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

Yifan Wang, Aleksander Holynski, Xiuming Zhang, Xuaner Zhang

1University of Washington 2Adobe Inc.
sunstage.cs.washington.edu

Abstract

A light stage uses a series of calibrated cameras and lights to capture a subject’s facial appearance under varying illumination and viewpoint. This captured information is crucial for facial reconstruction and relighting. Unfortunately, light stages are often inaccessible: they are expensive and require significant technical expertise for construction and operation. In this paper, we present SunStage: a lightweight alternative to a light stage that captures comparable data using only a smartphone camera and the sun. Our method only requires the user to capture a selfie video outdoors, rotating in place, and uses the varying angles between the sun and the face as guidance in joint reconstruction of facial geometry, reflectance, camera pose, and lighting parameters. Despite the in-the-wild un-calibrated setting, our approach is able to reconstruct detailed facial appearance and geometry, enabling compelling effects such as relighting, novel view synthesis, and reflectance editing.

1. Introduction

A light stage [11] acquires the shape and material properties of a face in high detail using a series of images captured under synchronized cameras and lights. This captured information can be used to synthesize novel images of the subject under arbitrary lighting conditions or from arbitrary viewpoints. This process enables a number of visual effects, such as creating digital replicas of actors that can be used in movies [1] or high-quality postproduction relighting [46].

In many cases, however, it is often infeasible to get access to a light stage for capturing a particular subject, because light stages are not easy to find: they are expensive and require significant technical expertise (often teams of people) to build and operate. In these cases, hope is not lost — one can turn to methods that are trained on light stage data, with the intention of generalizing to new subjects. These methods do not require the subject to be captured by a light stage but instead use a machine learning
model trained on a collection of previously acquired light stage captures to enable the same applications as a light stage, but from only one or several images of a new subject [6, 25, 30, 38, 40, 50, 52]. Unfortunately, these methods have difficulty faithfully reproducing and editing the appearance of new subjects, as they lack much of the signal necessary to resolve the ambiguities of single-view reconstruction, i.e., a single image of a face can be reasonably explained by different combinations of geometry, illumination, and reflectance.

In this paper, we propose an intermediate solution — one that allows for personalized, high-quality capture of a given subject, but without the need for expensive, calibrated capture equipment. Our method, which we dub SunStage, uses only a handheld smartphone camera and the sun to simulate a minimalist light stage, enabling the reconstruction of individually-tailored geometry and reflectance without specialized equipment. Our capture setup only requires the user to hold the camera at arm’s length and rotate in place, allowing the face to be observed under varying angles of incident sunlight, which causes specular highlights to move and shadows to swing across the face. This provides strong signals for the reconstruction of facial geometry and spatially-varying reflectance properties. The reconstructed face and scene parameters estimated by our system can be used to realistically render the subject in new, unseen lighting conditions — even with complex details like self-occluding cast shadows, which are typically missing in purely image-based relighting techniques, i.e., those that do not explicitly model geometry. In addition to relighting, we also show applications in view synthesis, correcting facial perspective distortion, and editing skin reflectance.

Our contributions include: (1) a novel capture technique for personalized facial scanning without custom equipment, (2) a system for optimization and disentanglement of scene parameters (geometry, materials, lighting, and camera poses) from an unaligned, handheld video, and (3) multiple portrait editing applications that produce photorealistic results, using as input only a single selfie video.

2. Related works

Face modeling. Extensive research has been devoted to the modeling of human faces, leading to various 3D morphable models (3DMMs) [2, 3, 5–7, 9, 10, 25, 31–33, 43]. These models are parametric (maybe in the form of neural networks [33]), allowing one to express variations compactly with a vector. They also encode strong priors learned from real scans. The groundbreaking face 3DMM is that of Blanz and Vetter [3] containing models for shape, expression, and appearance (the Phong model). Also influential is the FLAME model [25] that uses vertex-based Linear Blend Skinning (LBS). FLAME is described by a mapping from shape, pose, and expression vectors to a list of vertices. We refer the reader to the survey by Egger et al. [13] for different face morphable models.

