High-Quality Facial Geometry and Appearance Capture at Home

Yuxuan Han Junfeng Lyu Feng Xu

School of Software and BNRist, Tsinghua University

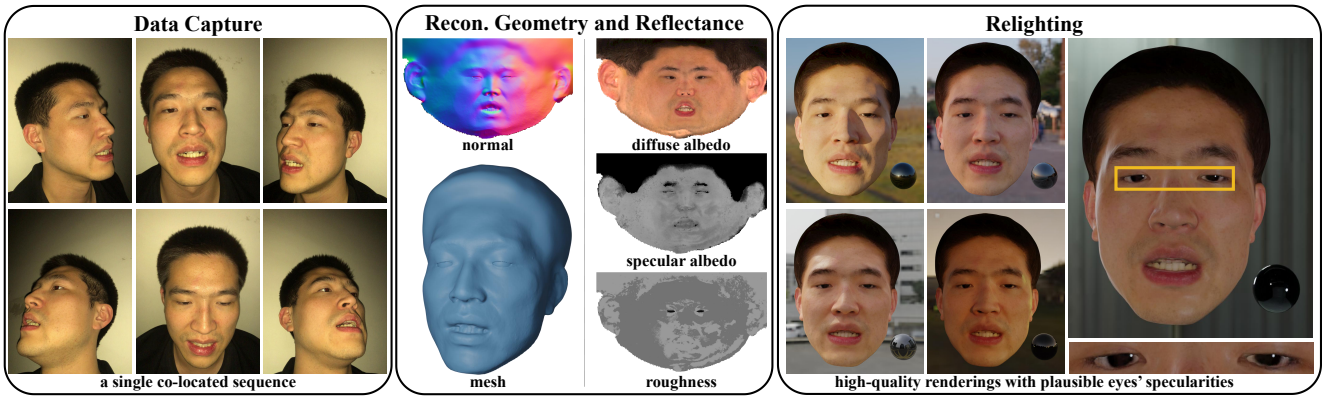

Figure 1. We propose a novel method for high-quality face capture, featuring a low-cost and easy-to-use capture setup and the capability to model the complete face with skin, mouth interior, hair, and eyes. Our method takes a single co-located smartphone flashlight sequence captured in a dim room (*e.g.* rooms with curtains or at night) as input. It reconstructs relightable 3D face assets from the recorded data. These can be used by common graphics software like Blender to create photo-realistic renderings in new environments.

Abstract

*Facial geometry and appearance capture have demonstrated tremendous success in 3D scanning real humans in studios. Recent works propose to democratize this technique while keeping the results high quality. However, they are still inconvenient for daily usage. In addition, they focus on an easier problem of only capturing facial skin. This paper proposes a novel method for high-quality face capture, featuring an easy-to-use system and the capability to model the complete face with skin, mouth interior, hair, and eyes. We reconstruct facial geometry and appearance from a single co-located smartphone flashlight sequence captured in a dim room where the flashlight is the dominant light source (*e.g. *rooms with curtains or at night). To model the complete face, we propose a novel hybrid representation to effectively model both eyes and other facial regions, along with novel techniques to learn it from images. We apply a combined lighting model to compactly represent real illuminations and exploit a morphable face albedo model as a reflectance prior to disentangle diffuse and specular. Experiments show that our method can capture high-quality 3D relightable scans. Our code will be released.*

1. Introduction

High-quality facial geometry and appearance capture are the core steps for cloning our human beings to get into the digital world. To achieve this, existing works [14, 18, 31, 40, 49] develop specialized and expensive apparatus in studios to 3D scan real humans. Although impressive results are demonstrated [1, 2], these techniques are currently only viable for a small number of professional users as on-site data capture is inconvenient and costly. Thus, low-cost but high-quality face capture is strongly in demand to connect broad daily users to the digital world.

A few recent works [3, 46] focus on democratizing the face capture process while keeping the results as close as possible to the studio-based techniques. The rationale of these methods is to exploit the high-frequency light sources in daily life, which is the key to recovering high-quality reflectance [39]. SunStage [46] exploits the sunlight where they reconstruct facial geometry and reflectance from a single selfie video of the subject rotating under the sun. PolFace [3] exploits the smartphone flashlight, *i.e.* they capture two co-located smartphone flashlight sequences with different polarization orientations in a darkroom to estimate facial geometry and reflectance. Although both methods ease the face capture process to a large extent compared to the stu-

dio, the requirements of sunlight [46], polarization filter [3], or darkroom [3] are still inconvenient for daily users to capture faces at home. In addition, as the reflectance property varies significantly across the face (*e.g.* the almost-rough skin *v.s.* the highly-specular eyes), these methods focus on an easier problem to only capture facial skin.

In this paper, we propose a novel method for low-cost high-quality facial geometry and appearance capture, which can model the complete face with skin, mouth interior, hair, and eyes. Firstly, we propose a novel hybrid face representation to adopt different models for different facial regions, *i.e.* eyeballs and other facial regions, due to their reflectance differences. For eyeballs, we adopt two sphere meshes with predefined specular reflectance while leaving the spatially varying diffuse albedo to be solved from the recorded data. The use of eyeball priors improves the reconstruction quality significantly since recovering geometry and reflectance for highly reflective objects (eyeballs in our case) is very challenging [30, 44]. For other facial regions including skin, mouth interior, and hair, we adopt neural field [51] considering its superior representation power and flexibility. Specifically, we adopt a neural SDF field [54] to represent geometry and a neural field to model the parameters of the Disney BRDF [11] as reflectance similar to previous works [12, 56]. To learn our hybrid representation from images, we design a novel mesh-aware volume rendering technique to integrate the eyeball meshes into the volume rendering process of the neural SDF field seamlessly.

