High-Quality Facial Geometry and Appearance Capture at Home
Supplementary Material

Yuxuan Han Junfeng Lyu Feng Xu

School of Software and BNRist, Tsinghua University

A. More Implementation Details

We provide some important implementation details in the following. The full training code and the data processing code will be released in the future.

A.1. Camera Calibration

We adopt MetaShape[1] to calibrate a shared camera intrinsic matrix for all the recorded frames and a camera extrinsic matrix for each frame. We translate and scale the mesh reconstructed by MetaShape to fit it into the $[-1, 1]^3$ cube; the camera extrinsic matrices are transformed the same way as the mesh. In this way, we ensure that a $[-1, 1]^3$ bounding box is enough for the neural field to represent the whole face. During the data capture process, we set the ISO to 300, the white balance to 4900K, and the FPS to 30.

A.2. Smartphone Flashlight Calibration

Recall that we parameterize the smartphone flashlight as a point light source with 3-channel intensity L. We further represent L as the multiplication of a 1-channel scale $s_L \in \mathbb{R}$ and a 3-channel RGB color $c_L \in \mathbb{R}^3$, *i.e.* $L = s_L \cdot c_L$. We calibrate c_L by capturing a smartphone flashlight image for a pure-white A4 page. We then adopt the mean color of a select patch on this image as c_L. We empirically set $s_L = 8$ and find it works well for all the subjects we captured following the camera calibration procedure in Section A.1.

A.3. Disney BRDF Implementation

We adopt a modified version of the Disney BRDF f_{pbr} [3], containing a diffuse term and a specular term. Both terms are implemented identically to WildLight [4]; see their paper and open-source code for more details.

A.4. Network Architecture

We implement our neural field on top of the multi-resolution hash grid [16]. The neural SDF field and the neural reflectance field are implemented as independent hash grids. The neural SDF field is initialized to a sphere to stabilize training [8]. In addition, we add a small MLP head on the neural SDF field to predict the view-dependent color.

During training, the photometric loss is computed for both the predicted one (*i.e.* the view-dependent color predicted by the small MLP head) and the physics-based one (*i.e.* the shading color computed from the predicted material, normal, the co-located flashlight, and the ambient light). We empirically find this strategy makes the geometry reconstruction more robust.

A.5. Modeling Photographer's Occlusion in Combined Light Representation

When capturing real-world data, we find the photographer would occlude the ambient light when he or she holds the camera moving around the target subject. In an environment with moderate-level ambient illumination (*e.g. noon w/o curtain* and *asym. ambient*), the photographer's occlusion becomes more apparent. In this scenario, using only K_{lm} to represent the ambient shading is inadequate, as the ambient light is changed as the photographer moves. Thus, we propose to explicitly model the photographer.

Inspired by Eclipse [26], we assign each training view a learnable occlusion mask parameterized as 2-order Spherical Harmonics (SH). Thus, the ambient shading for the i-th view becomes:

$$l_{amb} = c \cdot O^i_{amb}(\mathbf{n}) \cdot \text{SoftPlus}(\sum_{l=0}^{2} \sum_{m=-l}^{l} \cdot K_{lm} \cdot Y_{lm}(\mathbf{n})) \tag{1}$$

Here, $O^i_{amb}(\cdot)$ is the visibility mask for the i-th view to compensate for the occlusion caused by the photographer, parameterized as:

$$O^i_{amb}(\mathbf{n}) = \text{Sigmoid}(\sum_{l=0}^{2} \sum_{m=-l}^{l} \cdot O^i_{lm} \cdot Y_{lm}(\mathbf{n})) \tag{2}$$

Here, O^i_{lm} are the SH coefficients for the occlusion mask for the i-th view, which are learned together with the K_{lm}. In this way, we adopt K_{lm} to represent the global ambient shading and a per-view O^i_{lm} to represent the photograph's occlusion.

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

A.6. Losses

We detail all the loss functions we used in the following. At a specific training iteration, we cast n camera rays to the 3D scene. For the i-th ray, we sample k_i points along it according to the empty space skip strategy proposed by the Instant-NGP paper [16].

Photometric Terms. We adopt an L1 photometric loss:

$$\mathcal{L}_{L1} = \sum_{i=1}^{n} ||\hat{I}_i - I_i||_1 \tag{3}$$

Here, $\hat{I}_i$ is the rendered color for the i-th ray while I_i is the corresponding ground truth.

In addition, we adopt an LPIPS loss $\mathcal{L}_{lpips}$ [33] over the full image to reconstruct richer details.

Both photometric terms are computed in the linear space; we apply a gamma function to convert the sRGB space into the linear space and empirically set the gamma value to 2.2. Although the LPIPS network is trained on images in the sRGB space, we empirically find it also works well on images in the linear space.