Such parametric face models provide a low-dimensional space for optimization or learning algorithms. DECA [14] uses the FLAME model to estimate detailed facial geometry (and albedo) from single images, by predicting additional displacement maps and adding them to the estimated FLAME models. More recently, NextFace [12] employs the 3DMM geometry and albedo priors to learn an albedo residual that captures more facial details.

Without modeling 3D face geometry, researchers have also achieved photorealistic synthesis of portrait images using generative models and large-scale high-quality image datasets [22, 23].

Light stage capture. The light stage as described in Debevec et al., achieves impressive portrait reconstruction and relighting by capturing a series of images of the face under varying illumination [11]. Subsequent work made this process faster, more efficient, and explored different types of illuminants [15, 16, 28].

Given that a light stage is not always accessible, a number of methods have been proposed to achieve similar outputs from a single (or few) input portrait images [18, 19, 29, 30, 40, 41, 48–50]. These method rely on a dataset of light stage captures or synthetic examples as training data.

Our setup can be thought of as a “minimalist light stage” formed by just the sun and a rotating camera, without requiring the high construction and maintenance costs of building a light stage. This parameterization of a sun and skylight model has been shown to be effective in photometric stereo [17, 21] and scene factorization [27, 42]. In a similar spirit, Calian et al. [8] focus on lighting estimation using faces as “light probes”. Sengupta et al. propose to circumvent the need for a complicated light stage by recording the facial appearance responses to varying contents displayed on a desk monitor, and then perform portrait relighting [37]. Sevastopolsky et al. also attempt to simplify the capture setup from a light stage to a mobile phone camera with a co-located flash [39]. Unlike our work, which is physically-based, their approaches use neural rendering, and therefore have less direct control over lighting, material, and scene parameters.

3. Overview

Our method targets accurate reconstruction of scene lighting, subject geometry, and material properties from a handheld video sequence of a person rotating in place under the sun. Given a selfie video, we take a test-time optimization approach that uses the information from all frames of the video to solve for a physical model of the
scene: the geometry and material properties of the face, scene lighting, and camera parameters (Fig. 2). This physical model consists of a base face shape parameterized by a low-dimensional deformable model $X^b$, a displacement map $\Delta X$, a reflectance model with diffuse $R^d$ and specular components $R^s$, scene lighting $L$, and a perspective camera $C$. These components are explained in detail in Sec. 4.

After this model has been recovered, we can modify the scene and the subject parameters to re-render images. We show several editing applications in Sec. 7: editing skin reflectance, relighting with arbitrary environment map, improving harsh lighting conditions (by softening shadows and adding fill lights), and adjusting camera parameters to change viewpoint or manipulate perspective effects.

4. Formulation

Given an input video, our system reconstructs the parameters of a physical model: i.e., geometry and reflectance of the subject, the scene lighting parameters, and the camera parameters. In this section, we detail all of these parameters and describe the rendering process that turns these parameters into an image.

Geometry. We denote $X_j$ as the full mesh of the subject for frame $j$, composed of a per-frame coarse mesh $X^b_j(\beta, \theta_j, \psi_j)$ and a global displacement map $\Delta X$. The coarse mesh $X^b_j$ is a FLAME deformable face model [25] defined by global shape code $\beta$, per-frame pose code $\theta_j$, and per-frame expression code $\psi_j$. $X^b_j$ also contains per-vertex UV coordinates, which maintain correspondence across variations in $\theta_j$ and $\psi_j$. As such, we model all our global (per-subject) spatially varying parameters in UV space, and sample values per-fragment when rasterizing.

The displacement map $\Delta X$ is used to model fine details like wrinkles that cannot be represented by $X^b_j$. We displace the coarse geometry by $\Delta X$ at rasterization time, by sampling a displacement value per-fragment and displacing each fragment along the surface normal $N_j$. After displacement, the updated fragment positions are used to compute a new surface normal $N'_j$.