To make our method easily used at home, we propose to train our model from a single co-located smartphone flashlight sequence captured in a dim room where the flashlight is the dominant light source (*e.g.* rooms with curtains or at night). Compared to previous works [3, 46], our capture setup neither needs special equipment like the polarization filter and darkroom nor outdoor light sources like sunlight, making it more user-friendly. However, it poses a new challenge to disentangle reflectance from the observed colors. To this end, we involve both lighting and appearance priors to restrict the optimization. Firstly, we apply a combined lighting model to compactly represent both the low-frequency dim ambient light and the high-frequency smartphone flashlight. Then, to constrain the diffuse-specular disentanglement, we resort to the reflectance prior provided by AlbedoMM [42], a 3D morphable face albedo model trained on Light Stage scans [18, 31, 43]. After training, we export our hybrid face representation to 3D assets compatible with common CG software (see Figure 1). By combining our method with Reflectance Transfer [37], we demonstrate application on relightable facical performance capture in a low-cost setup. Our main contributions include:

- We propose a novel method for high-quality facial geometry and appearance capture, featuring a low-cost and easy-to-use capture setup and the capability to model the complete face with skin, mouth interior, hair, and eyes.
- We propose a novel hybrid representation to effectively model eyes and other facial regions and novel techniques to train it from images.
- We apply a combined lighting model to compactly represent the real illuminations and propose a reflectance constraint derived from AlbedoMM [42] to improve diffuse-specular disentanglement in our low-cost capture setup.

2. Related Work

Face Capture. Face capture has attracted much attention in the past two decades. Traditional methods have demonstrated very impressive results [1, 2] under the studio-capture setup. The seminal work of Debevec et al. [14] proposes to reconstruct the face reflectance fields by densely capturing One-Light-At-a-Time (OLAT) images of the human face using the Light Stage [13]. To capture 3D assets compatible with the graphics pipeline, Ma et al. [31] propose to capture the normal and albedo maps of faces leveraging polarized spherical gradient illumination. Subsequently, Ghosh et al. [18] extend this technique to support multi-view capture to obtain ear-to-ear assets. Another class of works captures faces under the single-shot setup. Beeler et al. [7] propose the first single-shot system to capture high-quality facial geometry, and then the follow-up works [19, 40] extend it to support appearance capture. However, all these methods require the users to travel to the studio for on-site capture, which is neither convenient nor low-cost for the broad daily users.

More recently, some works have proposed to democratize the process of face capture. Some methods propose to reconstruct facial geometry and reflectance from a single in-the-wild image [15, 21, 23, 24, 26, 36, 42, 52]. Although these methods are easy to use for daily users, the reconstruction quality is far behind the studio-capture methods. Another class of works attempts to capture faces in the multi-view setup. NeuFace [59] proposes to learn the facial geometry and a novel neural BRDF from the multi-view images captured under unknown low-frequency light. To keep the face capture results as close as possible to the studio, recent methods propose to exploit high-frequency light sources in daily life like sunlight [46] or smartphone flashlight [3]. They solve facial geometry and reflectance from a single selfie video of the subject rotating under the sun [46] or two co-located sequences with different polarization orientations captured in a darkroom [3]. In this paper, we use only a single co-located sequence for face capture, which is more easy to use by daily users. We propose a novel hybrid representation by combining neural SDF field and mesh to reconstruct high-quality complete facial geometry and appearance including skin, mouth interior, hair, and eyes.

Neural Fields for Inverse Rendering. Recent advances represent 3D scene attributes (*e.g.* density and

color) as a continuous function, *a.k.a* neural fields [51], achieving state-of-the-art results on various tasks including view synthesis [4–6, 32, 33] and geometry reconstruction [35, 45, 53, 54]. More recently, some works [9, 10, 12, 22, 29, 34, 55, 56, 58] extend neural fields to inverse rendering, where geometry and reflectance are modeled as neural fields and learned from the captured images. Among these works, the most relevant to us is WildLight [12]. It solves geometry and reflectance from two sequences, one with the flashlight turned on and one turned off. Similar to WildLight, we adopt a neural SDF field [45, 54] to represent geometry and a neural reflectance field to model the parameters of the Disney BRDF [11]. However, we apply a more compact lighting representation so that we require only a single flashlight turned-on sequence for training. In addition, as we focus on the human face rather than common objects, we can exploit face priors. We propose a hybrid representation to exploit eyeball priors to help reconstruction and a reflectance constraint derived from AlbedoMM [42] to regularize the neural reflectance fields.

Hybrid Representation for Digital Avatar. Recent works [16, 25, 60] propose hybrid representation to model digital avatar, considering that we humans are made of components of different properties, *e.g.* skin, hair, eyes, and clothes. Among these works, the most relevant to us is EyeNeRF [25]. In EyeNeRF, the predefined eyeball meshes are used to guide the volume rendering process to better model the ray reflection and refraction on the eyeball surface; the facial geometry is totally represented by the neural density fields [32]. However, in our method, we adopt the eyeball mesh as part of the facial geometry and propose novel strategies to constrain the combination of the eyeball mesh and the neural SDF field. Such a hybrid representation not only helps us to bypass the challenge of reconstructing the highly reflective eyeballs but also makes our method fully compatible with the graphics pipeline.

3. Method

In this Section, we first introduce our capture setup (Section 3.1). We then propose a novel hybrid representation for high-quality and complete face modeling (Section 3.2). To train it from the captured data, we propose a novel mesh-aware volume rendering technique (Section 3.3) and a set of carefully-designed training strategies (Section 3.4).

