Supplementary Materials for SunStage: Portrait Reconstruction and Relighting using the Sun as a Light Stage

Yifan Wang[1] Aleksander Holynski[1] Xiuming Zhang[2] Xuaner Zhang[2]
[1]University of Washington [2]Adobe Inc.
sunstage.cs.washington.edu

A. Formulation Details

We expand the diffuse and specular contribution derivations below:

Diffuse contribution. The diffuse contribution L_o^d is given by the diffuse terms of Equation 5 in the main paper:

$$\begin{aligned} &L_o^d(x) \\ &= \sum_{\omega_i} V(x,\omega_i) L_i^{\text{amb}}(\omega_i) \odot R^d(x,\omega_i)\,(\omega_i \cdot n(x))\,\Delta\omega_i \\ &+ \sum_{\omega_i} V(x,\omega_i) L_i^{\text{sun}}(\omega_i) \odot R^d(x,\omega_i)\,(\omega_i \cdot n(x))\,\Delta\omega_i \end{aligned} \tag{1}$$

$$= \sum_{\omega_i} V(x,\omega_i) L_i^{\text{amb}}(\omega_i) \odot \frac{a(x)}{\pi}(\omega_i \cdot n(x))\Delta\omega_i$$

$$+ V(x,p^{\text{sun}})k^{\text{sun}}[1,1,1] \odot \frac{a(x)}{\pi}(p^{\text{sun}} \cdot n(x))\Delta p^{\text{sun}} \tag{2}$$

$$\approx \sum_{\omega_i} L_i^{\text{amb}}(\omega_i) \odot \frac{a(x)}{\pi}(\omega_i \cdot n(x))\Delta\omega_i$$

$$+ V(x,p^{\text{sun}})k^{\text{sun}}[1,1,1] \odot \frac{a(x)}{\pi}(p^{\text{sun}} \cdot n(x))\Delta p^{\text{sun}} \tag{3}$$

where $a(x)$ is the albedo at point x, k^{sun} is the (optimized) sun intensity, and p^{sun} is the (optimized) sun direction. The sun is modeled as a directional light source, so the second summation can be simplified to a single term (i.e. only in the direction of p^{sun}). $E \in \mathbb{R}^{16\times32\times3}$ is the high-dynamic-range (HDR) environment map. We ignore the visibility term for the ambient lighting during optimization, as it is computationally intensive to compute the visibility for all light directions and the ambient intensity is much weaker than that of the sun.

Specular contribution. The specular contribution L_o^s at each pixel is given by the specular term of the sun, see Equation 5 in the main paper (recall that we ignore the specular term of the ambient due to its weak contribution):

$$\begin{aligned} &L_o^s(x,\omega_o) \\ &= \sum_{\omega_i} V(x,\omega_i) L_i^{\text{sun}}(\omega_i) \odot R^s(x,\omega_i,\omega_o)\,(\omega_i \cdot n(x))\,\Delta\omega_i \end{aligned} \tag{4}$$

$$\begin{aligned} &= V(x,p^{\text{sun}})k^{\text{sun}}[1,1,1]k_s\frac{s+2}{2\pi}\,(h(p^{\text{sun}},\omega_o)\cdot n(x))^s \\ &(p^{\text{sun}} \cdot n(x))\,\Delta p^{\text{sun}} \end{aligned} \tag{5}$$

where we have substituted Equation 2 in the main paper and reduced the summation to just one term at p^{sun} (i.e. L_i^{sun} is 0 elsewhere).

B. Optimization Details

B.1. Coarse optimization

The pose loss, $L_{\text{pose}} = L_{\text{mask}} + L_{\text{lmk}}$, is optimized using an ADAM optimizer [5] with the following learning rates for different parameters:

$(\beta, \theta_j, \psi_j)$	$1e{-}4$
S (Object scale)	$1e{-}2$
T_j (Object translation)	$1e{-}2$

We optimize for a total of 2,000 epochs on a single NVIDIA RTX 2080 Ti GPU. Each epoch consists of one optimization step for all training images. The coarse optimization takes 2.5 hours to converge for a sequence that contains 200 images with resolution 224×224.

To account for the fact that our geometry does not explicitly model hair and clothing, our mask loss L_{mask} consists of a foreground mask loss $L_{\text{mask\_foreground}}$ which corresponds to skin regions and a background mask loss $L_{\text{mask\_background}}$ which corresponds to background pixels.

$$L_{\text{mask}} = L_{\text{mask\_foreground}} + L_{\text{mask\_background}} \tag{6}$$

where the foreground mask loss is enforcing 1s in the skin region and the background mask loss is enforcing 0s in the background region. Both are L2 losses.