$X^b_j$ is optimized per-frame, since it accounts for subtle (and unavoidable) variations in expression and head pose during the capture, which are modeled by $\theta_j$ and $\psi_j$, $\Delta X$, on the other hand, is optimized in UV-space (i.e., globally per-subject), since the deformations it encodes are invariant to the changes in expression or pose. Formally, the final geometry $X_j$ is given by:

$$X_j = X^b_j(\beta, \theta_j, \psi_j) + \Delta X \odot N_j$$

where $\odot$ is the Hadamard product.

Reflectance. We model the skin reflectance, denoted as $R(x, \omega_i, \omega_o) \in \mathbb{R}^3$, where $x$ is a 3D point on the face geometry $X$, $\omega_i$ is the incoming light direction, and $\omega_o$ is the
outgoing direction, using a diffuse and a specular component: \( R = R^d + R^s \).

The diffuse component \( R^d(x, \omega_i, \omega_o) \in \mathbb{R}^3 \) is a Lambertian reflectance model consisting of an albedo map, \( a \), which we optimize as a per-subject UV-space image. For the skin’s specular component, we use the Blinn-Phong model [4].

\[
R^s(x, \omega_i, \omega_o) = k_s \frac{s+2}{2\pi} (h(\omega_i, \omega_o) \cdot n(x))^s
\]  

(2)

where \( h(\omega_i, \omega_o) = \text{normalize}(\omega_i + \omega_o) \) is the half vector, \( k_s \) is the specular intensity, \( s \) is the specular exponent, and \((s+2)/(2\pi)\) is the normalization term for the reflection lobe to integrate to 1. Following [47], we segment the UV-space map into 10 segmented specular reflectance clusters. We then optimize for a spatially-varying pair of values \((s, k_s)\) per-cluster, enabling varying shininess across the face.

While Blinn-Phong does not model many complex effects such as subsurface scattering, our experiments with other models for facial reflectance, such as microfacet models [44], show no significant quality improvements, and often introduce unstable training. More analysis is provided in the supplementary material.

**Lighting.** We use a sun-sky model to represent lighting as the sum of an “ambient” environment map and the sun: \( L_i(x, \omega_i) = L_{i,\text{amb}}(\omega_i) + L_{i,\text{sun}}(\omega_i) \). Note neither \( L_{i,\text{amb}}(\omega_i) \) nor \( L_{i,\text{sun}}(\omega_i) \) depends on the 3D point \( x \), since we model both as directional lights. Optimization-wise, our lighting parameters consist of a \( 16 \times 32 \times 3 \) environment map for ambient lighting, the sun direction \( p_{\text{sun}} \in S^3 \), and the scalar sun intensity \( k_{\text{sun}} \). We fix the sun color to white \([1, 1, 1] \) in our lighting model to resolve the albedo-illumination ambiguity.

### 4.1. Rendering

We calculate the outgoing radiance \( L_o \) at 3D location \( x \) as viewed from viewing direction \( \omega_o \) as:

\[
L_o(x, \omega_o) = \int_{\Omega} V(x, \omega_i) L_i(x, \omega_i) \odot R(x, \omega_i, \omega_o) (\omega_i \cdot n(x)) d\omega_i
\]  

(3)

\[
= \sum_{\omega_i} V(x, \omega_i) \left( L_{i,\text{amb}}(\omega_i) \odot R^d(x, \omega_i) \right. \\
+ \left. L_{i,\text{sun}}(\omega_i) \odot R^d(x, \omega_i) + L_{i,\text{amb}}(\omega_i) \odot R^s(x, \omega_i, \omega_o) \\
+ L_{i,\text{sun}}(\omega_i) \odot R^s(x, \omega_i, \omega_o) \right) (\omega_i \cdot n(x)) \Delta \omega_i
\]  

(4)

\[
+ \sum_{\omega_i} V(x, \omega_i) \left( \right. \\
+ \left. L_{i,\text{sun}}(\omega_i) \odot R^s(x, \omega_i, \omega_o) \right) (\omega_i \cdot n(x)) \Delta \omega_i
\]  