3.1. Data Capture

As illustrated in Figure 1, we capture a single video sequence around the subject in a dim room using the smartphone camera with its flashlight opened up. The capture takes around 25 seconds for a subject. We use an iPhone X to capture all the sequences in this paper. We resize the images to 960×720 resolution before processing. We calibrate the camera parameters for each frame using an off-the-shelf software[1]. We assume the only high-frequency light source is the smartphone flashlight and further assume it shares the same position as the camera; if not otherwise specified, the data is captured in a room at night[2]. Such a setup has several advantages: *i*) it is easy to fulfill for daily users at home, *ii*) the high-frequency flashlight provides rich cues to recover reflectance, and *iii*) there are no apparent shadows in the captured frames, avoiding extra efforts during optimization.

3.2. Hybrid Representation

Our goal is to capture the complete face with skin, hair, mouth interior, and eyes from the recorded frames. In addition, we require the resulting assets to be compatible with common CG software. In this context, neural fields become the first choice for their superior representation power and flexibility. Similar to existing works [12, 56], we can represent geometry as the neural SDF field and adopt a neural field to model the BRDF parameters as the reflectance; at test time, we can export it to 3D assets including a mesh and a set of UV maps. However, such a holistic representation fails to capture plausible geometry and reflectance for the highly reflective eyeballs, leading to unpleasing results around the eyes (see Figure 3 and 4). To overcome the challenging problem of eyeball reconstruction, we propose a hybrid representation to explicitly exploit eyeball priors. Specifically, we split the whole face F into the eyeballs region E and all the other regions S, *i.e.* $F{=}E \cup S$.

For the S region, including skin, mouth interior, and hair, we adopt a neural SDF field $f_{sdf} : \mathbf{x} \rightarrow sdf_S$ to represent geometry and a neural field $f_{brdf} : \mathbf{x} \rightarrow \{c, s_S, \rho_S\}$ to model the BRDF parameters as the reflectance. Here, $\mathbf{x} \in \mathbb{R}^3$ is the position of the sampled point; $c \in \mathbb{R}^3$, $s_S \in \mathbb{R}$, and $\rho_S \in \mathbb{R}$ are the diffuse albedo, specular albedo, and roughness of the Disney BRDF model [11] respectively.

For eyeballs, *i.e.* the E region, we make an assumption that they are two sphere meshes with the same radius. We further assign a specular lobe with predefined specular albedo s_E and roughness ρ_E to the eyeballs, while leaving the spatially varying diffuse albedo of the eyeballs to be solved from the recorded data. Thus, the only person-specific characteristics of the eyeballs are the two positions $\mathbf{p}_l, \mathbf{p}_r \in \mathbb{R}^3$, a shared scalar radius $r \in \mathbb{R}$, and the diffuse albedo. We reuse f_{brdf} to represent the spatially varying diffuse albedo of the two eyeballs. Although simple, we demonstrate decent rendering results with plausible specularities appearing on the eyes.

3.3. Mesh-Aware Volume Rendering

To train our hybrid representation, we need to render it into an image to compute the photometric loss against the

[1] https://www.agisoft.com/

[2] It is not a darkroom as the white wall and furniture would reflect light.

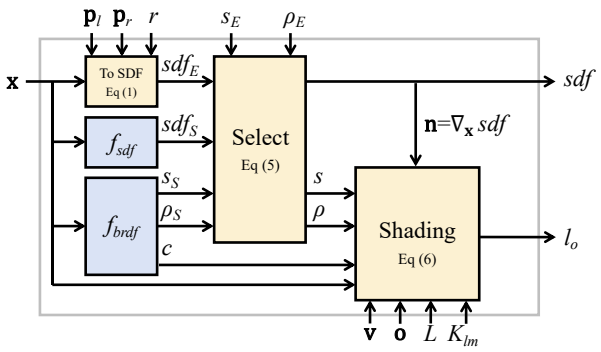

Figure 2. Illustration of the proposed mesh-aware volume rendering technique tailored to our novel hybrid face representation.

recorded frame. To this end, we propose a mesh-aware volume rendering technique tailored to our hybrid representation as illustrated in Figure 2. Specifically, given a 3D position $\mathbf{x} = \mathbf{o} + t \cdot \mathbf{d}$ sampled on the camera ray where $\mathbf{o} \in \mathbb{R}^3$ is the camera position, $\mathbf{d} \in \mathbb{R}^3$ is the opposite view direction, and $t \in \mathbb{R}$ is the viewing distance, we introduce how to compute its SDF value, normal, and observed color.

We first convert the eyeball meshes to the SDF field:

$$sdf_E = \min(||\mathbf{x} - \mathbf{p}_l||_2 - r, ||\mathbf{x} - \mathbf{p}_r||_2 - r) \quad (1)$$

Considering that the complete facial region F is the union of the eyeballs region E and the other region S, we compute its SDF value sdf, specular albedo s, and roughness ρ as:

$$s = \text{select}(s_E, s_S; sdf_E, sdf_S) \quad (2)$$
$$\rho = \text{select}(\rho_E, \rho_S; sdf_E, sdf_S) \quad (3)$$
$$sdf = \text{select}(sdf_E, sdf_S; sdf_E, sdf_S) \quad (4)$$

The intuition is that we select its attributes modeled by either the E region or the S region according to its geometry. We define the differentiable select operator as:

$$\text{select}(*_E, *_S; sdf_E, sdf_S) = \begin{cases} *_E & sdf_E \leq sdf_S \\ *_S & sdf_E > sdf_S \end{cases} \quad (5)$$

Note that Eq (4) is equivalent to setting the SDF value of the union region F as the minimum SDF value of its two components, *i.e.* E and S. The diffuse albedo of region F is directly set to c considering that we use f_{brdf} to represent both regions. We compute the normal $\mathbf{n} \in \mathbb{R}^3$ as the gradient of the SDF value *w.r.t* the position, *i.e.* $\mathbf{n} = \nabla_{\mathbf{x}} sdf$.