We compute the background mask using the matting model from RVM [6], and the skin mask from a modified

version[1] of BiSeNet [12, 13]. The landmarks used for the landmark loss are obtained from HRNet [9].

B.2. Photometric optimization

The final loss:

$$L = \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{lmk}} L_{\text{lmk}} + \lambda_E L_E + \lambda_{Es} L_{Es} + \lambda_{\text{VGG}} L_{\text{VGG}} + \lambda_{\text{photo}} L_{\text{photo}} \quad (7)$$

is optimized using ADAM for 4,000 epochs. The initial learning rates are

$(\beta, \theta_j, \psi_j)$	$1e{-}4$
ΔX	$1e{-}4$
S	$1e2$
T_j	$1e2$
a	$1e{-}2$
s	$1e{-}2$
k_s	$1e{-}2$
p^{sun}	$1e{-}3$
E	$1e{-}3$
k^{sun}	$1e{-}3$

Every 1,000 epochs, we decay the learning rate of all parameters to 10% of their previous values. The optimization runs on a single NVIDIA RTX 2080 Ti GPU and takes 2 days to converge for a sequence of 200 images of resolution 224×224. The VGG loss L_{VGG} in the photometric optimization follows the implementation of pix2pixHD [10].

C. Capture Details

Our video sequences are captured by an iPhone 12 Pro Max or an iPhone 13 Pro Max. While we have validated that our method works equally well with keypoint-based pose estimation techniques, we opt for using the phone's integrated visual-inertial SLAM system, since we found empirically that it produces more reliable camera orientations. In practice, to capture our sequences, we use CamTrackAR[2], an app that captures synchronized camera intrinsics[3], poses, and video frames.

During capture, we impose constraints on the subject expression to account for the limitations of the face morphable model we use for optimization, e.g. not modeling the teeth (see more discussions in Section H). Prior to capture, our subjects are instructed to try to maintain a constant expression, to face forward, and to rotate in place. Naturally, it is nearly impossible to remain perfectly centered and to keep a constant pose and expression, but fortunately, our formulation is tolerant to variations in both. In fact, our reconstructed geometry can typically model subtle expression variations like smiles and twitches. Still, to reduce noise in optimization, we filter out the input frames in which the subject has an open mouth or is blinking, or frames which contain significant glare (i.e., when the camera is facing the sun).

While a single rotation provides sufficiently many constraints on the shape of the face (through shadows and specular reflections), the quality of the geometry and texture at the boundaries of the face (i.e., at the edge of the jaw and the side of the face) can be improved by capturing additional frames where the camera is rotating independently from the face. In these cases, we capture a full 360 degree rotation, then stop, and continue rotating the camera along a 15 degree arc back and forth. When using this capture technique, we typically sample around 200 frames from the full video for optimization – 100 from the initial 360 degree rotation, and 100 from the arc sequence. Since all images from the arc sequence provide similar photometric constraints (i.e. they do not give strong constraints on material properties, lighting, or normal, given that the lighting conditions are identical), and in order to avoid degenerate optimization, we only optimize for at most five arc images per epoch.

D. Evaluation Details

In this section, we describe how we compare SunStage with ablated variants and baselines, for both the tasks of relighting and view synthesis. Since different baseline approaches make different assumptions, we clarify the necessary adaptations made for fair comparisons. We also provide analysis of the results.

D.1. Held-out frames

In addition to capturing a single 360-degree rotation, we capture two testing sequences: one for evaluating view synthesis, and one for evaluating relighting quality.

Immediately after completing the capture, the subject hands the camera to another person, who captures a multi-view video of the subject (i.e., translating and rotating the camera to capture the face from different viewpoints). The subject remains still during this process. Of these captured frames, the first (which is typically the frame facing the subject head-on), is included in the "training" sequence (and is used as the input for other methods that only operate on single frames), and the remainder are held out as testing images for view synthesis. For the purposes of evaluation, we assume the subject is entirely stationary for this portion of the capture.

Additionally, we ask the subject to capture a second sequence, either in a different location, at a different time of day, or at a different relative angle from the sun. We use

[1] https://github.com/zllrunning/face-parsing.PyTorch

[2] https://fxhome.com/product/camtrackar

[3] since the iPhone camera uses optical image stabilization, and therefore the principal point and focal lengths vary as the camera moves

these examples to evaluate our method's ability on relighting. Both sequences are reconstructed separately, without using a shared model for the subject's geometry, reflectance, or appearance. For evaluation, we swap the estimated lighting conditions between the two models, i.e., we render an image using the estimated sun direction and environment map (as well as a given frame's pose and expression) from sequence A and the subject geometry and materials from sequence B, and compare the result to the corresponding real image from sequence A. These target poses, expressions, and lighting conditions are shared by all ablations and baseline experiments.

