Monocular Facial Appearance Capture in the Wild

Yingyan Xu[1,2] Kate Gadola[1] Prashanth Chandran[2]
Sebastian Weiss[2] Markus Gross[1,2] Gaspard Zoss[2] Derek Bradley[2]
[1]ETH Zürich [2]DisneyResearch|Studios
{yingyan.xu,grossm}@inf.ethz.ch kgadola@student.ethz.ch
{prashanth.chandran,sebastian.weiss,gaspard.zoss,derek.bradley}@disneyresearch.com

Abstract

We present a new method for reconstructing the appearance properties of human faces from a lightweight capture procedure in an unconstrained environment. Our method recovers the surface geometry, diffuse albedo, specular intensity and specular roughness from a monocular video containing a simple head rotation in-the-wild. Notably, we make no simplifying assumptions on the environment lighting, and we explicitly take visibility and occlusions into account. As a result, our method can produce facial appearance maps that approach the fidelity of studio-based multi-view captures, but with a far easier and cheaper procedure.

1. Introduction

3D facial scanning is a fundamental tool for the creation of realistic digital humans in several industries like film and video game entertainment, communication and telepresence, medical applications, and the new trend of AI-driven digital characters. For decades, practitioners have relied on high-quality 3D face scans in order to bring people into virtual worlds. Much of the technology evolution has focused on reconstructing the surface geometry, where initial scanners could create detailed triangle meshes in controlled studio settings with many cameras and lights, and then more recent efforts focused on fast and lightweight face reconstruction from monocular inputs in less-constrained, so-called “in-the-wild” settings. While the facial geometry is extremely important, the shape alone is not enough to re-render the subject in novel environments with photorealistic quality. For this task, we must additionally recover the appearance properties of the face, which dictate how light interacts with the skin surface. As such, in today’s high-end facial scanning pipelines the desired result includes a high-resolution facial surface mesh with corresponding appearance textures for properties like the diffuse albedo, specular intensity and specular roughness, which are compatible with modern skin shaders.

Like facial geometry reconstruction, the field of skin appearance estimation is also well-studied in controlled studio environments, where accurate appearance maps can be reconstructed from large setups that obtain multi-view or multi-shot images under calibrated lighting [55]. Following the geometry trend, current research aims to allow facial appearance capture in less constrained settings, for example outdoors using the sun as a single point light [62]. Unfortunately, these methods often make simplifying assumptions, and thus there still exists a large gap in reconstruction quality between current in-the-wild methods and production-ready studio-based capture.

In this work we present a new method for facial appearance capture in the wild, surpassing the level of fidelity of existing lightweight methods. Our approach requires only a short video sequence of a simple head rotation, captured from a single camera in any environment, including indoors or outdoors, on a sunny day or in shadow. Our approach is built on traditional inverse rendering optimization, where a fast differentiable renderer is used to solve for the geometry and appearance parameters together with the environment lighting simultaneously. Different from previous methods, we do not make any assumptions on the lighting condition (*e.g.* we do not require a sun in the sky), and as our main contribution we explicitly take visibility into account, effectively removing baked-in shading by correctly modeling self-occlusion in our appearance solver. The result is a detailed geometry mesh with textures for diffuse albedo, specular intensity and roughness. As we will show, our approach leads to more faithful recovery of the appearance properties than existing techniques in the wild. As a particular application, our approach allows fast capture of actors on a film set, with resulting assets that can be used directly in traditional VFX pipelines. In summary, we make the following key contributions:

- A new state-of-the-art method for in-the-wild facial appearance capture that makes no assumption on the scene lighting condition.
- A novel shading model that explicitly handles visibility

and self-occlusion for inverse rendering pipelines, achieving high-quality appearance reconstruction from monocular input.

2. Related Work

In the following section, we first outline relevant works around in-the-wild inverse rendering which do not necessarily focus on the human face. Second, we highlight works which specifically tackle facial appearance capture.

Inverse rendering in the wild. Inverse rendering is the process of decomposing the scene into 3D shape, material and illumination by simulating the rendering process and comparing the results against captured images. It has been a popular research topic with the recent advances in novel view synthesis using neural implicit representations [46, 47, 61], 3D Gaussian splatting [36], mesh-based differentiable renderers (rasterizers [37] and path tracers [30, 40]) and using diffusion models as prior [21, 41, 43, 65]. While some existing techniques target a very challenging scenario where the lighting can differ across different images [8, 9, 20], we restrict our discussion here to methods that assume a static unknown environment lighting. PhySG [70] utilizes spherical Gaussians to approximately and efficiently evaluate the rendering equation in closed form. Munkberg *et al.* [48] propose an efficient end-to-end framework for joint learning of topology, triangle meshes and materials, achieving much faster training and inference compared to previous NeRF-based factorization methods [7, 59, 72]. They also introduce a differentiable formulation of the split-sum approximation of environment lighting to efficiently recover all-frequency lighting. Follow-up work [29] shows that material and lighting decomposition can be further improved with a more realistic shading model, incorporating ray tracing and Monte Carlo integration. Recently, 3D Gaussian Splatting techniques were used in conjunction with physical-based rendering to allow for scene relighting [5, 23, 42, 64, 74].