Mask Loss. We adopt an L1 mask loss:

$$\mathcal{L}_{mask} = \sum_{i=1}^{n} ||\hat{O}_i - O_i||_1 \tag{4}$$

Here, $\hat{O}_i$ is the rendered occupancy for the i-th ray; O_i is the corresponding pseudo ground truth computed from an off-the-shelf face parsing network [11].

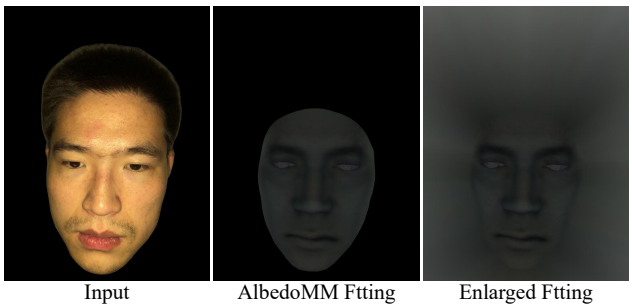
Eikonal Loss. We add an Eikonal term [8] to the sampled points to regularize the SDF value predicted by f_{sdf}:

$$\mathcal{L}_{eikonal} = \sum_{i=1}^{n} \sum_{j=1}^{k_i} (||\nabla f_{sdf}(\mathbf{x}_{ij})||_2 - 1)^2 \tag{5}$$

Normal Smooth Loss. We add a regularization term to encourage smooth normal by constraining the normal of a sampled point $\mathbf{x}$ to be similar to its nearby point $\mathbf{x}^{\epsilon}$ [22, 34]:

$$\mathcal{L}_{eps} = \sum_{i=1}^{n} \sum_{j=1}^{k_i} (1 - \mathbf{n}(\mathbf{x}_{ij}) \cdot \mathbf{n}(\mathbf{x}_{ij}^{\epsilon})) \tag{6}$$

During training, the nearby points are sampled following the strategy of PermutoSDF [22].



Input AlbedoMM Fting Enlarged Fting

Figure 1. We enlarge the AlbedoMM fitted specular albedo to the whole image as the pseudo ground truth.

Composition Loss. To constrain the training of our hybrid representation, inspired by ObjectSDF++ [30], 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 parsing network [11] over the n sampled rays:

$$\mathcal{L}_{comp} = \sum_{i=1}^{n} ||\hat{O}_i^E - O_i^E||_1 + \sum_{i=1}^{n} ||\hat{O}_i^S - O_i^S||_1 \tag{7}$$

Reflectance Regularization. We exploit the morphable face albedo model – AlbedoMM [23] – 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 (see Figure 1) to obtain I^s as pseudo ground truth to supervise the volume-rendered one $\hat{I}^s$ on the sampled rays:

$$\mathcal{L}_{ref} = \sum_{i=1}^{n} ||k \cdot \hat{I}_i^s - I_i^s||_1 \tag{8}$$

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 [11], 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.

A.7. Training Schedule

We adopt a two-stage training strategy. In the first stage, we volume render the neural field to optimize the geometry and reflectance network, *i.e.* f_{sdf} and f_{brdf}, and the shading coefficients K_{lm} (and the optional O_{lm}^i) jointly. We set ω_{L1} to 1, ω_{mask} to 1, $\omega_{eikonal}$ to 1, ω_{eps} to 0.5 in the hair region while 0.02 in other region, ω_{comp} to 1, and ω_{ref} to 0.5.

In the second stage, we extract the mesh from the neural field and perform surface rendering; we fix the geometry

network f_{sdf} while only optimizing the reflectance network f_{brdf} and the shading coefficients K_{lm} (and the optional O^i_{lm}) in this stage. Since geometry is fixed, only photometric loss and reflectance regularization are adopted. We set ω_{L1} to 1, ω_{lpips} to 0.1, and ω_{ref} to 0.01. Note that we turn down the weight of the statistical prior, *i.e.* ω_{ref}, in the second stage to encourage the network to recover more person-specific specular details from the observations.

We train our method by 40000 iterations. The first 30000 iterations are for the first stage and the last 10000 iterations are for the second stage. We adopt the Adam optimizer with an initial learning rate of 0.001. The learning rate is annealed by 0.3 for every 15000 iterations. Our method can be trained within 70 minutes using a single Nvidia RTX 3090 graphics card. In our experiment, the loss weights are shared for all the captured sequences.

A.8. Automatic 3D Assets Extraction

For the S region, we extract the 0.001 iso-surface of the neural SDF field using the Marching Cubes [13] algorithm as we find the geometry is biased in VolSDF; a similar observation can be found in BakedSDF [31]. For the E region, we directly use the sphere meshes as its geometry.

We rasterize the extracted meshes into all the training views and compare the rendered occupancy mask to the face parsing mask. For the triangle faces unseen from every training view, we delete them from the extracted meshes. Then, we find all the connection areas on the mesh using an existing tool [6]. We only keep the largest one while deleting all other connection areas.

