5

[56] 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, 3

[57] 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, 3, 7

[58] 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. 4

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

[60] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. Face normals in-the-wild using fully convolutional networks. In *CVPR*, pages 38–47, 2017. 3

[61] 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

[62] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *ICCV*, pages 4364–4372, 2017. 7

[63] 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

[64] 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

[65] 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, 3, 4, 5, 6, 7, 8

[66] 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. 1

[67] 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, 3

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

[69] Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *CVPR*, pages 4601–4609, 2019. 1, 3

[70] 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

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

[72] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE TPAMI*, 21(8):690–706, 1999. 3, 4

[73] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013. 5

[74] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 5

[75] 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, 4, 5, 6, 7, 8

[76] 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. 1, 3

[77] 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, 3

[78] 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. 3

[79] 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