D.2. Metrics

To quantitatively compare these methods, we compute relighting and view synthesis errors measured in Peak Signal-to-Noise Ratio (PSNR), Similarity Index Measure (SSIM) [11], and Learned Perceptual Image Patch Similarity (LPIPS) [15]. As Table 1 in the main paper shows, SunStage achieves the best performance in both relighting and view synthesis across all three error metrics.

For the task of relighting, we composite all methods' results onto the ground-truth frames using skin masks extracted using a modified version[4] of BiSeNet [12, 13], since our method does not relight non-skin regions, e.g. hair and clothes. The composited result is used for comparison against the ground-truth images.

D.3. Baseline Comparisons

In the following paragraphs, we provide the details about how we train and test the baseline models, as well as our analysis of these baseline results.

DECA. As a very naïve baseline, we use DECA [3], a single-image face reconstruction method, for the tasks of novel-view synthesis and relighting. From a single image, DECA predicts facial geometry, spherical harmonic lighting, and albedo. For view-synthesis, we run DECA on two images (separately) to get two sets of albedo, geometry, and lighting. To render an image from a new viewpoint, we simply swap the shape code, expression code and albedo from one image to the other. While this gives DECA a significant advantage, since the lighting and pose are estimated directly from the ground-truth frame, we find in practice that the rendered results seldom resemble the ground-truth images. Additionally, although DECA contains a deformation map to model fine details, we find in practice that this seldom accurately models subject-specific geometry details such as wrinkles. One reason for the poor performance at this task is the orthographic assumption. As shown in Figure 6 in the main paper, DECA's pose and shape estimates significantly deteriorate upon introduction of strong perspective effects.

[4] https://github.com/zllrunning/face-parsing.PyTorch

For the task of relighting, we similarly run DECA on a pair of images, and swap the lighting conditions between the two. We find that since DECA assumes a Lambertian model, the resulting images are far from photorealistic.

GCFR. GCFR [4] is a single-image relighting method that aims to handle hard shadows in new lighting scenarios. It predicts a shadow mask from an estimated depth map of the face. We use the pretrained model from GCFR for the baseline comparison.

Given the input and target image, we first use GCFR's Shadow Mask Estimation module to estimate the shading map, then we estimate the albedo map from the input image using GCFR's hourglass network (albedo decoder). To render the target relit image, we compose the *target shading map* (advantageous to the baseline) with input albedo map following Equation 6 in [4].

However, GCFR fails to accurately relight images even with the target shading map. Its hourglass network is not able to estimate a good albedo from the input image — the estimated albedo often has shadows and specular highlights baked in. This is likely because the training dataset of GCFR does not contain enough images with hard shadows.

DPR. DPR [17] is a single-image learning system that operates entirely in the 2D image space. We use the pretrained model from DPR as a baseline for the task of relighting.

For simplicity, we evaluate DPR only on same-environment relighting. In other words, we ask DPR to render the same environment as the input sequence, but under different incident sun angles. To render a relit image, we first run DPR to estimate the spherical harmonics coefficients of the input environment map, then rotate these towards the lighting of the target image, and finally feed those SH coefficients to perform relighting. DPR additionally requires the input and output to be spatially aligned. To facilitate this, we use the geometry estimated by SunStage as a proxy for reprojecting an input frame into the pose of the target frame.

We find that DPR is unable to accurately relight images, failing to synthesize accurate shadows and specularity, largely as a result of the lack of explicit 3D reasoning of the subject.

Total Relighting. Total Relighting [7] (TR) is a state-of-the-art single-image method trained on high-quality light state data, achieving impressive shading and specular highlights.

We use Total Relighting as a baseline for relighting. As input, we provide the same reprojected images as we do for DPR, and provide our target environment maps (the same ones used for evaluating our method) as the target lighting conditions.

Since TR does not model cast shadows explicitly, it of-

ten has difficulty removing cast shadows from input images, and does not produce accurate cast shadows in the relit images. Additionally, Total Relighting produces color tones that do not match the target images. We suspect this occurs as a result of a number of factors: (1) there is an inherent ambiguity in (particularly single-frame) decomposition of lighting and albedo, and Total Relighting may simply be decomposing the two differently when compared to our method, and (2) the lighting maps provided for our quantitative relighting tasks (i.e. those in the main paper, not the HDRI renderings shown in the supplement) are the result of our method's decomposition, and therefore may not match exactly the characteristics of the HDRI images used for training TR. We use the pretrained Total Relighting model (from the authors) to run inference on our images.