We adopt Blender's UV Unwrap function to create the UV mapping function for the meshes. Then, we generate a normal map, diffuse albedo map, specular albedo map, and roughness map accordingly.

A.9. Relightable Performance Capture

In this Section, we show how we combine our method with the Reflectance Transfer technique [20] to construct a simple and powerful baseline for the challenging problem of relightable facial performance capture in a low-cost setup.

Preliminary of Reflectance Transfer. The core idea of Reflectance Transfer is to use optical flow to transfer the lighting effects of a source frame to a target frame. This way, one can capture a relightable scan for only one facial expression as the source frame. Then, a new performance sequence of the same person (or a different person who has a similar appearance to the source person) can be relit.

Specifically, the relightable scan for the source frame is the densely sampled light transport function captured by the Light Stage [7]. The target performance sequence $\{I^i_{src}\}^n_{i=1}$ is captured under a known lighting L_{src}.

To obtain the relit target performance sequence $\{I^i_{tgt}\}^n_{i=1}$ under a new lighting L_{tgt}, they first render the source relightable scan under L_{src} and L_{tgt} to obtain the corresponding renderings I^0_{src} and I^0_{tgt}. The lighting effects R^0 is defined as the ratio image of I^0_{src} and I^0_{tgt}:

$$R^0 = \frac{I^0_{tgt}}{I^0_{src}} \tag{9}$$

Then, they warp the lighting effects R_0 to a target frame I^i_{src} using the warping function computed from I^0_{src} and I^i_{src} to obtain R_i:

$$R^i = \text{warp}(R^0) \tag{10}$$

Here, $\text{warp}(\cdot)$ is the optical flow computed from I^i_{src} to I^0_{src}; note that the lighting condition of I^0_{src} and I^i_{src} are the same, which is the key to compute reliable optical flow.

By multiplying the warped light effects R^i with the target frame I^i_{src}, a relit frame I^i_{tgt} can be obtained:

$$I^i_{tgt} = R^i \odot I^i_{src} \tag{11}$$

See their paper for other details including refining the warping function, filtering the ratio image, aligning the head pose, and the keyframe propagation technique to enhance temporal consistency. Although the Reflectance Transfer method is not physically based, it works well for a large body of low or mid-frequency illuminations as demonstrated by their paper [20].

Combine Our Method to Reflectance Transfer. Recall that the Reflectance Transfer method requires a relightable scan for the source frame, a facial performance sequence captured under known lighting L_{src}, and a target lighting L_{tgt}. Our goal is to construct a low-cost version of the Reflectance Transfer to support relightable facial performance capture in the low-cost setup. To this end, we modify their method in several aspects.

For its first requirement – the relightable scan for the source frame, we can directly replace it with our method's results. Given the target lighting L_{tgt}, we can directly render I^0_{tgt}. However, its second requirement, *i.e.* capturing the facial performance sequence under known lighting, is hard to fulfill in the low-cost setup. Thus, we propose to capture the performance sequence under unknown but low-frequency lighting and solve it using our relightable scan.

Specifically, we render the source frame under the first 2-order Spherical Harmonics (SH) basis lighting to obtain $\{I^0_i\}^8_{i=0}$. We parameterize the lighting as the linear combination weights $\{c_i\}^8_{i=0}$ of these SH bases. Given a target frame I^i_{src} captured under the unknown lighting L_{src}, our goal is to estimate $\{c_i\}^8_{i=0}$ to minimize the following photometric loss:

$$\mathcal{L}_{pho} = ||\text{warp}(\hat{I}^0_{src}) - I^i_{src}||_1 \tag{12}$$

Here, $\text{warp}(\cdot)$ is the optical flow computed from I^i_{src} to $\hat{I}^0_{src}$ using RAFT [25]; $\hat{I}^0_{src}$ is computed as the linear combination of $\{I^0_i\}^8_{i=0}$ weighted by $\{c_i\}^8_{i=0}$:

$$\hat{I}^0_{src} = \sum_{i=0}^{8} c_i \cdot I^0_i \tag{13}$$

In our scenario, RAFT can be seen as a fully differentiable optical flow solver. Thus, the photometric loss function is fully differentiable *w.r.t* the unknowns, *i.e.* $\{c_i\}^8_{i=0}$. We adopt gradient descent to minimize $\mathcal{L}_{pho}$. Note that in each iteration step, $\hat{I}^0_{src}$ are updated since $\{c_i\}^8_{i=0}$ are updated. Thus, the warping function also needs to be recomputed in each iteration.

At the beginning of the optimization, the lighting of $\hat{I}^0_{src}$ may be far from I^i_{src}, which violates the optical flow assumption. However, we empirically find RAFT can also produce reasonable results in this scenario. During the optimization process, we observe that as the lighting of $\hat{I}^0_{src}$ becomes closer to I^i_{src}, the estimated optical flow becomes more accurate.