Combined Lighting Model. We represent lighting in our capture setup as a combination of the high-frequency smartphone flashlight and the low-frequency dim ambient light. We parametrize the flashlight as a point light with predefined 3-channel intensity L. For the ambient light, we only consider its contribution to the diffuse term; we parameterize the diffuse shading under the ambient light as the 2-order Spherical Harmonics (SH) [38] in the SoftPlus output space to ensure its non-negativity. Then, given the material parameters and the normal, we can compute its shading as:

$$l_o = l_{flash} + l_{amb}, \text{where} \quad (6)$$
$$l_{flash} = \frac{L}{||\mathbf{x} - \mathbf{o}||_2^2} \cdot f_{pbr}(\mathbf{l}, \mathbf{v}; c, s, \rho) \cdot \max(\mathbf{n} \cdot \mathbf{v}, 0) \quad (7)$$
$$l_{amb} = c \cdot \text{SoftPlus}(\sum_{l=0}^{2} \sum_{m=-l}^{l} \cdot K_{lm} \cdot Y_{lm}(\mathbf{n})) \quad (8)$$

Here, f_{pbr} is the Disney BRDF, $\mathbf{v} = -\mathbf{d}$ is the view direction, $\mathbf{l}$ is the light direction and we have $\mathbf{l} = \mathbf{v}$ in the co-located setup, K_{lm} are the SH coefficients for the diffuse shading under the ambient light, and $Y_{lm}(\cdot)$ are the SH bases. Compared to WildLight [12] which uses NeRF [32] to represent the ambient shading, our representation is more compact. Thus, in our scenario, it is feasible to estimate it from a single flashlight video. In addition, it makes our capture process faster, enabling capturing more challenging facial expressions.

Given the SDF value sdf, normal $\mathbf{n}$, and the observed color l_o of a sample point $\mathbf{x}$, we can volume render the camera ray following VolSDF [45].

3.4. Training

We learn the two neural fields f_{sdf} and f_{brdf} and the ambient shading parameters K_{lm} from the captured frames. Similar to EyeNeRF [25], the eyeball position $\mathbf{p}_l, \mathbf{p}_r$, and radius r are set manually, which can be easily done in CG software like Blender. We leave incorporating automatic method [48] into our system as the future work. We emphasize that our goal is to exploit priors to overcome the challenge problem of reconstructing the highly reflective eyeballs, rather than reconstructing high-quality eyeballs with accurate positions and characteristics [8].

To train the neural fields from images, we adopt photometric loss, mask loss, and Eikonal loss [20] similar to previous works [12]. We also adopt a loss to enforce the normal of the nearby sampled points to be the same [41, 58]. See more details in our *supplementary material*. However, all these routine loss functions provide no explicit constraints to regularize the eyeball meshes and the neural SDF field to only represent their own region. Thus, it leads to unnatural results as shown in Figure 3. In addition, our method focuses on capturing the face rather than common objects so that we can exploit face-specific priors for regularization. To this end, we propose two novel and well-designed losses tailored to our task and hybrid representation.

Composition Loss. To constrain the training of our hybrid representation, inspired by ObjectSDF++ [50], we render the occlusion-aware object opacity mask $\hat{O}^E$ and $\hat{O}^S$ for the E and S region and compare them to the corresponding ground truth O^E and O^S obtained from an off-the-shelf face

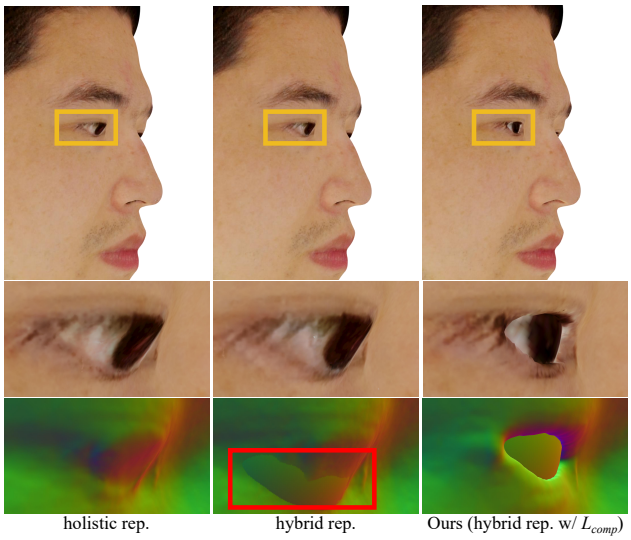

Figure 3. Qualitative evaluation of the hybrid representation and $\mathcal{L}_{comp}$ on geometry reconstruction around eyes. Texture and normal close-ups are shown in the second and third rows respectively.

parsing network [28] over the n sampled rays:

$$\mathcal{L}_{comp} = \sum_{i=1}^{n} ||\hat{O}_i^E - O_i^E||_1 + \sum_{i=1}^{n} ||\hat{O}_i^S - O_i^S||_1 \quad (9)$$