NLT. Like SunStage, NLT [16] is a "test-time optimization" approach that learns a subject-specific appearance model from multiple observations of the same subject.

We use NLT as a baseline for both relighting and view synthesis. Since NLT expects that the incoming light directions and viewing directions are known and a geometry proxy is provided, we train NLT on the input images, along with the camera poses, sun directions, and face geometry estimated by SunStage. At test time, we query the trained NLT with novel sun directions for relighting and novel viewpoints for view synthesis.

We find that for both tasks, NLT produces less sharp specular highlights and an overall less accurate rendering than SunStage. NLT produces blurry results and ghosting shadows, likely due to the discrepancy in the number of images used for training and the number of images typically captured by a light stage (NLT has been shown to work on $300 \times 50 = 15,000$ images, as opposed to ours that uses only 200).

NextFace. Similar to SunStage, NextFace [2] is an optimization-based face reconstruction method. It learns to decompose the input image into shape, lighting, and material properties from multiple observations of the same subject.

We evaluate NextFace on both relighting and view synthesis. We first train NextFace on our training images to estimate the shape, lighting, and material properties. For the test set, we train another NextFace model to get the estimated lighting as the target for relighting, and shape parameters as the target for view synthesis. To render the final image for relighting, we use the lighting at test time, and shape and materials at training time. For view synthesis, we use the shape at test time, and lighting and material at training time. We use the ray tracer in NextFace as the renderer.

We find that for both tasks, NextFace fails to model self-cast shadows. This is due to the lighting formulation it adopts. NextFace uses spherical harmonics (SH) to model the scene lighting, which is unlikely to model high-frequency lighting such as the hard sunlight.

D.4. Ablations

In this section, we examine different ablated versions of our method.

Ours w/o coarse. In this ablation, we directly optimize for all parameters without the coarse alignment stage. In practice, we find that optimization seldom converges to a reasonable solution due to the ill-posed nature of our optimization problem: different combinations of geometry, reflectance, lighting, and camera poses may lead to the same observed image.

As Figure 1a shows, the optimization result, without coarse alignment, often gets trapped in local optima. Quantitatively, this ablated version of SunStage falls far behind the full model.

Ours without spatially varying (SV) specular coefficients k_s, s. This ablation study explores how spatially-varying shininess (i.e., the k_s and s parameters of our reflectance model) is critical for recovering a photorealistic facial reflectance model. Since different face regions possess different shininess factors, using a global k_s and s value leads to an averaged solution, where no area is estimated to be strongly shiny to avoid large re-rendering errors.

Another artifact we observe is that the eyeballs are estimated to have overly wide specular lobes (similar to that of the skin), as shown in Figure 1b. Quantitatively, this variant performs worse than our full model but still achieves reasonable errors (ranked the third best for relighting). This is likely due to the fact that the specular component is numerically insignificant.

Ours w/o L_{mask}, L_{lmk}. This model variant ablates the contribution of mask loss and landmark loss in the second (i.e., photometric) stage of our optimization. For this variant, we preserve the first stage solution and disable the mask and landmark losses for the second stage of optimization.

Similar to the "w/o coarse" ablation, this model variant solves a less constrained optimization problem than our full model does, e.g., without the facial keypoints, there is no constrains on the mouth image pixels to align with the geometry corresponding to the mouth. We observe similar qualitative (Figure 1c) and quantitative results (Table 1 in the main paper) as in the "w/o coarse" ablation.

Ours w/o L_{mask}. Similar to the previous ablation, we turn off L_{mask} only in the second stage of our optimization. This ablation suffers from alignment issues as seen before, and therefore we skip its visualization in Figure 1. Quantitatively, as shown in Table 1 in the main paper , this ablation