Facial appearance capture. Traditional face capture studios often employ a multi-view setup with controlled and calibrated lighting conditions to reconstruct the skin appearance [14, 25, 26, 55]. Similar setups were used for facial appearance decomposition with neural and gaussian primitives [56, 57, 66, 67, 69]. Lighter alternative setups were also explored by Lattas *et al.* [39] and Choi *et al.* [13] but they still require the subject to be seating in a dedicated space. Recently, the research community has investigated more lightweight setups which are easily accessible to everyone. However, most of these techniques still pose some constraints on the capture environment. CoRA [28] reconstructs relightable 3D face assets from a single co-located smartphone flashlight sequence captured in a dim room. Similarly, Azinović *et al.* [1] additionally attach polarization foils and capture two such sequences with perpendicular polarization orientation to separate skin surface and subsurface reflectance. Using a co-located light and camera setup, these methods assume the position of the dominant light source is known and no shadowing term needs to be modeled. Instead of using a smartphone flashlight, SunStage [62] takes a selfie video rotating under the sun as input and uses the varying angles between the sun and the face as guidance. Cast shadows from the sun are modeled by shadow mapping. It also jointly optimizes the sun's position together with face geometry and appearance. All these methods assume the specular reflection (and the shadowing) of the face comes from a single dominant point or directional light source in the capture environment, and the ambient light contributes only to a low-frequency diffuse term. Another line of work [15–18, 27, 38, 51] aims to reconstruct shape and reflectance properties from a single portrait image, often relying heavily on statistical shape and appearance priors [6, 24, 58] which limits its expressiveness. Our work instead focuses on accurate personalized reconstruction. Rainer *et al.* [54] use tiny shading networks to disentangle shading from explicit reflectance maps but assume a known high-quality geometry and smooth lighting. NeuFace [73] represents the face with a neural SDF and proposes to learn appearance factorization under unknown low frequency light with a novel neural BRDF basis. However, it requires multi-view input similar to those from a light stage setup. Closest to ours, FLARE [4] builds relightable head avatars from monocular videos. It adopts the split-sum approximation [32, 48] for relighting and a neural version of it during training. This rendering model ignores the self-occlusion and bakes part of the shading into the albedos. In contrast, we propose a modified formulation of the split-sum approximation which explicitly handles light visibility and combine it with ray tracing, leading to higher quality shape and appearance reconstruction.

3. Monocular Appearance Capture

We now describe our method for facial appearance capture given a monocular head rotation sequence. We assume the expression of the subject does not change throughout the sequence. As a pre-processing step, we run monocular tracking based on landmarks [11] and a photometric loss [53] to obtain an initial canonical mesh from a 3DMM fit. We also estimate a fixed camera pose, per-frame rigid head poses and neck rotation. For more details about the pre-processing step please see the supplemental PDF. The output of our inverse rendering system will be a 3D asset of the subject containing a high-quality mesh, and diffuse albedo, specular intensity and roughness as 2D texture maps. In the following, we first describe our geometry optimization formulation in Section 3.1, our novel occlusion-aware shading model in Section 3.2, and then optimization details in Section 3.3.

3.1. Geometry Optimization

Balancing the updates to both geometry and textures in inverse rendering can be a challenging task. More specifically, the optimization might overfit too quickly on the textural components before learning the correct geometry. This is often the case when Laplacian shape regularization is applied to enforce geometric smoothness, making the geometry update too slow. Related work such as FLARE [4] addresses this issue by using a two-stage approach. First, a detailed geometry and only blurry textures are learned, and then, the shape is fixed and textures are learned after re-initialization in the second stage. Our method, however, optimizes geometry and textures at the same time. To do so, we adopt a preconditioning framework similar to Nicolet *et al.* [50], which biases gradient steps towards smooth solutions. The vertex positions $\mathbf{v}$ in each iteration are updated by

$$\mathbf{v} \leftarrow \mathbf{v} - \eta(\mathbf{I} + \lambda_{\text{geo}}\mathbf{L})^2 \frac{\partial \mathcal{L}}{\partial \mathbf{v}}, \tag{1}$$

where $\mathbf{v} \in R^{N\times 3}$ collects mesh vertex positions along rows, η is the learning rate, $\mathbf{I}$ is an identity matrix, $\mathbf{L}$ is the uniform Laplacian, and $\mathcal{L}$ is our loss function described in Section 3.3. The hyper-parameter $\lambda_{geo} > 0$ balances between the original objective of matching the input images and a smooth mesh. We set $\lambda_{geo} = 19$ in all our experiments. This way, we can apply a large learning rate for the geometry optimization while keeping the mesh smooth and self-intersection free.

3.2. Occlusion-Aware Shading Model

Following the rendering equation [31], we compute the outgoing radiance $L(\omega_o)$ at location $\mathbf{x}$ from direction ω_o by:

$$L(\omega_o) = \int_\Omega f(\mathbf{x}, \omega_i, \omega_o) L_i(\omega_i)(\omega_i \cdot \mathbf{n}) d\omega_i. \tag{2}$$

We decompose the BRDF $f(\mathbf{x}, \omega_i, \omega_o)$ into the sum of a diffuse term f_d and a specular term f_s. We use the simple Lambertian model for the diffuse term:

$$f_d(\mathbf{x}) = \frac{\rho(\mathbf{x})}{\pi}, \tag{3}$$

where ρ is the diffuse albedo. We use a specular BRDF similar to Kelemen and Szirmay-Kalos [35], which has been shown to be well suited for rendering human skin [49, Chapter 14]:

$$f_s(\mathbf{x}, \omega_i, \omega_o) = \frac{DGF}{4(\omega_i \cdot \mathbf{n})(\omega_o \cdot \mathbf{n})}, \tag{4}$$

where D, G, and F are functions representing the Beckmann normal distribution, geometric attenuation and Fresnel terms, respectively.

Accounting for self-occlusion. We consider direct illumination where $L_i(\omega_i)$ comes only from the light sources. Spherical Harmonics [12, 44] or Spherical Gaussians [68] are popular representations used by prior methods [70, 73] but they can only model low- to medium-frequency lighting. We therefore follow other work, *e.g.* FLARE [4], and use a differentiable split-sum approximation [33, 45, 48], which allows to capture all-frequency lighting. Unfortunately this approach ignores self-shadowing, and so we propose a novel visibility-modulated split-sum approximation to account for self-shadowing, which we describe in the following.