(5)

where \( V(x, \omega_i) \) is the light visibility at \( x \) from \( \omega_i \), and \( L_i(x, \omega_i) \) is the incoming radiance reaching \( x \) from \( \omega_i \). We ignore the specular reflection caused by the ambient lighting, i.e., \( L_{i,\text{amb}}(\omega_i) \odot R^s(x, \omega_i, \omega_o) \), since it is much weaker than the specular reflection of the sun. In the next subsections, we will group the terms into a diffuse contribution \( L_{i,\text{diffuse}}^d \) and a specular contribution \( L_{i,\text{specular}}^s \; L_o = L_{i,\text{diffuse}}^d + L_{i,\text{specular}}^s \). For the final rendered color value, we apply the Reinhard operator [35] and a gamma correction of \( \gamma = 2.2 \) to \( L_o \) to convert from linear to sRGB space.

**Diffuse contribution.** The diffuse contribution \( L_{i,\text{diffuse}}^d \) is then given by only the diffuse terms of Equation 5:

\[
L_{i,\text{diffuse}}^d(x) = \sum_{\omega_i} L_{i,\text{amb}}(\omega_i) \odot \frac{a(x)}{\pi} (\omega_i \cdot n(x)) \Delta \omega_i
\]  

\[
+ V(x, p_{\text{sun}}) L_{i,\text{sun}}[1, 1, 1] \odot \frac{a(x)}{\pi} (p_{\text{sun}} \cdot n(x)) \Delta p_{\text{sun}}
\]  

(6)

where \( a(x) \) is the albedo at point \( x \), \( k_{\text{sun}} \) is the (optimized) sun intensity, and \( p_{\text{sun}} \) 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, only in the direction of \( p_{\text{sun}} \). We additionally optimize for a high-dynamic-range (HDR) environment map \( E \in \mathbb{R}^{16 \times 32 \times 3} \), from which values of \( L_{i,\text{amb}} \) are sampled.

**Specular contribution.** The specular contribution \( L_{i,\text{specular}}^s \) at each pixel is given by only the specular term due to the sun in Equation 5 (recall that we ignore the specular ambient term due to its weak contribution):

\[
L_{i,\text{specular}}^s(x, \omega_o) = V(x, p_{\text{sun}}) L_{i,\text{sun}}[1, 1, 1] k_s \frac{s+2}{2\pi} (h(p_{\text{sun}} \cdot \omega_o) \cdot n(x))^s (p_{\text{sun}} \cdot n(x)) \Delta p_{\text{sun}}
\]  

(7)

where we have substituted Equation 2 and reduced the summation to just one term at \( p_{\text{sun}} \) (since \( L_{i,\text{amb}} \) is 0 elsewhere).

**Shadow map.** In order to generate a map of self-occluded shadows, we perform two passes of rasterization: first, we render a z-buffer from a virtual orthographic camera aligned with the sun direction, \( p_{\text{sun}} \), and then, when rasterizing a given camera viewpoint, compare all fragment positions \( d_{\text{hit}} \) to the light’s z-buffer \( d_{\text{shadow}} \). To avoid precision issues and ensure smooth gradients for back-propagation, we implement a soft comparison as follows in generating shadow/visibility maps:

\[
V(x, \omega_i) = 1 - \text{sigmoid}(k(d_{\text{hit}} - d_{\text{shadow}} \times b))
\]  

(8)

where \( k \) is the falloff slope, and \( b \) the tolerance. We use \( k = 800 \) and \( b = 0.0015 \).

### 5. Optimization

The described physical model contains a large number of parameters to be optimized, controlling scene elements like lighting, geometry, pose, and texture. Unfortunately,
 naïvely optimizing all these parameters from scratch does not result in an optimal solution, since the final observed appearance of the face can often be explained variously through changes to geometry, material properties, lighting or camera parameters, making optimization severely under-constrained and ambiguous. Therefore, we adopt a two-stage optimization approach, through which parameters are gradually enabled. In this section, we describe this process and the relevant losses that guide optimization.