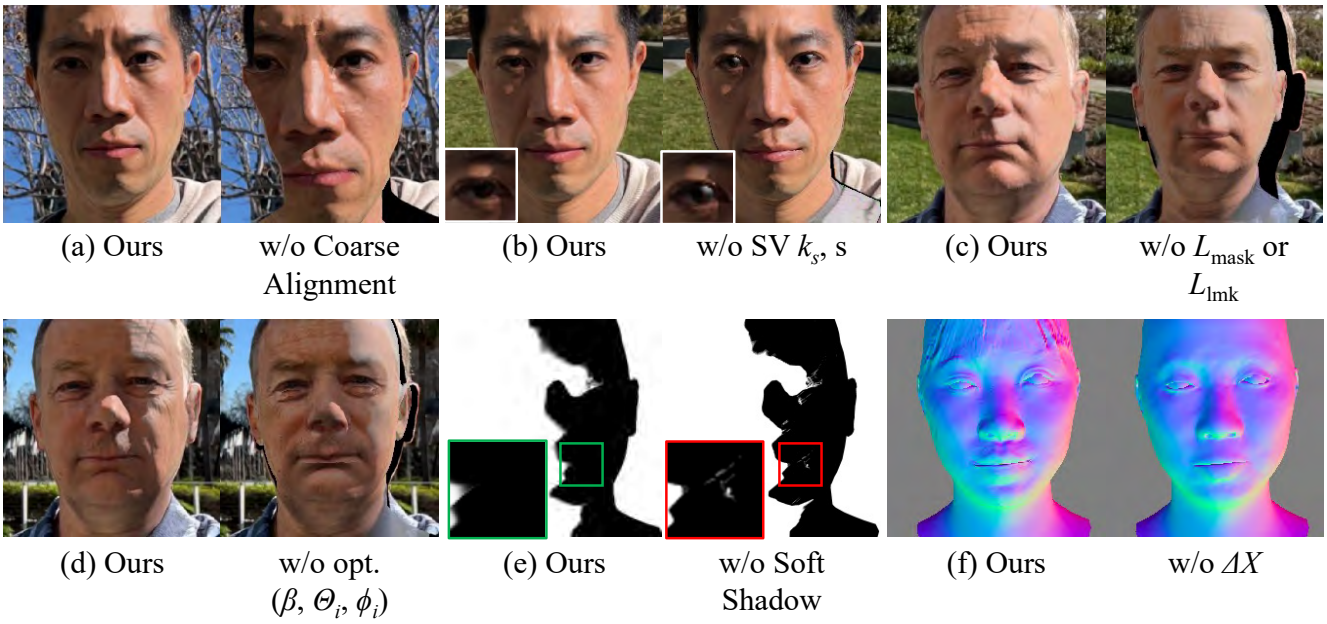


Figure 1. Visualization of common artifacts produced by the ablated versions of SunStage.

outperforms "w/o L_{mask}, L_{lmk}" by having more optimization constraints from L_{lmk} but still underperforms our full model by a large margin.

Ours w/o opt. $(\beta, \theta_i, \phi_i)$. In this experiment, we preserve the initialized shape provided by DECA without refining it. Since the images captured under our setup are selfies, the perspective effects are not accounted for by DECA, which assumes an orthographic camera model. As such, the shape estimated by DECA is not well-aligned with our input images. As shown in Figure 1d and Table 1 in the main paper , our shape optimization strategy improves the initialized DECA shape.

Ours w/o soft shadow. In this variant, instead of doing the soft comparison as Equation 8 in the main paper states, we use a hard z-buffer comparison in producing the shadow maps. Although Table 1 in the main paper shows that this ablated version of our model achieves reasonable quantitative performance, as Figure 1e demonstrates, using a hard comparison produces spurious shadows, especially when the sun is at grazing angles. Additionally, the optimized shadows (and the sun position) are less accurate, which is likely due to the instability in optimization as the gradients are not continuous for the hard shadow comparison formulation.

Ours w/o ΔX. We also explore the quality of our method without optimizing for a displacement map. As Figure 1f illustrates, this ablation is unable to model geometric details such as wrinkles and pores. Consequently, such effects are baked into the albedo, causing artifacts in applications such as relighting and material editing. Additionally, this ablation produces blurrier renderings, since the high-frequency appearance change is harder to be explained by other factors such as reflectance and lighting.

Figure 2. Comparison on different reflectance models. A more complex model [8] does not significantly improves visual quality. On the other hand, it is hard to optimize and introduces instability in training.

Ours w/ microfacet reflectance model. Finally, we change the Blinn-Phong reflectance model to a more complex microfacet reflectance model [8]. As shown in Figure 2, microfacet model produces comparable specular highlights

Figure 3. **Synthetic OLAT**. By rendering the recovered face with a single distant light source (where geometry artifacts are exposed), we can simulate the One-Light-at-A-Time (OLAT) data that was only possibly captured with a light stage.

with that from the Blinn-Phone model. The microfacet model [8] describes the complicated light paths that depend on incoming light direction, surface normal and material properties. The gradients on these parameters, which contribute to multiple terms in the equation, are more noisy comparing to the simple Blinn-Phong reflectance model. Using the same optimization scheme as in Blinn-Phong, we find it impossible for the scene parameters converge to a reasonably steady state. Therefore, we turn off the specular highlights for the first 100 epochs to reduce the parameter entanglement, and start optimizing all variables in the microfacet model once the light (i.e., the sun) position is converged. We observe little visual quality difference between using microfacet and Blinn-Phong reflectance models, while the former involves a much more unstable and difficult optimization scheme. SunStage thus uses the simple Blinn-Phong reflectance model.