We start by introducing the split-sum formulation from Karis [32] which approximates Eq. 2 as

$$L(\omega_o) \approx \int_\Omega f(\mathbf{x}, \omega_i, \omega_o)(\omega_i \cdot \mathbf{n}) d\omega_i \int_\Omega L_i(\omega_i) D(\mathbf{h})(\omega_i \cdot \mathbf{n}) d\omega_i. \tag{5}$$

The first term is the integral of the BRDF under a solid white environment map, *i.e.* $L_i(\omega_i) = 1, \forall \omega_i$. It only depends on $\cos\theta = \omega_i \cdot \mathbf{n}$ and the roughness r. Therefore, it can be precomputed and stored as a 2D look-up texture. The second integral is a pre-filtered environment map, where D is the normal distribution function of the BRDF and $\mathbf{h} = \frac{\omega_i + \omega_o}{\|\omega_i + \omega_o\|_2}$ is the half vector. This term can also be precomputed as a mipmap by convolving the environment map with D at different roughness values. At rendering time, the outgoing radiance can then be efficient computed using only two texture lookups. Note that when evaluating Eq. 5, the contribution of the prefiltered environment map is only dependent on the reflected view direction ω_r and the roughness r. This is not physically accurate as the contribution should be reduced if the light is (partially) blocked from the upper hemisphere at a specific point. To account for self-occlusion, we must modulate $L_i(\omega_i)$ differently at different locations based on the light visibility $V(\mathbf{x}, \omega_i)$, *i.e.*, making the second integral in Eq. 5 $\int_\Omega L_i(\omega_i) V(\mathbf{x}, \omega_i) D(\mathbf{h})(\omega_i \cdot \mathbf{n}) d\omega_i$. This new integral cannot be precomputed anymore since it's different for different $\mathbf{x}$. However, we notice that when $r \ll 1$, the new integral is 0 unless ω_i is close to ω_r. Therefore, we can approximate it as

$$\begin{aligned} &\int_\Omega L_i(\omega_i) V(\mathbf{x}, \omega_i) D(\mathbf{h})(\omega_i \cdot \mathbf{n}) d\omega_i \\ \approx &\tilde{V}(\mathbf{x}, \omega_r) \int_\Omega L_i(\omega_i) D(\mathbf{h})(\omega_i \cdot \mathbf{n}) d\omega_i. \end{aligned} \tag{6}$$

Here, $\tilde{V}(\mathbf{x}, \omega_r)$ is the view-dependent visibility. The simplest choice of $\tilde{V}(\mathbf{x}, \omega_r)$ is to only evaluate the light visibility at ω_r, *i.e.*, $\tilde{V}(\mathbf{x}, \omega_r) := V(\mathbf{x}, \omega_r)$. Intuitively, this means the outgoing radiance is 0 if the reflected view direction is occluded. Note that this approximation is exact