5.1. Coarse alignment

Our system begins by using an off-the-shelf network (DECA [14]) to generate, for each input image, a set of shape parameters \( \beta \), pose parameters \( \theta_j \), and expression parameters \( \psi_j \) of a FLAME face model [25], as well as the relative pose parameters of the virtual camera observing the 3D face. Unfortunately, as with many other single-image facial geometry estimators, DECA assumes an orthographic projection model and therefore cannot accurately recover geometry for our selfie capture sequences, which contain heavy perspective effects (Figure 6). Without a good initialization for geometry, optimization of lighting and material properties seldom converges to an optimal solution due to the heavily ambiguous nature of our optimization problem.

To circumvent this issue, we employ a first stage of optimization where we only optimize for the parameters of a perspective camera (with a known focal length, extracted from input metadata) and the face geometry parameters \( \beta, \theta_j, \psi_j \). As initialization for this optimization process, we use the predicted DECA values for each frame’s pose \( \theta_j \) and expression \( \psi_j \), but set all frames to the average predicted shape \( \beta_{\text{avg}} = \frac{1}{N} \sum_{j} \beta_j \), since the identity remains constant across all frames. To convert DECA’s orthographic camera to a perspective camera, we additionally optimize for an unknown scale \( S \) and translation \( T_j \), which are initialized to empirically chosen values \( S = 2.6e4, T_j = (0, 0, 1.5e5) \). During optimization, the face shape \( \beta \) and scale \( S \) are shared across all frames, while camera pose \( T_j \), expression \( \psi_j \), and pose \( \theta_j \) are optimized per-frame. Note that DECA controls the relative orientation of the camera and the face by varying the pose code \( \theta \) instead of the camera rotation. We adopt this formulation and keep the camera orientation fixed relative to the face. The global orientation of the camera at each frame (and therefore the face) is extracted from the capture video, either through a structure-from-motion system or IMU measurements commonly available on a smartphone.

We use two losses to guide this optimization: a mask loss \( L_{\text{mask}} \) and a landmark loss \( L_{\text{link}} \). The FLAME model includes 3D facial landmark points, corresponding to the standard 68-point facial landmarks set [36] used in facial tracking. Our landmark loss minimizes the L1 distance between the 2D projection of these 3D landmarks (into the input camera viewpoint) and 2D landmarks estimated from the input frame by a 2D landmark detector HRNets [45].

The facial landmarks provide a strong constraint on facial feature alignment, but are sparse, and therefore cannot constrain the overall shape or boundary of the mesh. To supplement it, we include a silhouette loss \( L_{\text{mask}} \), which penalizes the L2 difference between the rasterized mask of the mesh \( I_{\text{sil}} \) and the semantic segmentation mask \( I_{\text{mask}} \) of the input image, using an off-the-shelf semantic segmentation network [26] trained to segment humans in portrait photographs.

The final pose loss is then: \( L_{\text{pose}} = L_{\text{mask}} + L_{\text{link}} \), optimized using an ADAM optimizer [24]. See supplemental for optimization parameters.

5.2. Photometric optimization

Once the 3D model and camera parameters are approximately aligned, we proceed to the second stage of optimization, in which we optimize the precise facial geometry, lighting, and reflectance properties. All the parameters optimized in the first stage (Section 5.1) remain as free variables. In total, the parameters optimized during this stage include: Lighting parameters: \( \rho^\text{sun} \), the global sun direction, \( E \), the global environment map, \( k^\text{sun} \), the global sun intensity, Facial geometry parameters: \( \beta \), the global FLAME shape code, \( \psi_j \), the per-frame expression code, \( \theta_j \), the per-frame pose code, \( \Delta X \), the global deformation map, Material properties: \( s \), the global, spatially-varying specular intensity, \( \alpha \), the global, spatially-varying specular roughness, Camera pose parameters: \( T_j \), the per-frame perspective camera translation, and \( S \), the global scene scale.