So far, we have introduced how to compute I^0_{src} and I^0_{tgt}. Thus, we can define the light effect R_0 as Equation (9). In this way, we can directly adopt Reflectance Transfer to relight the whole facial performance sequence.

Results. See our *supplementary video* for the facial performance relighting results.

Limitations and Discussions. Although impressive results are demonstrated, this baseline method for low-cost facial performance capture has several limitations.

Similar to the original Reflectance Transfer method, all the shading effects, *e.g.* specularities, shadows, are transferred in the image space via optical flow. Thus, there is no guarantee that these effects are conform to the geometry. For example, we observe the hard shadow flickered across the frames when relit under a high-frequency target illumination due to the unstable optical flow estimation. We emphasize that facial performance relighting under high-frequency illuminations is very challenging even has access to the Light Stage data [14, 19, 32]. When relit under low-frequency illuminations, we observe it can often produce plausible results.

Another limitation inherited from the original Reflectance Transfer method is that, as there is only one source frame, in a target frame there may be some regions that can not find correspondence in the source frame. For example, if eyes in the source frame are opened up, a closed-eye target frame's relighting effects around eyelids cannot be found in the source frame; in this case, the eyelids' lighting effects are hallucinated via the warping function.

	PSNR ↑	SSIM [28] ↑	LPIPS [33] ↓
NextFace++	17.62	0.7339	0.2727
Wildlight	23.80	0.8205	0.2798
Ours	**26.12**	**0.8808**	**0.1642**

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

Nevertheless, we believe our solution can serve as a strong baseline in the field of low-cost facial performance relighting to motivate future work.

B. More Experiments

B.1. More Evaluations

Hybrid Face Representation and $\mathcal{L}_{comp}$. We show more qualitative evaluations of our hybrid representation and the composition loss $\mathcal{L}_{comp}$ in Figure 2 and Figure 3.

Combined Light Representation. In Figure 4, we show the photo of the three captured environments mentioned in Section 4.1 (Combined Light Representation) of our main paper, *i.e. night w/ curtain*, *noon w/ curtain*, and *noon w/o curtain*. In the following, we present a more thorough evaluation of this design choice.

We conduct experiments on a challenging real scene with asymmetric ambient illumination; we dub it *asym. ambient*. In this real scene, a red lamp is placed on the right side of the face. The photo of this scene and one sampled captured frame can be found in Figure 4. We explicitly model the occlusion caused by the photographer (see Section A.5) when testing on this scene. We show the reconstructed shading component of the ambient and the smartphone flashlight in Figure 5. We observe that our method can still disentangle the ambient and flashlight contributions from the captured images in a plausible way. As shown in Figure 6, we observe that the baseline variant, *i.e. Ours w/o* l_{amb}, bakes the red ambient light into the diffuse albedo while our method can obtain a cleaner one.

Reflectance Regularization. We show more qualitative evaluations of our reflectance regularization $\mathcal{L}_{ref}$ in Figure 7.

B.2. More Comparisons

Comparison to WildLight. We compare our method to WildLight [4], a state-of-the-art inverse rendering method for generic objects that takes two co-located smartphone flashlight sequences as input, one with the flashlight opened and the other closed. In WildLight, they learn a NeRF [15]