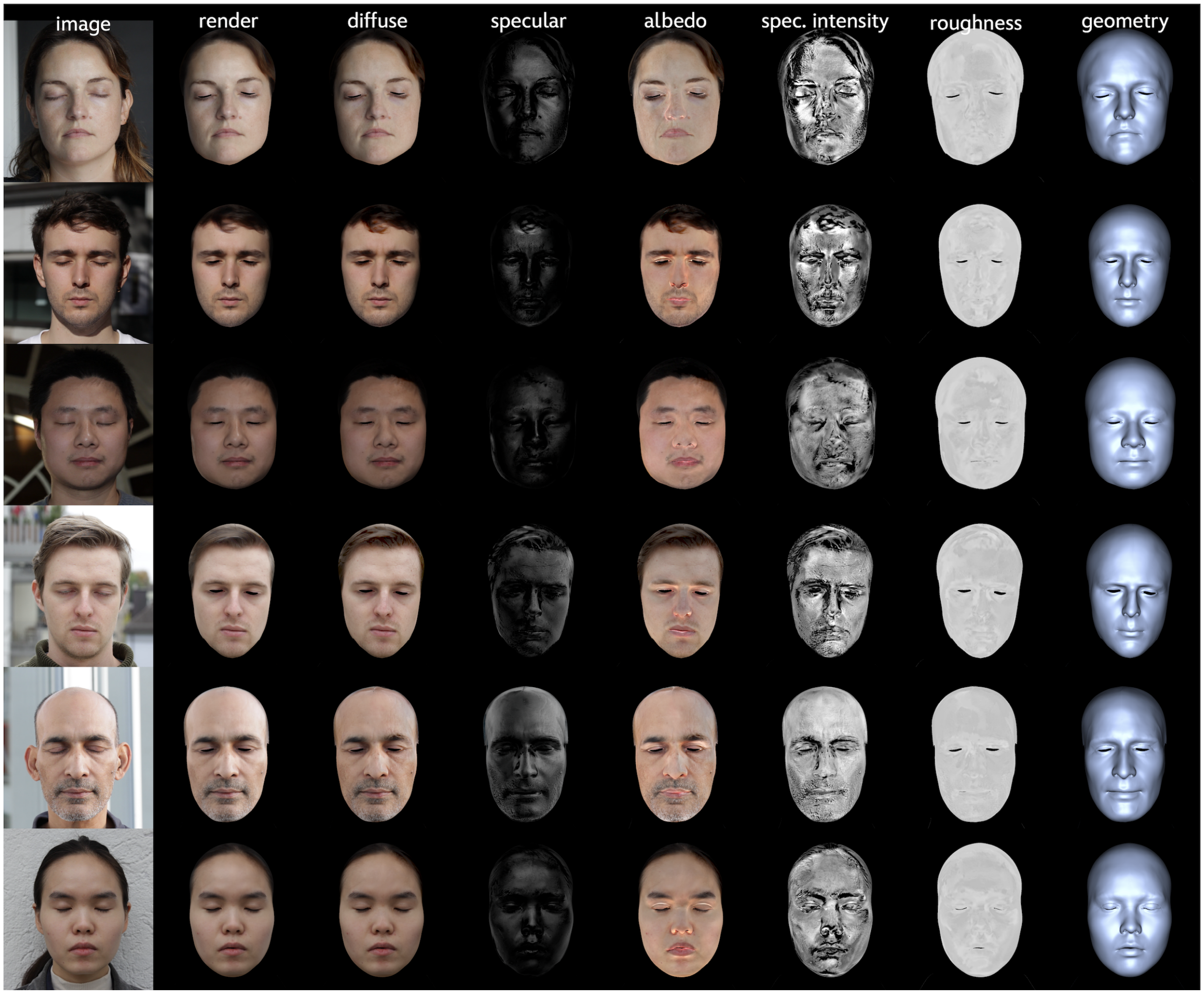

Figure 1. A selection of facial appearance reconstruction and decomposition results for different subjects in different environments, both indoors and outdoors, with sunny and cloudy sky.

for perfect specular reflection (*i.e.* mirrors). To add some softness in the visibility term (instead of considering it as a binary function), we approximate it by Monte Carlo integration,

$$\tilde{V}(\mathbf{x}, \omega_r) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \frac{V(\mathbf{x}, \omega_k)}{D(\mathbf{n}, \omega_k, \omega_r, r)}, \tag{7}$$

where the samples are drawn following the normal distribution of the BRDF. This bears some resemblance to the appearance models of Saito *et al*. [56]. However, they parameterize the view-dependent specular visibility term using a neural network as they work with a large amount of studio data where the lighting is controlled and calibrated. Note however, Eq. 6 introduces large errors for rough surfaces (*i.e.*, when $r \ll 1$ does not hold). We therefore use it only for the specular component, and implemented ray tracing with multiple importance sampling [60] for the diffuse component using the OptiX [52] engine.

3.3. Optimization Details

Our reconstruction loss consists of an L_1 data term $\mathcal{L}_{\text{img}}$ and some regularization terms. We employ an L_1 mask loss $\mathcal{L}_{\text{mask}}$ between the mask obtained from MODNet [34] and the predicted binary mask. Although we use a parameterization similar to Nicolet *et al*. [50], we find it helpful to still employ a Laplacian regularizer to stabilize the geometry optimization such that we do not need to set different λ_{geo} for different datasets. This regularizer encourages the Laplacian of the optimized mesh to stay close to the Laplacian of the initial 3DMM fit

$$\mathcal{L}_{\text{Lap}} = \left\| \mathbf{L}(\mathbf{v} - \mathbf{v}_{\text{init}}) \right\|_2^2. \tag{8}$$

We apply a white light regularization $\mathcal{L}_{\text{light}}$ on the environment map as in [48] and the roughness texture is regularized to be smooth via a total variation loss $\mathcal{L}_{\text{rough}}$. We noticed in our experiments that part of the specular signal tends to be baked into the diffuse albedo. We thus apply a weak regularization to encourage the diffuse render I_{diffuse} to be small if possible, as

$$\mathcal{L}_{\text{diffuse}} = \left\| I_{\text{diffuse}} \right\|_2^2. \tag{9}$$

The final loss is then

$$\begin{aligned}\mathcal{L} :=& \mathcal{L}_{\text{img}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{Lap}}\mathcal{L}_{\text{Lap}} + \\ & \lambda_{\text{light}}\mathcal{L}_{\text{light}} + \lambda_{\text{rough}}\mathcal{L}_{\text{rough}} + \lambda_{\text{diffuse}}\mathcal{L}_{\text{diffuse}}.\end{aligned} \tag{10}$$

4. Experiments

We now show several results of our appearance capture method, evaluate its performance compared to previous work, and offer several ablation studies to validate our design choices. Please refer to our supplemental material for additional results.

4.1. Appearance and Geometry Reconstruction

We begin by highlighting the versatility of our approach by showing several appearance capture results of different subjects in different environments in Fig. 1. Each row of the figure illustrates one of the input images, the corresponding render using our recovered appearance properties, and then a breakdown of the reconstructed appearance maps (diffuse albedo, specular intensity, specular roughness) and geometry. Our method can be applied indoors or outdoors, with sunny or cloudy skies. We show how the recovered geometry and appearance maps can be used to render the subject in a new environment by relighting them. Fig. 2 shows two subjects relit in two different environments (see Fig. 1 for the original environment). Please refer to the supplemental video for more results.

Figure 2. Relighting results of multiple frames on two different subjects in two different environments. The environment is shown top left of each row. See Fig. 1 for the original environment.

4.2. Comparisons

We compare our results to related methods for facial appearance capture in the wild: FLARE [4], NextFace [15], and SunStage [62].

Fig. 3 shows a qualitative comparison of our method to FLARE on two different subjects. While the combined final render is similar between FLARE and our method, it is clear that FLARE fails to separate the diffuse and specular components, baking most of the specular signal in the diffuse map resulting in a nearly zero specular render (Fig. 3, 3rd column). In contrast, our shading model is completely physically-based and correctly separates the diffuse and specular components. The normals reconstructed by FLARE portray a lot of spatial noise and the geometry contains self-intersections, unlike ours (4th column). The diffuse albedo from FLARE contains more baked-in diffuse shading than our result (5th column). As a result, FLARE performs worse when relighting the subject under a novel environment map than our approach (columns 6, 7 and 8).