During optimization, we randomly select a frame \( j \), render the face using a differentiable rasterizer [34] and the equations described in Section 4 to get the rendered image \( \hat{I} \). In addition to the previously defined landmark and mask losses, we include L2 and VGG [20] photometric losses, comparing the original and reconstructed images:

\[
L_{\text{photo}} = || \hat{I}_j - (I_{\text{sil}} - I_j \cdot I_{\text{mask}}) ||_2
\]

We also include an L2 regularization \( L_E \) and L2-smoothness regularization \( L_{Es} \) on the reconstructed environment map, to encourage the majority of the lighting to be explained by direct sunlight and to aid in disentanglement of the sun and ambient lighting. The total optimized loss becomes:

\[
L = \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{link}} L_{\text{link}} + \lambda_E L_E + \lambda_{Es} L_{Es} + \lambda_{\text{VGG}} L_{\text{VGG}} + \lambda_{\text{photo}} L_{\text{photo}}
\]

with \( \lambda_{\text{mask}}, \lambda_{\text{link}} = 0.05, \lambda_{\text{VGG}} = 0.005, \lambda_E = 0.01, \lambda_{Es}, \lambda_{\text{photo}} = 1. \) Additional optimization details are provided in the supplemental materials.
Figure 3. **Qualitative: Relighting.** A comparison of our method at rendering a new (unseen) lighting environment (h). Our method is able to realistically synthesize the novel lighting condition, including cast shadows and specularity, and nearly matches the (unseen) target reference image. See supplement for additional details on experimental setup and analysis of results.

Figure 4. **Qualitative: view synthesis.** A comparison of our method at the task of generating an image from an unseen viewpoint (f), having only seen a limited collection of input viewpoints. See supplement for more details on experimental setup and analysis of results.

### 6. Evaluation

In this section, we detail quantitative and qualitative experiments comparing our approach with state-of-the-art methods and ablated variants of our method.

**Baseline comparisons.** We evaluate our method on the tasks of novel-view synthesis and relighting. For novel-view synthesis, we compare our method with DECA [14], Neural Light Transport (NLT) [51], and NextFace [12]. For relighting, we compare with DECA, NLT, NextFace, GCFR [18], image-based methods Deep Single Image Portrait Relighting (DPR) [52], Total Relighting (TR) [30] and NVPR [49]. Additional comparisons and details on the experimental setups are provided in the supplemental materials.

We present qualitative comparisons for relighting in Figure 3 and novel view synthesis in Figure 4. Quantitative comparisons on these images are provided in Table 1. These testing images consist of (1) a multi-view capture of the face, in which the subject remains still and the camera is moved to novel viewpoints in the same environment as the original capture, and (2) front-facing sequences in novel environment lighting and unseen sun positions. All testing images are not seen during training of our method, NLT or NextFace. The results shown in Figures 3 and 4 as well as Table 1 clearly demonstrate that our method outperforms all the baselines at both relighting and view synthesis. Single-image methods (DECA, GCFR, DPR, TR) can generalize to other subjects, but fail to recover more faithful and physically accurate facial details. Comparison with multi-image methods (NLT, NextFace) demonstrates that SunStage is a better reconstruction system. Additional analysis of the comparisons is provided in the supplemental materials.
Diffuse Specular Normal Albedo Ambient Reconstruction Ground Truth

Figure 5. Decomposition. We show all the components which comprise our final rendered image to demonstrate that our method not only closely recreates the ground truth image (reproducing realistic highlights and shadows), but also performs a meaningful decomposition of lighting components and facial geometry. Note that our reconstructed surface normals include high frequency details specific to each subject, like wrinkles and birthmarks, which are used in computation of the shadows and specular reflections.

Disentanglement. In Figure 5, we demonstrate how SunStage decomposes the appearance of a portrait photograph into different components: specular, diffuse, and ambient. We also visualize the surface normal, albedo, and other intermediate representations to show that our method is able to effectively recover a physically plausible reconstruction of the real world and disentangle the different components that contribute to the final appearance. We further validate the quality of the reconstructed geometry and materials in the supplementary material.