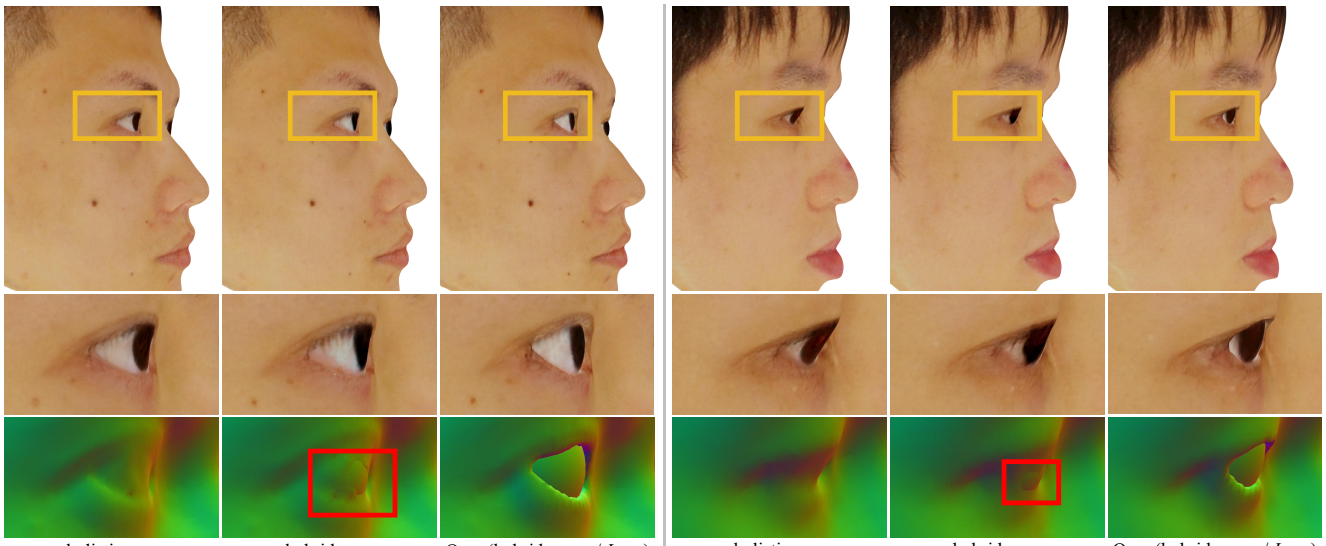


Figure 2. More qualitative evaluation of the hybrid representation and $\mathcal{L}_{comp}$ on geometry reconstruction around eyes. The close-up rendered texture and normal are shown in the second and third rows respectively.

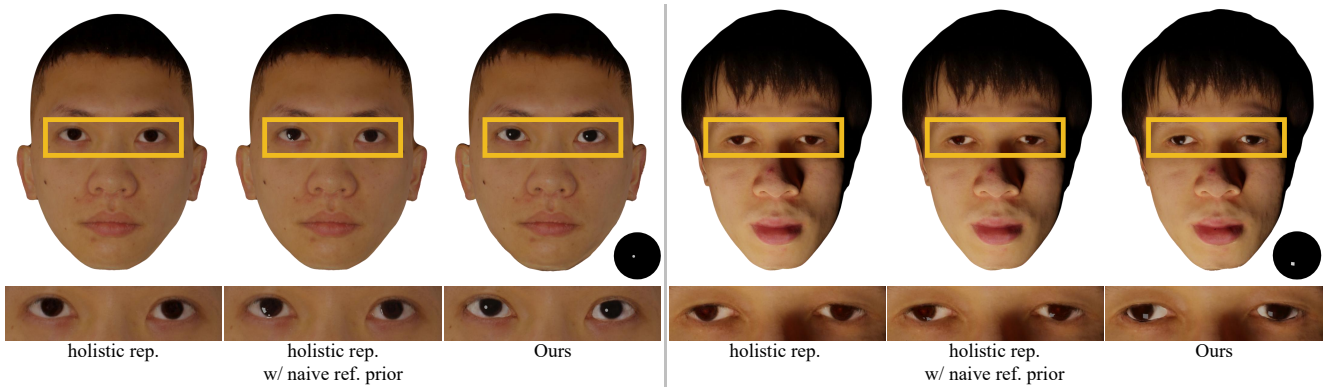


Figure 3. More qualitative evaluation of our hybrid representation and the baseline variants on relighting.

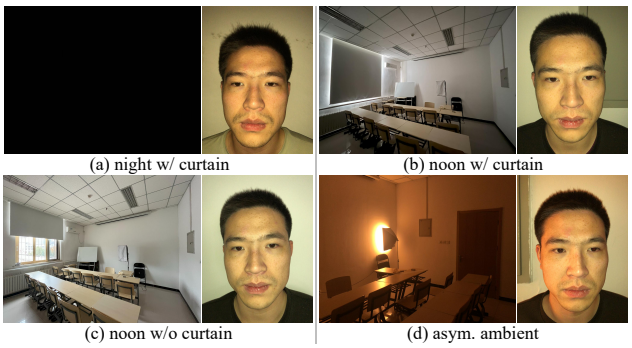


Figure 4. The photo of the scenes we conducted experiments on and the corresponding example images captured in the scene. Note that *asym. ambient* is only referenced in the *supplementary material* while the other three scenes are mentioned in the main paper.

to model the ambient shading and directly use the flashlight-closed sequence to supervise it. Compared to WildLight, our method only needs a single flashlight-opened sequence for training as our combined light representation is more compact. In addition, WildLight cannot model the indirect illumination brought by the smartphone flashlight as their ambient shading NeRF is supervised by the flashlight-closed images.

We first conduct experiments on the data captured in the room at night, *i.e. night w/ curtain*. As shown in Figure 4, in this scenario the images are totally black if the flashlight is closed[2]. Thus, we turn off the ambient shading NeRF in WildLight. We compare the reconstructed geometry and reflectance in Figure 8. Our method obtains superior re-

[2]In this case, the task of the ambient shading term l_{amb} in our combined light representation is to model the indirect illumination caused by the smartphone flashlight.

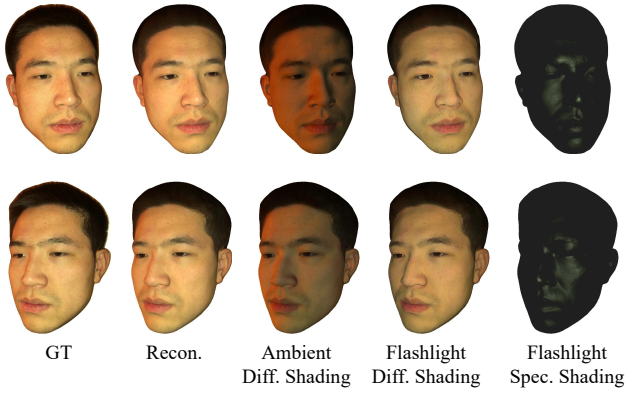


Figure 5. The reconstructed shading contribution of the ambient and the smartphone flashlight on two viewpoints. This experiment is conducted on the data captured in *asym. ambient*.