We also perform qualitative comparisons to NextFace and SunStage in Fig. 4. NextFace relies on a statistical prior, leading to inaccurate shapes and blurry appearance for different subjects (Fig. 4, rows 1 and 4). It produces very similar shapes for the two subjects shown in Fig. 4, while our geometry preserves the identity and likeness of the subjects. SunStage assumes a single point light in the scene, and the shadows and specular components come only from this point light. This leads to poor diffuse and specular separation and incorrect shadows in generic environments like the examples shown here (Fig. 4, rows 2 and 5). In the first subject, we see orange artifacts on the forehead of the diffuse albedo in the SunStage result (row 2). In the second subject, the specular component is completely missing (row 5). The shape and textures from SunStage also have lower resolution than ours. The final row of Fig. 4 illustrates the appearance details for zoomed-in regions shown by the red and blue squares, indicating that our method produces the most accurate details.

As a quantitative comparison, we show reconstruction errors of to NextFace, SunStage and FLARE compared to ours in Table 1. For fairness, we compute errors only on skin regions and average over all subjects in the dataset. Our method prevails in all metrics. The metrics are computed in linear RGB space and averaged over all tested subjects.

4.3. Evaluations

Effects of light visibility. We first evaluate the effects of accounting for self-occlusion by comparing the relit renders of our method with the original split-sum approximation (denoted as *Ours (w/o vis)*) in Fig. 5. We can see that when self-occlusion is not accounted for, shadows under the capture lighting got baked into the albedo (column 1 with zoomed-in patches focusing around the nose), leading to

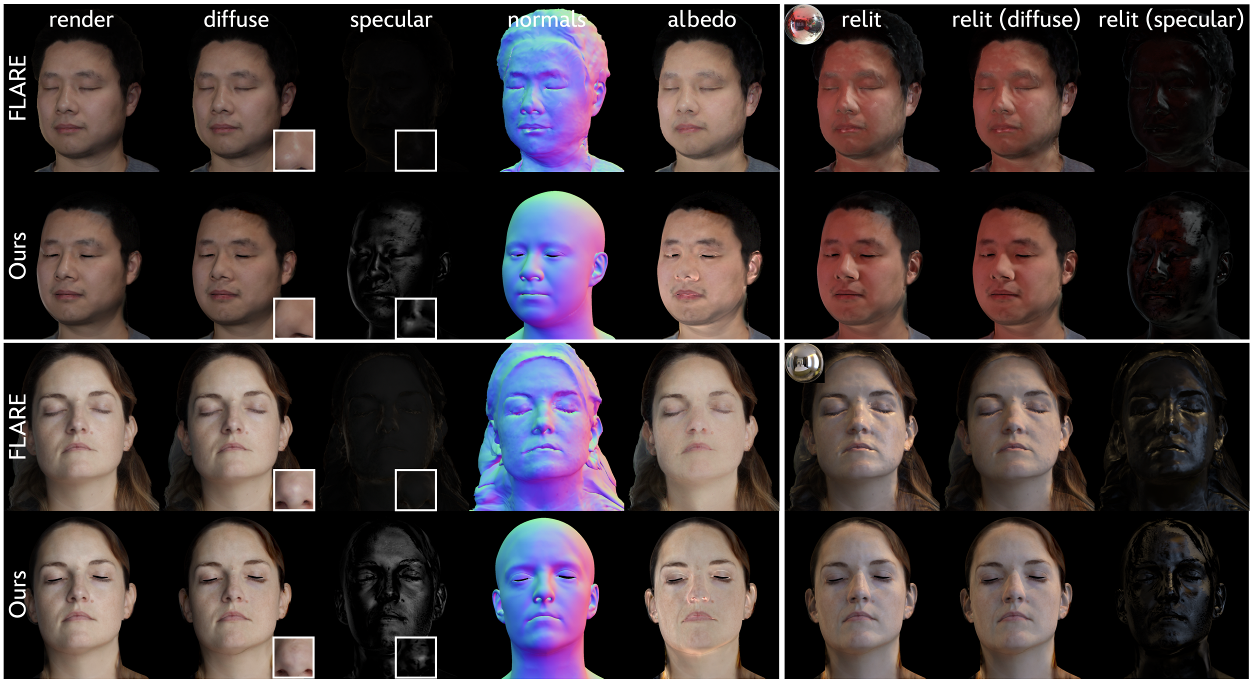

Figure 3. Qualitative comparisons with FLARE [4]. We show one training frame on the left, with separated diffuse, specular, normals and diffuse albedo renders. On the right we show a relit example using the reconstructed appearance. FLARE fails to separate the diffuse and specular signals, produces noisy normals, and bakes shading into the diffuse albedo, all leading to poor results under relighting.

	PSNR ↑	MAE ↓	SSIM [63] ↑	LPIPS [71] ↓
NextFace [15]	25.30	10.63	0.78	0.31
SunStage [62]	29.47	5.28	0.88	0.14
FLARE [4]	30.40	2.01	0.94	0.15
Ours (w/o vis)	34.55	1.79	0.96	**0.10**
Ours	**38.09**	**1.18**	**0.97**	**0.10**

Table 1. Reconstruction errors computed over the skin region averaged for all the subjects.

wrong shadows when relit (row 1, column 2, 3). The specular component also exhibits unrealistic sharp highlight (row 1, column 2, 4). Moreover, when the major light source is behind the subject, the baseline renders show artifacts in the form of strong specular highlights on the side of the nose and under the chin (row 3, column 2, 4) since self-occlusion is not properly handled. In contrast, our proposed model correctly removes baked-in shading and produces more realistic relit renders (row 2, 4). We also quantitatively evaluated the performance gain of accounting for self-occlusion in terms of reconstruction errors in Table 1.

Evaluation on synthetic data. To better evaluate the quality of the reconstructed mesh and textures against ground truth, we created a synthetic dataset with a monocular head rotation sequence similar to the real data. The ground truth mesh and albedo are from a studio appearance capture method [55] (Fig. 6 column 1). We use a natural outdoor environment map on a cloudy day as lighting. Please refer to the supplementary document for more details on the synthetic dataset. We show reconstructed diffuse albedo and geometry of our method with and without light visbility accounted for in columns 2 and 3 of Fig. 6, and the error maps compared with the ground truth are shown in the last two columns. We can see that our method produces a more accurate reconstruction of the diffuse albedo while the baseline result contains a lot of baked-in shading. In terms of shape reconstruction, the two perform similarly. However, the baseline method does not handle shadows well, leading to slightly worse shape recovery in regions where self-occlusion plays an important role, *e.g.*, the lips.