Reflectance Regularization. We exploit the morphable face albedo model – AlbedoMM [42] – as the reflectance prior. Specifically, we devise a multi-view AlbedoMM fitting algorithm to reconstruct the specular albedo for each frame. Then, we enlarge the solved specular albedo to the whole image to obtain I^s as pseudo ground truth to supervise the volume-rendered one $\hat{I}^s$ on the sampled rays:

$$\mathcal{L}_{ref} = \sum_{i=1}^{n} ||k \cdot \hat{I}_i^s - I_i^s||_1 \quad (10)$$

Here, $k \in \mathbb{R}$ is a learnable scalar to compensate for the scale ambiguity stemming from our predefined light intensity L. For pixels from the eyeballs region E, we do not compute $\mathcal{L}_{ref}$ since we already have predefined prior s_{eye}. For pixels from the hair region indicated by the parsing mask [28], we constrain its specular albedo to be 0 to obtain a diffuse appearance as we empirically find fitting a specular lobe produces artifacts when rendered in novel environments.

4. Experiments

We implement our method on top of the multi-resolution hash grid [33] and VolSDF [54] using NerfAcc [27]. Our method can be trained within 70 minutes using a single Nvidia RTX 3090 graphics card. After training, we automatically export our hybrid representation to a triangle mesh with corresponding UV maps for normal, diffuse albedo, specular albedo, and roughness as shown in Figure 1. We adopt Blender to re-render these assets in novel environments. We urge the readers to check our *supplementary video* and *supplementary material* for more implementation details, experimental results, and illustrations.

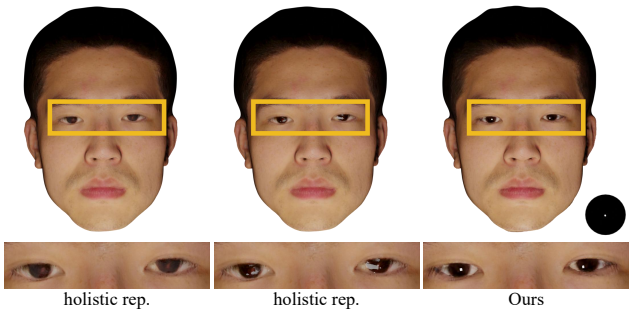

Figure 4. Qualitative evaluation of our hybrid representation and the baseline variants on relighting under a frontal point light.

4.1. Evaluations

Hybrid Face Representation and $\mathcal{L}_{comp}$. Recall that our motivation for proposing the hybrid face representation is to alleviate the challenging problem of reconstructing the highly reflective eyeballs' geometry and appearance from images. To evaluate its effectiveness, we compare a baseline variant *holistic rep.*, where the whole facial geometry and reflectance are represented by f_{sdf} and f_{brdf}. In addition, we compare to a baseline variant *hybrid rep.* where we remove $\mathcal{L}_{comp}$ from our full method to evaluate its efficacy.

We show the geometry reconstruction results in Figure 3. Without eyeballs prior, *holistic rep.* fails to reconstruct reasonable eyeball geometry. Without the composition loss $\mathcal{L}_{comp}$, we cannot ensure the eyeball meshes and the neural SDF represent their own region as we expected. Our method, *i.e. hybrid rep. w/* $\mathcal{L}_{comp}$ obtains the best results. It seamlessly integrates the eyeballs' geometry and reflectance prior to the hybrid face representation and constrains its learning via the composition loss $\mathcal{L}_{comp}$.

We show the relighting results in Figure 4. The baseline variant *holistic rep.* models eyeballs' reflectance the same way as the other facial regions. It fails to reconstruct a plausible specular lobe for eyeballs, leading to unnatural diffuse-looking relighting results around the eyes. We further enhance it by manually setting the specular albedo and roughness of the eyeballs region on the exported UV maps to s_{eye} and ρ_{eye} respectively; we dub this one *holistic rep. w/ naive ref. prior*. Although reflectance prior is utilized, it still cannot generate plausible specularities due to the erroneous geometry reconstruction. Our method produces plausible specularities in eye renderings since we exploit both geometry and reflectance eyeball priors.

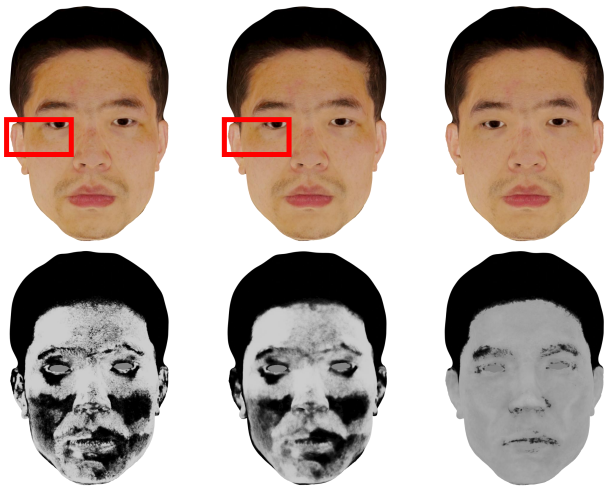

Figure 5. Qualitative evaluation of our reflectance regularization loss $\mathcal{L}_{ref}$ and the baseline variants on diffuse (the first row) and specular (the second row) albedo estimation. See our *supplementary video* for more illustrations.

Combined Light Representation. To evaluate the effectiveness of our combined light representation, we compare a baseline variant that only uses a point light to represent the dominant flashlight while ignoring the ambient. We report the performance gain (in terms of PSNR) on face reconstruction of our method over this baseline on 3 capture environments with increasing ambient intensity: 0.08dB for *night w/ curtain*, 0.05dB for *noon w/ curtain*, and 0.61dB for *noon w/o curtain*. See the photo of these environments and more evaluations in *supplementary material*.