E. Additional Results

In Figure 5, Figure 6, Figure 7 and Figure 8, we show more comparisons with Neural Video Portrait Relighting (NVPR) [14] and Total Relighting (TR) [7]. Both NVPR and TR are image based relighting methods which leverage the priors learned from light stage data. At test time, the model takes in an arbitrary input image and a target HDR environment map, and generates a relit result. We find that neither of the baselines fully preserves the identity of the subject, changing facial geometry or missing some of the detailed reflectance properties (e.g. accurate specular highlights) that are unique to each individual subject. Both NVPR and TR also leave harsh traces on the relit results at the locations where the shadow boundary exists in the input image (see Figure 5 row 1 and row 3). This is likely a result of the lack of such images (i.e., with harsh shadows) in the training dataset.

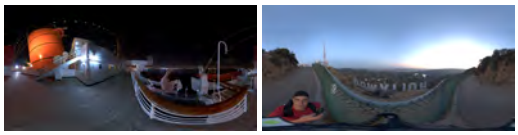


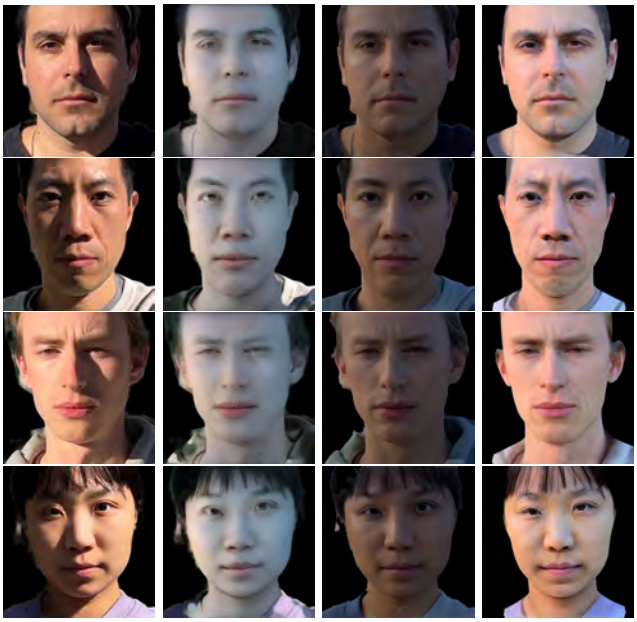
Figure 4. Target lighting: Queen Mary (left) and Hollywood (right) for the following comparisons with Neural Video Portrait Relighting and Total Relighting. Both environment maps are tone mapped for visualization.

F. Applications

OLAT. To further validate the quality of the reconstructed geometry and the material properties, we simulate the One-Light-at-A-Time (OLAT) lighting setup typically seen in light stage captures [1]. Our results in Figure 3 show that we can plausibly recreate this challenging lighting setup, which typically exposes most errors in the estimated geometry and reflectance.

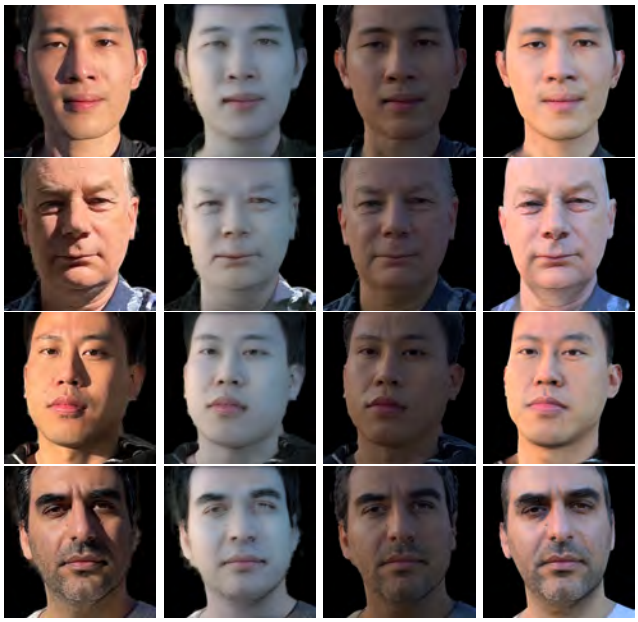
Relighting: Soften shadows. To soften harsh shadows, we increase the size of the light source by applying a random offset j to the (optimized) sun position p_{sun}. This offset can be interpreted as the radius of a virtual area light – j controls the size of the light and thus the softness of the shadow. We sample n new sun positions, and average these n renders to produce the resulting rendering with softened shadows.