Ours *vs* raytracing for specular. While a ray-tracer handles visibility inherently, we find that our modified split-sum approximation behaves more stably in our inverse rendering setting when we need to solve for shape, textures, lighting at the same time from merely monocular input. Ray tracing produces flickery specular images (Fig. 7 row 2) where part of the skin abruptly changes from very bright to very dark while the head is rotating. We also notice high frequency artifacts, *e.g.* sharp boundary between bright and dark pixels, in the ray traced render, as denoted by the red arrows. In contrast, our model gives visually smoother

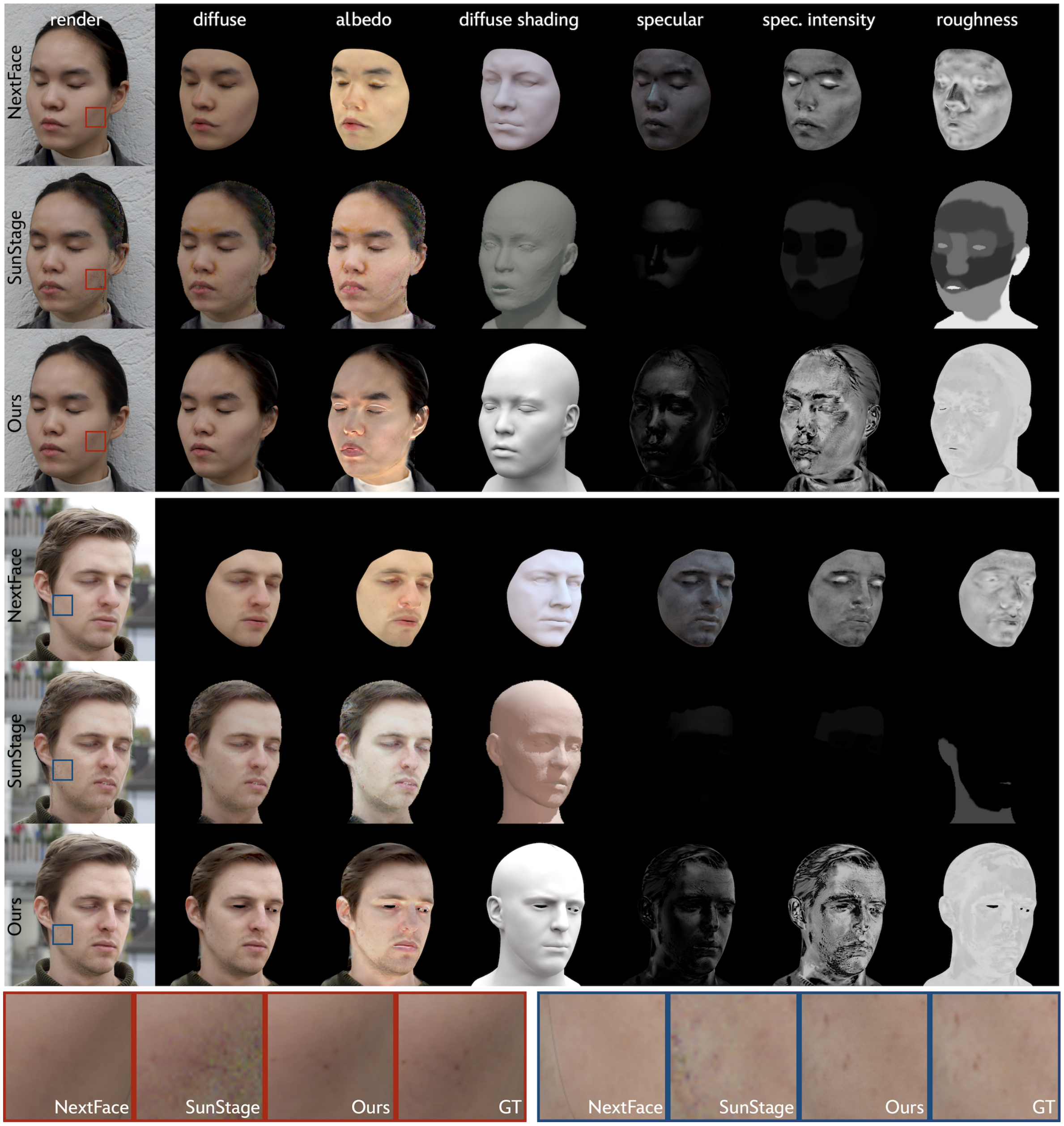

Figure 4. Qualitative comparisons with NextFace [15] and SunStage [62] on two different subjects. The first column is the resulting render overlaid on one input image, and the remaining columns indicate the recovered appearance maps. NextFace produces inaccurate shapes and blurry appearance, where SunStage produces poor shadows and incorrect diffuse/specular separation. Our method produces the most accurate results, also indicated by the zoom region in the final row.

specular renders (Fig. 7, row 3). Please refer to the supplemental videos for an animated visualization.

Optimizing vertex positions *vs* blendweights. We choose to optimize vertex positions directly instead of parameterizing the shape using blendweights of a 3DMM as in NextFace [15]. We show in Fig. 8 that our model achieves the least reconstruction error in the central part of the face compared to the initial mesh or the one optimized through a 3DMM in the same inverse rendering setting. Even though the silhouette of the side face is never shown in the training data, our method recovers the correct shape of the nose from shading. The ground truth reference mesh is a 3D facial scan from a multi-view face scanner [3].