Reflectance Regularization. To regularize the estimated reflectance, we exploit AlbedoMM [42] as priors. We qualitatively evaluate this term as it is a regularizer. As shown in Figure 5, without our reflectance prior loss $\mathcal{L}_{ref}$, the estimated diffuse and specular albedo have a degraded quality compared to the full method. We also compare a widely-used smooth prior that constrains the nearby points' estimated specular albedo to be the same [22, 34, 58] in Figure 5. Again, our method obtains superior quality on diffuse and specular albedo estimation over this naive prior.

4.2. Comparisons

We compare state-of-the-art inverse rendering methods in various problem setups, from low-cost systems to studios. The first class of works takes multi-view images captured in an unconstrained environment as input, which is a bit easier to use than our method; we compare the latest one, *i.e.* NeRO [30], as the representative work. The second class takes the same input as our method; we involve a model-based method NextFace++ for comparison. We also test WildLight [12] under this setup. The last class is

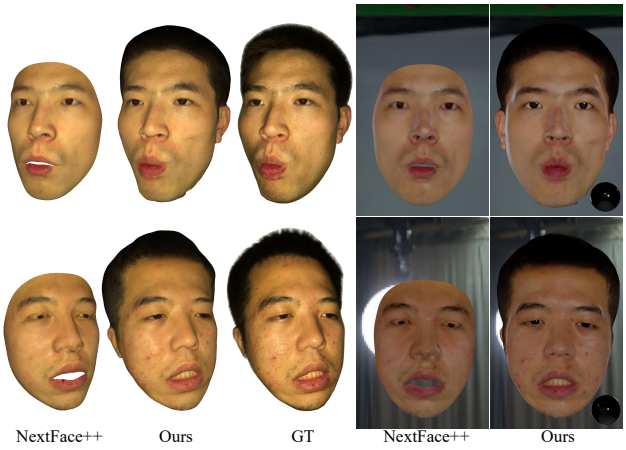

Figure 6. Qualitative comparison of our method and NextFace++ on face reconstruction and relighting.

	PSNR ↑	SSIM [47] ↑	LPIPS [57] ↓
NextFace++	17.62	0.7339	0.2727
Ours	**26.12**	**0.8808**	**0.1642**

Table 1. Quantitative comparison of our method and NextFace++ on face reconstruction. The metric is averaged on 5 subjects.

studio-based methods; we involve a Light Stage-based solution [18]. Due to space limitation, we put the comparison to NeRO and WildLight in our *supplementary material*. We do not compare to PolFace [3] as it is closed-source.

Comparison to NextFace++. NextFace [15] takes single or multiple in-the-wild face images as input. It first fits 3DMM to the images by estimating the lighting, camera parameters, head pose, BFM geometry parameter [17], and AlbedoMM reflectance parameter [42]. Then, they refine the statistical reflectance maps on a per-texel basis. For a fair comparison, we enhance NextFace in the following aspects: (1) we provide our camera parameters to NextFace and introduce a learnable 1D scalar to compensate for the scale difference between the BFM canonical space and our camera frame, and (2) we implement our combined lighting model in NextFace. We dub it NextFace++.

We compare the face reconstruction and relighting results of our method to NextFace++. As shown in Figure 6, NextFace++ can only represent facial skin as it relies on the BFM geometry while our method can represent the complete face thanks to the proposed hybrid representation. On face reconstruction, NextFace++ is confined to the space of the BFM model, which cannot represent person-specific characteristics around the eyes, nose, and mouth, while our method can better fit the captured images as our hybrid geometry representation is more powerful and flexible. On face relighting, our method achieves more realistic results

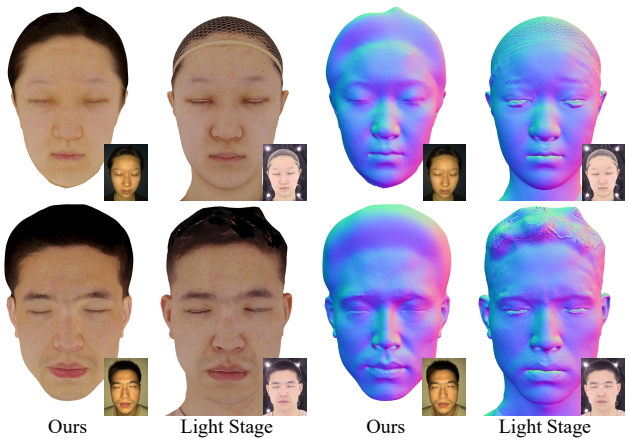

Figure 7. Qualitative comparison of our method and the Light Stage-based solution [18] on diffuse albedo and normal reconstruction. We show one frontal reference frame sampled from the recorded data at the right bottom corner of each image.

around eyes as we exploit eyeballs' reflectance prior while NextFace++ models them the same way as skin. In Table 1, we show quantitative metrics on the hold-out validation images sampled from the captured co-located sequence; we obtain superior results over NextFace++.

Comparison to Light Stage. In high-budget production, Light Stage [13] has demonstrated tremendous success in 3D scanning real humans [1, 2]. To evaluate the performance gap between our low-cost method and the state-of-the-art in the studio, we compare it to the Light Stage-based solution [18] implemented by SoulShell[3]. In their Light Stage, polarization filters are equipped on lights and cameras to capture the diffuse albedo while sphere gradient illuminations are activated to capture the normal; their system cannot capture specular albedo currently. We invite volunteers to their studio for on-site capture.