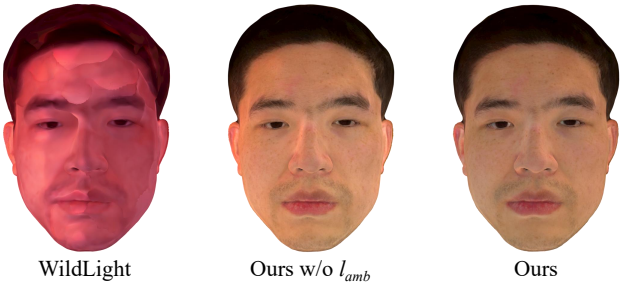


Figure 6. Comparison on diffuse albedo reconstruction of WildLight, our method, and the baseline variant. This experiment is conducted on the data captured in *asym. ambient*.

sults over WildLight since (1) we naturally integrate facial geometry and reflectance priors into our method, leading to better eyeball reconstruction and reflectance estimation, and (2) other design choices of our method, *e.g.* the view-dependent color head and the two-stage training strategy, make it more robust on real-world data and produce more detailed textures. In Table 1, we present quantitative results on face reconstruction; to ensure a fair comparison, all the metrics are computed on the mesh renderings. Not surprisingly, our method obtains better results than WildLight again. We also copy the NextFace++'s results from the main paper for reference.

We then conduct experiments on the data captured in the challenging *asym. ambient* (see the photo of this scene in Figure 4). We turn on the ambient shading NeRF in WildLight. We use the same data as our method to train WildLight, *i.e.* a single co-located sequence with the flashlight opened. As shown in Figure 6, without direct supervision on the ambient shading NeRF, WildLight cannot disentangle the ambient illumination and the smartphone flashlight from the captured images in a plausible way, leading to an unreasonable diffuse albedo. Our method obtains better results as our combined light representation is compact enough to disentangle the ambient and flashlight contributions solely from the captured data.

Comparison to NeRO Some methods [9, 12, 17, 34] in inverse rendering take multi-view images of an object as input; from these images, they estimate the environment lighting and the object's geometry and reflectance. Compared to our method, these works have an even more easy-to-use capture setup for daily users; it neither requires the ambient illumination to be low frequency nor needs the flashlight to be opened up during capture. Among these works, NeRO [12] is the state-of-the-art. In this part, we compare our method to NeRO to see whether our setup is necessary to reconstruct high-quality facial geometry and appearance.

We capture two videos for the same identity, one in our capture setup (*i.e.* co-located smartphone flashlight video captured around the subject in a dim room) and the other following NeRO's capture setup (*i.e.* smartphone video captured around the subject in an unconstrained environment). Some recorded data for training NeRO is shown in Figure 9. The comparison results are shown in Figure 10.

On diffuse albedo reconstruction, NeRO bakes shadow in the diffuse albedo as it is very challenging to simulate the global light transport to model shadow in the inverse rendering process. Our method obtains a cleaner one as in our capture setup the recorded images are almost shadow-free. On geometry reconstruction, our method obtains better results again since our setup makes the inverse rendering problem easier compared to NeRO as we have prior knowledge on lighting, *i.e.* the combination of a low-frequency ambient and a high-frequency flashlight. On face relighting, our method demonstrates better results for two reasons: (1) our method can estimate more accurate geometry and reflectance, and (2) our hybrid face representation can better model the eyes than using a single neural SDF to represent the whole face.

Discussion of Other Related Works NeuFace [35] is a recent method proposed to reconstruct geometry and neural BRDF from multi-view images of a subject captured in a studio with the synchronized multi-camera system. We do not compare it as our goal is to reconstruct high-quality relightable 3D face assets compatible with common graphics software under a low-cost and easy-to-use capture setup. However, NeuFace's neural BRDF representation and the corresponding customized shader cannot achieve our goal.

A concurrent work [21] proposes a method for facial inverse rendering from smartphone-captured multi-view images captured in arbitrary unknown lighting. However, similar to PolFace [1] and SunStage [27], they only focus on facial skin capture, while our method proposes a hybrid face representation that can efficiently represent the complete face with skin, mouth interior, hair, and eyes.