Regularize diffuse component. Last, we apply a weak regularization to encourage the diffuse render to be small which prevents too much specular from being baked into the diffuse component, as we show in Fig. 9. This gives us better separation of the diffuse and specular signals.

Figure 5. Relighting evaluation of our method with and without accounting for self-occlusion. We also visualize the corresponding diffuse albedo, diffuse render and specular render.

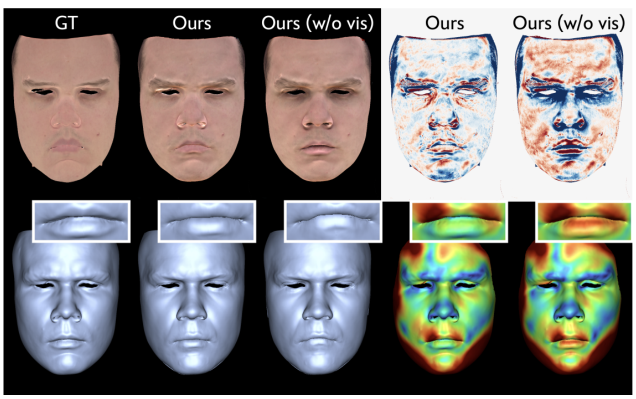

Figure 6. Comparing our method with and without accounting for light visibility on a synthetic dataset. The mesh errors are displayed with a scale of 0mm 5mm and the albedo error with a scale of -0.1 0.1.

5. Limitations

We assume the head poses (and neck rotations) are provided as input to our method. Inaccuracies in the head pose estimation impair the reconstruction quality of our method substantially (see failure cases in the supplemental). Although we make no assumptions on the lighting condition, we cannot recover the appearance if part of the face is in extreme shadows in all frames. Our current face template geometry does not model eyes, however switching to a different template model would allow to reconstruct eyes better. Also,

Figure 7. Comparing the specular renders from our method and a ray-tracer across frames in the same head rotation sequence.

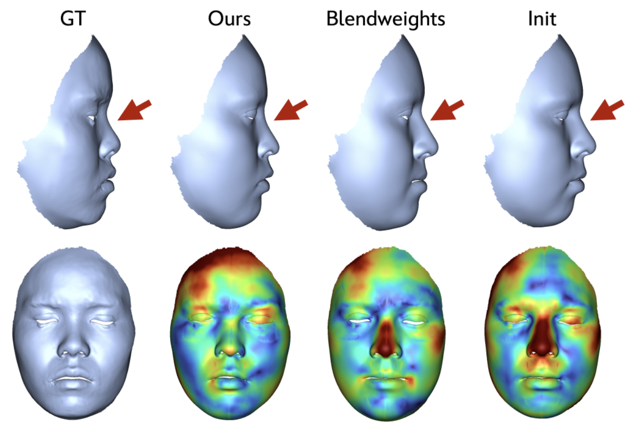

Figure 8. Our geometry optimization pipeline recovers the correct nose shape from shading, achieving the better reconstruction quality compared to 3DMM fit. The mesh errors are displayed with a scale of 0mm 5mm.

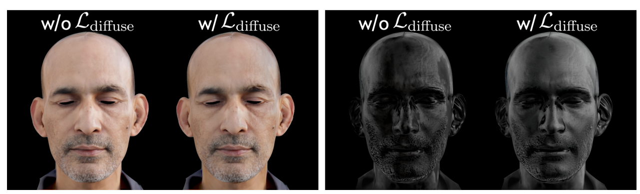

Figure 9. Diffuse and specular component without and with the regularization term $\mathcal{L}_{\text{diffuse}}$.

correct skin tone recovery is not guaranteed due to the ambiguity between the illumination and appearance.

6. Conclusion

We propose a new lightweight facial appearance capture method, surpassing the quality of existing lightweight approaches. It works truly in the wild, making no assumption on the environment illumination. Our novel shading model explicitly accounts for self-occlusion, leading to faithful recovery of the shape and appearance properties from a monocular video containing a simple head rotation.

References

[1] 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. 2, 1

[2] Kelian Baert, Shrisha Bharadwaj, Fabien Castan, Benoit Maujean, Marc Christie, Victoria Abrevaya, and Adnane Boukhayma. Spark: Self-supervised personalized real-time monocular face capture. *arXiv preprint arXiv:2409.07984*, 2024. 3

[3] 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. 7

[4] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J Black, and Victoria Fernandez-Abrevaya. Flare: Fast learning of animatable and relightable mesh avatars. *arXiv preprint arXiv:2310.17519*, 2023. 2, 3, 5, 6, 1

[5] Zoubin Bi, Yixin Zeng, Chong Zeng, Fan Pei, Xiang Feng, Kun Zhou, and Hongzhi Wu. Gs^3: Efficient relighting with triple gaussian splatting. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 2

[7] 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. 2

[8] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34: 10691–10704, 2021. 2

[9] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022. 2

[10] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Semantic deep face models. In *2020 international conference on 3D vision (3DV)*, pages 345–354. IEEE, 2020. 1

[11] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, and Derek Bradley. Continuous landmark detection with 3d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16858–16867, 2023. 2, 1

[12] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Adv. Neural Inf. Process. Syst.*, 32, 2019. 3

[13] Seokjun Choi, Seungwoo Yoon, Giljoo Nam, Seungyong Lee, and Seung-Hwan Baek. Differentiable display photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11831–11840, 2024. 2

[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. 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, 5, 6, 7, 1

[16] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12819–12829, 2021.