In Figure 7, we compare the diffuse albedo and normal reconstructed from our method and the Light Stage. Although impressive results are achieved, our method still legs behind the Light Stage results in several aspects: (1) Light Stage can better disentangle the diffuse and specular components due to its usage of the polarization filters, leading to a cleaner diffuse albedo map, and (2) Light Stage can reconstruct higher resolution maps with pore level details since it uses high-definition DSLR camera to capture data, while our method is limited by the quality of the smartphone video camera and the inevitable subtle movement of the subject during our longer capture process (around 25 seconds).

4.3. Results and Application

We present the face capture results of our method on different identities and facial expressions in Figure 10. Our

[3] http://soulshell.cn/

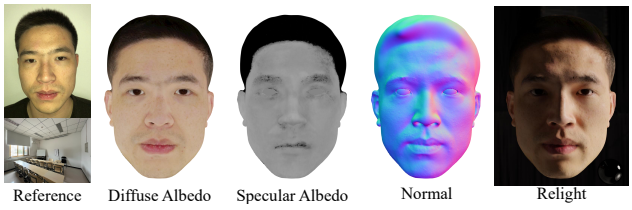

Figure 8. Our method can reconstruct high-quality facial geometry and reflectance even if apparent ambient exists. We show a frontal view sampled from the recorded video and the photo of the scene where we capture data (*noon w/o curtain*) in the leftmost column.

Figure 9. Qualitative face performance relighting results obtained by combining our method with the Reflectance Transfer [37]. We show the origin performance sequence and the relit one in the first and second row respectively.

method can disentangle the diffuse and specular components from the images in a plausible way, leading to an authentic high-quality relightable scan. In addition, our method can capture various challenging facial expressions thanks to the strong representation power and flexibility of our hybrid face representation. In Figure 8, we demonstrate the robustness of our method by training it from the data captured at noon in a room with the curtain opened, *i.e. noon w/o curtain*. In this challenging scenario with apparent ambient, our method still obtains high-quality results.

By replacing the Light Stage scan in the Reflectance Transfer technique [37] as our method's result, we build a simple but strong baseline for the challenging task of relightable face performance capture in the low-cost setup. We record a performance sequence under an unknown but low-frequency lighting and make it relightable as shown in Figure 9. See the *supplementary material* for more details.

4.4. Limitations and Discussions

Although our method demonstrates high-quality results in an easy-to-use manner, it still has several limitations. Similar to EyeNeRF [25], the position and radius of the eyeball meshes are manually set in our method, which incurs

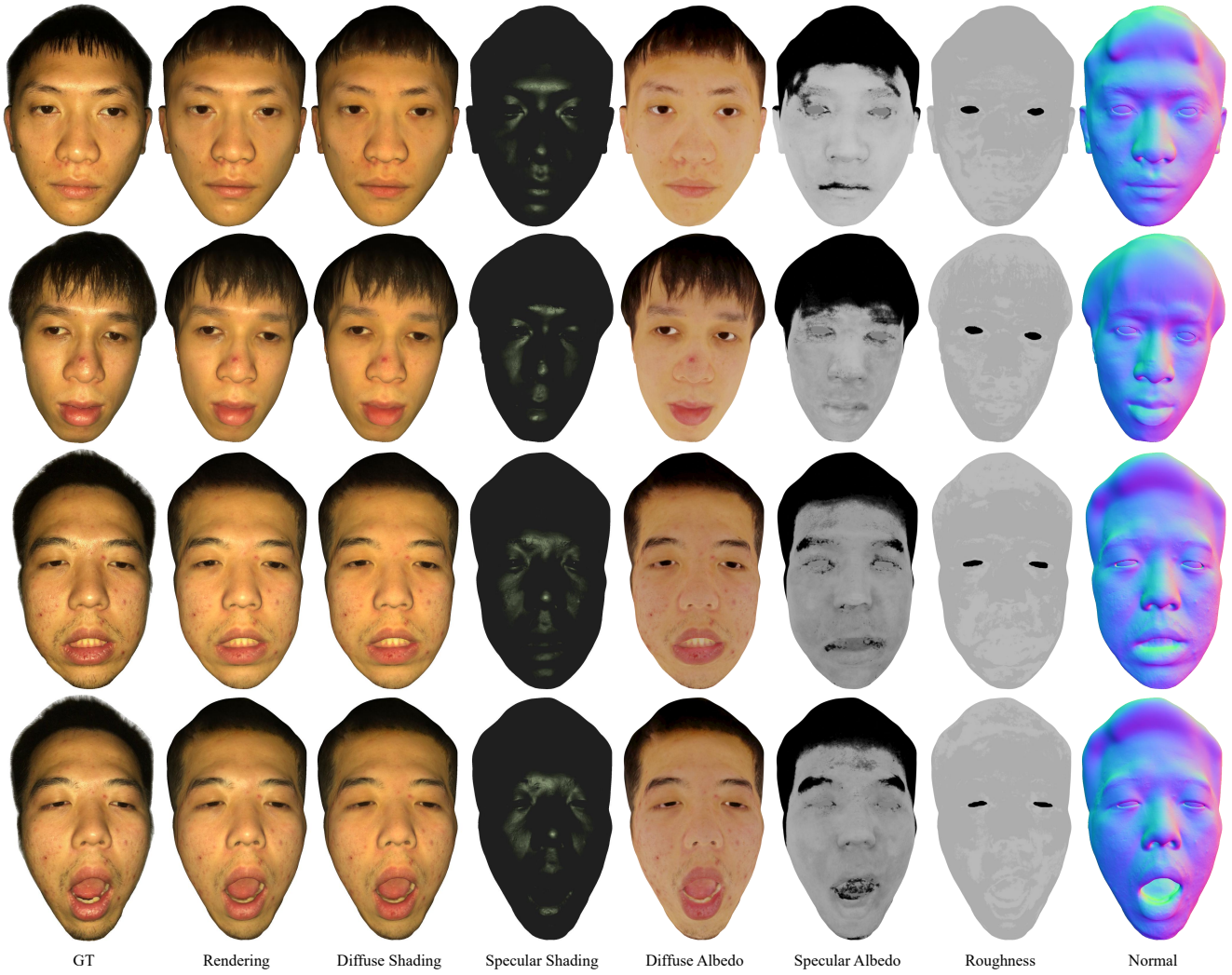

Figure 10. Facial geometry and appearance capture results of our method on different identities and facial expressions.