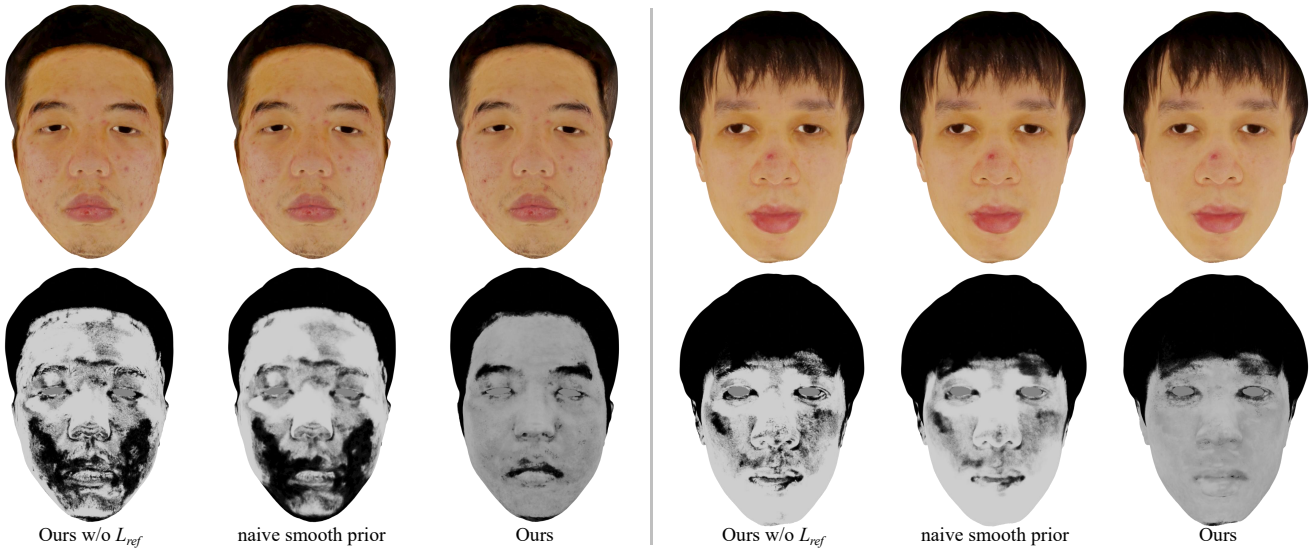


Figure 7. More qualitative evaluation of our reflectance regularization loss $\mathcal{L}_{ref}$ and the baseline variants on diffuse (the first row) and specular (the second row) albedo estimation.

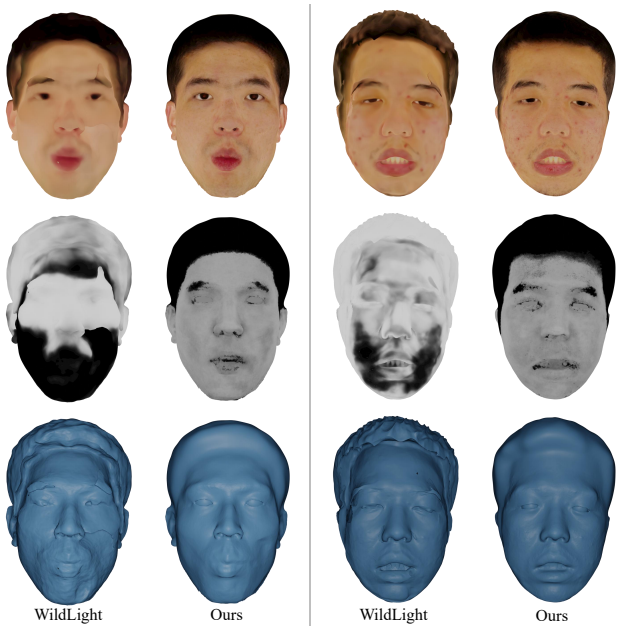


Figure 8. Qualitative comparison of WildLight [4] and our method on diffuse albedo reconstruction (the first row), specular albedo reconstruction (the second row), and geometry reconstruction (the last row).

Figure 9. Example multi-view images to train NeRO. These images are captured in an uncontrolled indoor environment.

C. Limitations, Discussions, and Future Works

Although our method demonstrates high-quality facial geometry and appearance capture results under a low-cost and easy-to-use setup, it still has some limitations.

Similar to EyeNeRF [10], the position and radius of the eyeball meshes are manually set in our method, which incurs some manual effort to the whole pipeline. We have tried to optimize the eyeballs' position and radius in the training process in our preliminary experiment. However, we find the results are not always plausible since the gaze direction is the same across the captured frames, which cannot provide enough cues to solve accurate eyeballs. In future work, we plan to capture a sequence of multi-gaze data before the face capture process to solve the position and radius of the eyeball automatically [29].