Relighting: Lighting replacement. We can use the learned properties to realistically render the subject with a new input environment map. The input environment map is downsampled to 16×32. To render, each pixel in the environment map is treated as a directional light source. The diffuse and specular contribution is calculated following Equation 6 and Equation 7 in the main paper.



(a) Input (b) NVPR [14] (c) TR [7] (d) Ours

Figure 5. Comparison with Neural Video Portrait Relighting and Total Relighting on the target lighting "Hollywood". Both NVPR and TR leverage face priors by training on large-scale light stage data. While being able to generalize to an arbitrary input, these methods can not model the individual skin reflectance properties. Both NVPR and TR also leave traces of visible shadow boundaries from the input images.



(a) Input (b) NVPR [14] (c) TR [7] (d) Ours

Figure 6. Comparison with Neural Video Portrait Relighting and Total Relighting on the target lighting "Hollywood". Both NVPR and TR leverage face priors by training on large-scale light stage data. While being able to generalize to an arbitrary input, these methods can not model the individual skin reflectance properties. Both NVPR and TR also leave traces of visible shadow boundaries from the input images.

Relighting: Different time of day. We show that we are able to simulate relit faces from an arbitrary time of day, including the fleeting "golden hour" and "blue hour" lighting that is favoured by many portrait photographers. To do so, we look up the correlated color temperature (in Kelvin) for different time of day, convert the color temperature into color matrices in the sRGB space, and use these to change the color of the sun in rendering.

View Synthesis. We can change the (optimized) camera parameters to synthesize novel views of the subject. We can also render the subject with different amounts of perspective effects by changing the (optimized) camera focal length. In practice, we linearly scale the focal length and the subject distance to preserve the size of the face in the frame, as is done in a dolly zoom.

G. Visualization Details

G.1. Compositing background

We use different compositing methods to combine the rendered foreground subject and background for different applications.

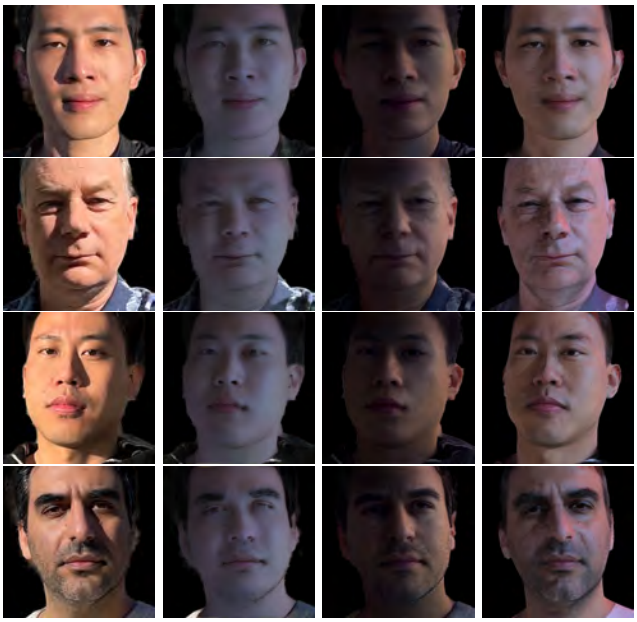
Black background. We use a black background (i.e. do not do compositing) for the One-Light-At-a-Time rendering (Figure 3), to mimic the capture setup of a light stage. We also use a black background for more dramatic lighting setups that are similar to studio lighting, like the blue fill light shown in Figure 1b in the main paper. We find that significant changes in color to the original scene's lighting tend to look unrealistic when composited onto the original background.

Original background. Whenever possible, we use the original input image as the background in compositing. Note that the original background contains a portion of the hair that is not modeled physically, and thus does not respect changes in lighting, viewpoint, or other parameters. As such, the cases in which we can realistically composite onto the original background are limited, and only include shadow softening and subtle changes to lighting direction and magnitude.

Panorama background. For the remainder of cases, when we would like the subject to remain in the original scene, but the lighting or viewpoint have changed significantly from

(a) Input (b) NVPR [14] (c) TR [7] (d) Ours

Figure 7. Comparison with Neural Video Portrait Relighting and Total Relighting on the target lighting "Queen Mary". Both NVPR and TR leverage face priors by training on large-scale light stage data. While being able to generalize to an arbitrary input, these methods can not model the individual skin reflectance properties. Both NVPR and TR also leave traces of visible shadow boundaries from the input images.