some manual effort to the whole pipeline. Pre-capturing a multi-gaze video to automatically estimate the eyeball position and size [48] is an interesting direction. Our method takes around 25 seconds to capture a subject with a fixed facial expression. During this period, subtle movement of the subject is inevitable, which would blur the reconstructed texture or bake the eyelids into the eyeball's texture. Speeding up the capture process is an interesting direction. See our *supplementary material* for more detailed discussions.

5. Conclusion

We propose a low-cost and easy-to-use technique for high-quality facial geometry and appearance capture, which takes a single co-located smartphone flashlight sequence captured in a dim room as input. Our method can model the complete face with skin, hair, mouth interior, and eyes. We propose a novel hybrid face representation by combining meshes and neural SDF field and techniques to train it from images. We apply a combined lighting model to compactly model the illumination and propose to exploit AlbedoMM [42] as priors to constrain the estimated reflectance. Our method reconstructs high-quality 3D relightable scans compatible with common CG software.

Acknowledgement

This work was supported by the National Key R&D Program of China (2023YFC3305600, 2018YFA0704000), the NSFC (No.62021002), and the Key Research and Development Project of Tibet Autonomous Region (XZ202101ZY0019G). This work was also supported by THUIBCS, Tsinghua University, and BLBCI, Beijing Municipal Education Commission. Feng Xu is the corresponding author.

References

[1] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *Acm siggraph 2009 courses*, pages 1–15. 2009. 1, 2, 7

[2] Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, et al. Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*, pages 1–1. 2013. 1, 2, 7

[3] Dejan Azinović, Olivier Maury, Christophe Hery, Matthias Nießner, and Justus Thies. High-res facial appearance capture from polarized smartphone images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16836–16846, 2023. 1, 2, 6

[4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[6] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 3

[7] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9. 2010. 2

[8] Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus H Gross. High-quality capture of eyes. *ACM Trans. Graph.*, 33(6):223–1, 2014. 4

[9] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. 3

[10] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 3

[11] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm Siggraph*, pages 1–7. vol. 2012, 2012. 2, 3

[12] Ziang Cheng, Junxuan Li, and Hongdong Li. Wildlight: In-the-wild inverse rendering with a flashlight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023. 2, 3, 4, 6

[13] Paul Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia*, 2(4):1–6, 2012. 2, 7

[14] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 1, 2

[15] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe Gosselin, Marco Romeo, and Louis Chevallier. Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum*, pages 153–164. Wiley Online Library, 2021. 2, 6

[16] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[17] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 6

[18] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)*, 30(6):1–10, 2011. 1, 2, 6, 7

[19] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. 2018. 2

[20] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 4

[21] Yuxuan Han, Zhibo Wang, and Feng Xu. Learning a 3d morphable face reflectance model from low-cost data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8598–8608, 2023. 2

[22] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Advances in Neural Information Processing Systems*, 35:22856–22869, 2022. 3, 6

[23] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020. 2

[24] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8629–8640, 2023. 2

[25] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 3, 4, 7

[26] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3410–3419, 2020. 2

[27] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 5

[28] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, 112:104190, 2021. 5

[29] Jingwang Ling, Zhibo Wang, and Feng Xu. Shadowneus: Neural sdf reconstruction by shadow ray supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2023. 3

[30] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *arXiv preprint arXiv:2305.17398*, 2023. 2, 6

[31] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*, 2007 (9):10, 2007. 1, 2

[32] 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*, 2020. 3, 4

[33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3, 5

[34] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, 2022. 3, 6

[35] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[36] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[37] Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production facial performance relighting using reflectance transfer. *ACM Transactions on Graphics (TOG)*, 26(3):52–es, 2007. 2, 7

[38] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 4

[39] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 1

[40] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. 2020. 1, 2

[41] Radu Alexandru Rosu and Sven Behnke. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2023. 4

[42] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020. 2, 3, 5, 6, 8

[43] Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 611–618. IEEE, 2011. 2

[44] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[45] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 3, 4

[46] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20792–20802, 2023. 1, 2

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

[48] Quan Wen, Derek Bradley, Thabo Beeler, Seonwook Park, Otmar Hilliges, Junhai Yong, and Feng Xu. Accurate real-time 3d gaze tracking using a lightweight eyeball calibration. In *Computer Graphics Forum*, pages 475–485. Wiley Online Library, 2020. 4, 8

[49] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (ToG)*, 25(3):1013–1024, 2006. 1

[50] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21764–21774, 2023. 4

[51] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tomp-

kin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 2, 3

[52] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[53] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 3

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2, 3, 5

[55] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 3

[56] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022. 2, 3

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[58] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 3, 4, 6

[59] Mingwu Zheng, Haiyu Zhang, Hongyu Yang, and Di Huang. Neuface: Realistic 3d neural face rendering from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16868–16877, 2023. 2

[60] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 3