Although plausible results are obtained, the geometry and reflectance model of the eyeball are still not very ac-

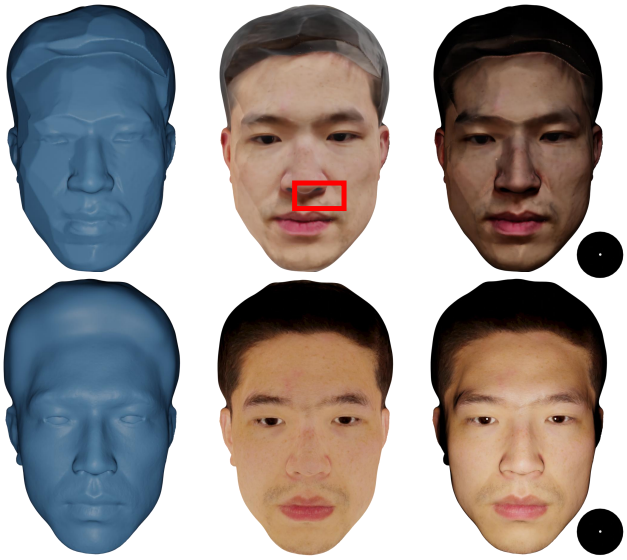


Figure 10. Qualitative comparison of NeRO (the first row) and our method (the second row) on geometry reconstruction, diffuse albedo reconstruction, and relighting results.

curate. In the previous work on high-quality eyeball capture [2], a more complex model was designed for eyeballs. In fact, our hybrid representation makes no assumption on the eyeballs' mesh; to support other eyeball geometry we only need to modify the way to convert the meshes to the SDF field. We chose the sphere meshes because it is easy to use by daily users while the more complex and accurate one [2] is not publicly available. Replacing the sphere eyeball model with this more complex and accurate one in our method to enhance the realism of eye rendering is an interesting direction.

We observe artifacts in the close-up view of the reconstructed eyeballs and/or other face regions' diffuse albedo. We attribute this to the following reasons: (1) our data capture setup requires around 25 seconds to record the sequence for a subject, some face parts (*e.g.* the eyelids, the lips, and the tongue) would move inevitably during the capture although we manually remove some frames with apparent movement like eyes blinking, (2) our method relies on off-the-shelf methods to provide segmentation mask and camera calibrations, which are not perfect, and (3) our method does not model eyelashes so they are often baked in the eyeballs' diffuse albedo. However, we find such artifacts are often imperceptible in the whole face's scale.

We assume the hair is a diffuse surface currently. We believe using a more accurate reflectance model [5] to represent hair is a very interesting future work. However, it is quite challenging in the low-cost setup since we do not have enough cues to reconstruct accurate hair geometry and reflectance [24]. Our method produces artifacts when representing flying hairs as shown in Figure 11. We emphasize that it is a very challenging problem even in studio-based hair reconstruction methods [18, 24].

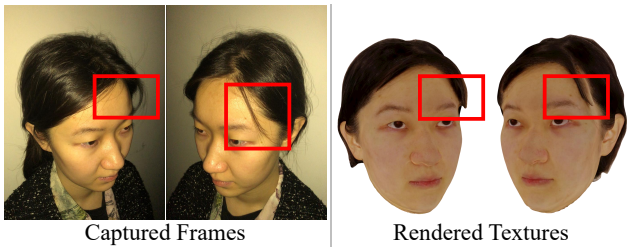


Figure 11. Our method cannot reconstruct plausible flying hairs.

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. 6

[2] 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. 8

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

[4] 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. 1, 4, 7

[5] Matt Jen-Yuan Chiang, Benedikt Bitterli, Chuck Tappan, and Brent Burley. A practical and controllable hair and fur model for production path tracing. In *ACM SIGGRAPH 2015 Talks*, pages 1–1. 2015. 8

[6] Dawson-Haggerty et al. trimesh. 3

[7] 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. 3

[8] 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. 1, 2

[9] 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. 6

[10] 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. 7

[11] 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. 2

[12] 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. 6

[13] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 3

[14] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 4

[15] 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. 4

[16] 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. 1, 2

[17] 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. 6

[18] Giljoo Nam, Chenglei Wu, Min H Kim, and Yaser Sheikh. Strand-accurate multi-view hair capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 155–164, 2019. 8

[19] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 4

[20] 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. 3

[21] 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. 6

[22] 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. 2

[23] 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

[24] Tiancheng Sun, Giljoo Nam, Carlos Aliaga, Christophe Hery, and Ravi Ramamoorthi. Human hair inverse rendering using multi-view photometric data. 2021. 8

[25] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4

[26] Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T Barron, Todd Zickler, and Pratul P Srinivasan. Eclipse: Disambiguating illumination and materials using unintended shadows. *arXiv preprint arXiv:2305.16321*, 2023. 1

[27] 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. 6

[28] 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. 4

[29] 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. 7

[30] 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. 2

[31] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P Srinivasan, Richard Szeliski, Jonathan T Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdfs for real-time view synthesis. *arXiv preprint arXiv:2302.14859*, 2023. 3

[32] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 802–812, 2021. 4

[33] 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. 2, 4

[34] 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, 6

[35] 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. 6