(a) Input (b) NVPR [14] (c) TR [7] (d) Ours

Figure 8. Comparison with Neural Video Portrait Relighting and Total Relighting on the target lighting "Queen Mary". Both NVPR and TR leverage face priors by training on large-scale light stage data. While being able to generalize to an arbitrary input, these methods can not model the individual skin reflectance properties. Both NVPR and TR also leave traces of visible shadow boundaries from the input images.

the observed input frames, we instead composite the subject onto a panorama of the original scene. This panorama is automatically stitched from the input video frames (masking out the subject in each frame, i.e., $I_j \cdot (1 - I_{\text{mask}})$). See examples in Figure 9.

H. Discussions and Limitations

Physical model. Our method inherits the limitations of existing morphable models that do not model hair, teeth, clothes, or accessories. Figure 10 (a, b) shows a reconstruction that does not model the hair, and Figure 10 (c, d) shows an example where the reconstruction fails to model the clothes.

Capture. Our capture setup is not always comprehensive enough to model the full reflectance of the face. There are regions of the face that may not observe changes in lighting during the entire capture, like the bottom of the chin, which is often under shade. This causes ambiguity in our reconstruction, since the observed color can be explained by different combinations of albedo and lighting. Shown in Figure 10 (e, f) and (g, h), this can result in highlights baked into the albedo.

Additionally, our optimization makes assumptions about the scene lighting: 1) the sun must be the dominant light source (i.e. the method does not work for a cloudy day capture), Figure 10 (i, j) shows an example where the video captured under a cloudy day does not produce a reasonable albedo, as the face is always observed under shade, without any specular or shadow constraints. and 2) the sun's color temperature must be roughly in the range of 5500K-6500K (i.e. daylight around noon). Otherwise our reconstruction can not resolve the ambiguity between the illuminant and albedo. Figure 10 (k, l) shows a video captured at golden hour, a strongly tinted lighting, which induces ambiguity in the recovered albedo. The result of the same identity captured under the required lighting condition is shown in Figure 10 (g, h), which has a much more reasonable albedo reconstruction.

Figure 9. Example stitched panorama used for video background composite.

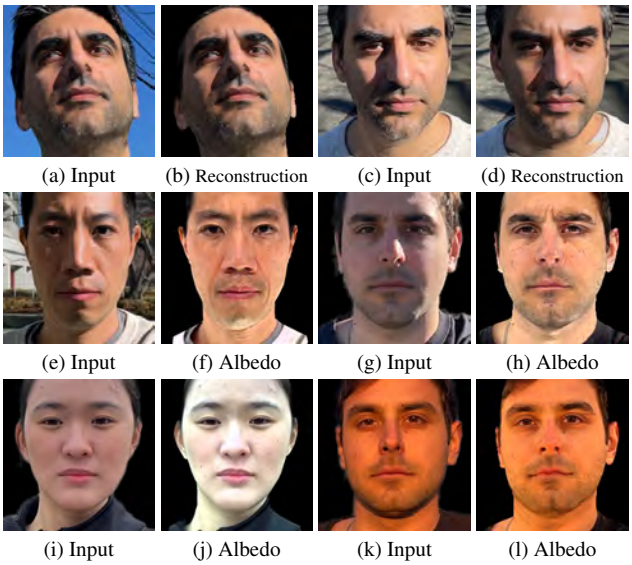


Figure 10. **Limitations**. SunStage has limitations in its physical model and capture setup. (a, b) and (c, d) show reconstructions that fail to model hair and clothes. (e, f) and (g, h) illustrate the reconstructed albedo entangled with highlights around the chin region, which does not see lighting variations at capture time. (i, j) and (k, l) show failure cases when the capture lighting requirement breaks. The former is captured under a cloudy day and the latter is captured under a strongly tinted lighting condition. (k, l) shows the same identity as in (g, h) captured under different lighting conditions. The difference in the predicted albedo demonstrates the albedo-illuminant ambiguity, and the need for the assumption of mid-day (or otherwise known) sun color.

References

[1] 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. 6

[2] 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. 4

[3] 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. 3

[4] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2022. 3

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 1

[7] 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. 3, 6, 7, 8

[8] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007. 5, 6

[9] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[11] 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. 3

[12] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 2, 3

[13] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2, 3

[14] 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. 6, 7, 8

[15] 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. 3

[16] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021. 4

[17] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7194–7202, 2019. 3