[17] Abdallah Dib, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. *arXiv preprint arXiv:2203.07732*, 2022. 1

[18] Abdallah Dib, Luiz Gustavo Hafemann, Emeline Got, Trevor Anderson, Amin Fadaeinejad, Rafael MO Cruz, and Marc-André Carbonneau. Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1770–1780, 2024. 2

[19] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 1

[20] Andreas Engelhardt, Amit Raj, Mark Boss, Yunzhi Zhang, Abhishek Kar, Yuanzhen Li, Deqing Sun, Ricardo Martin Brualla, Jonathan T Barron, Hendrik Lensch, et al. Shinobi: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19636–19646, 2024. 2

[21] Yuto Enyo and Ko Nishino. Diffusion reflectance map: Single-image stochastic inverse rendering of illumination and reflectance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[22] 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. 2

[23] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. *arXiv preprint arXiv:2311.16043*, 2023. 2

[24] 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. 2

[25] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2011. 2

[26] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Transactions on Graphics (ToG)*, 37(6):1–13, 2018. 2

[27] 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

[28] Yuxuan Han, Junfeng Lyu, and Feng Xu. High-quality facial geometry and appearance capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 697–707, 2024. 2, 3

[29] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380*, 2022. 2

[30] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 2

[31] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3

[32] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 2, 3

[33] Jan Kautz, Pere-Pau Vázquez, Wolfgang Heidrich, and Hans-Peter Seidel. A unified approach to prefiltered environment maps. In *Rendering Techniques 2000: Proceedings of the Eurographics Workshop in Brno, Czech Republic, June 26–28, 2000 11*, pages 185–196. Springer, 2000. 3

[34] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022. 4

[35] Csaba Kelemen and Laszlo Szirmay-Kalos. A microfacet based coupled specular-matte brdf model with importance sampling. In *Eurographics (short presentations)*, 2001. 3

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[37] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 2, 1

[38] 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

[39] Alexandros Lattas, Yiming Lin, Jayanth Kannan, Ekin Ozturk, Luca Filipi, Giuseppe Claudio Guarnera, Gaurav Chawla, and Abhijeet Ghosh. Practical and scalable desktop-based high-quality facial capture. In *European Conference on Computer Vision*, pages 522–537. Springer, 2022. 2

[40] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37 (6):222:1–222:11, 2018. 2

[41] Ruofan Liang, Zan Gojcic, Merlin Nimier-David, David Acuna, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Photorealistic object insertion with diffusion-guided inverse rendering. *arXiv preprint*, 2024. 2

[42] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 2

[43] Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimküehler, and Christian Theobalt. Diffusion posterior illumination for ambiguity-aware inverse rendering. *ACM Transactions on Graphics*, 42(6), 2023. 2

[44] T. M. MacRobert. Spherical harmonics: An elementary treatise on harmonic functions and their applications, 1928. 3

[45] David K McAllister, Anselmo Lastra, and Wolfgang Heidrich. Efficient rendering of spatial bi-directional reflectance distribution functions. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*, pages 79–88, 2002. 3

[46] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[47] 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. 2

[48] 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. 2, 3, 5

[49] Hubert Nguyen. *GPU Gems 3*. Addison-Wesley Professional, 2007. 3

[50] Baptiste Nicolet, Alec Jacobson, and Wenzel Jakob. Large steps in inverse rendering of geometry. *ACM Transactions on Graphics (TOG)*, 40(6):1–13, 2021. 3, 4

[51] 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*, pages 8806–8817, 2023. 2

[52] Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. Optix: a general purpose ray tracing engine. *Acm transactions on graphics (tog)*, 29(4):1–13, 2010. 4, 1

[53] Shenhan Qian. Versatile head alignment with adaptive appearance priors. 2024. 2, 1

[54] Gilles Rainer, Lewis Bridgeman, and Abhijeet Ghosh. Neural shading fields for efficient facial inverse rendering. In *Computer Graphics Forum*, page e14943. Wiley Online Library, 2023. 2, 1

[55] Jérémy Riviere, Paulo FU Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph.*, 39(4):81, 2020. 1, 2, 6

[56] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2024. 2, 4

[57] Kripasindhu Sarkar, Marcel C Bühler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo Gotardo, Thabo Beeler, et al. Litnerf: Intrinsic radiance decomposition for high-quality view synthesis and relighting of faces. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2

[58] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5011–5020, 2020. 2

[59] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[60] Eric Veach and Leonidas J Guibas. Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 419–428, 1995. 4

[61] 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. *arXiv preprint arXiv:2106.10689*, 2021. 2

[62] 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, 5, 6, 7

[63] 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

[64] Tong Wu, Jia-Mu Sun, Yu-Kun Lai, Yuewen Ma, Leif Kobbelt, and Lin Gao. Deferredgs: Decoupled and editable gaussian splatting with deferred shading. *arXiv preprint arXiv:2404.09412*, 2024. 2

[65] Chen Xi, Peng Sida, Yang Dongchen, Liu Yuan, Pan Bowen, Lv Chengfei, and Zhou. Xiaowei. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. *arxiv: 2404.11593*, 2024. 2

[66] Yingyan Xu, Gaspard Zoss, Prashanth Chandran, Markus Gross, Derek Bradley, and Paulo Gotardo. Renerf: Relightable neural radiance fields with nearfield lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22581–22591, 2023. 2

[67] Yingyan Xu, Prashanth Chandran, Sebastian Weiss, Markus Gross, Gaspard Zoss, and Derek Bradley. Artist-friendly relightable and animatable neural heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2457–2467, 2024. 2

[68] Ling-Qi Yan, Yahan Zhou, Kun Xu, and Rui Wang. Accurate translucent material rendering under spherical gaussian lights. *Comput. Graph. Forum*, 31(7):2267–2276, 2012. 3

[69] Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. Towards practical capture of high-fidelity relightable avatars. In *SIGGRAPH Asia 2023 Conference Proceedings*, 2023. 2

[70] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[72] 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. 2

[73] 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, 3

[74] Zuo-Liang Zhu, Beibei Wang, and Jian Yang. Gs-ror: 3d gaussian splatting for reflective object relighting via sdf priors. *arXiv e-prints*, pages arXiv–2406, 2024. 2