Ablation studies. In addition to our comparisons with the state-of-the-art baselines, we also compare with ablated variants of our own method. In particular, we include seven such experiments in Table 1: our method (1) without the initial first stage of coarse geometric alignment, i.e., directly optimizing both geometric and photometric parameters from the start, (2) without the spatially varying specular parameters, instead using a single global scalar $s$ and $k_s$, (3) without the geometric alignment losses $L_{mask}$ and $L_{lmk}$, (4) without just $L_{mask}$, (5) without shape optimization, i.e., keeping the initial shape code predicted by DECA, (6) without soft shadow computation, i.e., using a hard z-buffer comparison to compute a shadow map instead of our soft

<table>
<thead>
<tr>
<th></th>
<th>Relighting</th>
<th>Novel view synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
</tr>
<tr>
<td>DECA [14]</td>
<td>16.41</td>
<td>0.69</td>
</tr>
<tr>
<td>GCFR [18]</td>
<td>16.97</td>
<td>0.70</td>
</tr>
<tr>
<td>DPR [52]</td>
<td>19.03</td>
<td>0.72</td>
</tr>
<tr>
<td>NLT [51]</td>
<td>20.15</td>
<td>0.75</td>
</tr>
<tr>
<td>Total Relighting [30]</td>
<td>20.24</td>
<td>0.79</td>
</tr>
<tr>
<td>NextFace [12]</td>
<td>22.98</td>
<td>0.76</td>
</tr>
<tr>
<td>Ours</td>
<td>23.64</td>
<td>0.83</td>
</tr>
<tr>
<td>Ours w/o coarse</td>
<td>17.83</td>
<td>0.66</td>
</tr>
<tr>
<td>Ours w/o SV $k_s$, $s$</td>
<td>21.31</td>
<td>0.77</td>
</tr>
<tr>
<td>Ours w/o $L_{mask}$, $L_{ink}$</td>
<td>16.46</td>
<td>0.61</td>
</tr>
<tr>
<td>Ours w/o $L_{mask}$</td>
<td>20.13</td>
<td>0.75</td>
</tr>
<tr>
<td>Ours w/o opt. $(\beta_i, \theta_i, \phi_i)$</td>
<td>18.67</td>
<td>0.69</td>
</tr>
<tr>
<td>Ours w/o soft shadow</td>
<td>21.46</td>
<td>0.77</td>
</tr>
<tr>
<td>Ours w/o $\Delta X$</td>
<td>21.16</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 1. Quantitative comparison. Comparison of our model on the tasks of novel view synthesis and relighting. See Section 6 for a description of the ablated variants.

Figure 6. Perspective. DECA’s assumption of an orthographic camera is broken by the strong perspective effects in selfies, causing poor alignment (b) with input images (a). Our first stage of optimization (c) (Sec. 5.1) improves alignment by solving for the parameters of a perspective camera and refined shape parameters. In the second stage we additionally optimize for a displacement map $\Delta X$ to produce our final shape with finer geometric details like wrinkles (d). Red line added to highlight alignment with (a).
7. Applications

**Relighting.** We demonstrate two types of relighting applications: (1) lighting modification and (2) lighting replacement. Practical lighting modification is common in portrait photography when the lighting conditions are not ideal, e.g., when direct sunlight casts undesirable harsh shadows with high contrast. A common practice is to make the light source larger and more diffuse by using a scrim or bounce card. In Fig. 7b, we show that by virtually increasing the size and spreading the energy of our reconstructed lighting source (i.e., the sun), we are able to soften the shadows and re-render a more visually pleasing face. Another approach to reducing the effects of harsh shadows is adding local fill lights, which reduces the contrast between the lit and shaded regions (Fig. 7c). Alternatively, fill lights can also be used for artistic purposes, to create dramatic lighting effects (Fig. 1b). Finally, replacing the scene lighting with that of a novel environment (Fig. 7d) is a necessary step in realistically inserting a captured subject into a virtual scene, which is useful for visual effects and VR applications.

**View Synthesis.** In Figure 8, we show that our reconstructed 3D model of the face can be used to synthesize new views by manipulating the viewpoint of the camera. We can also change other camera parameters, such as the focal length, to reduce the perspective effects on the face, which is often desirable for selfie images that contain significant facial distortion due to perspective.

**Skin Reflectance Editing.** We are also able to edit the reflectance components of the subject. As shown in Figure 1e, we can adjust the optimized albedo to add freckles, stickers, or other textures that realistically interact with reflections, shadows, and other elements of scene lighting, or we can adjust the specular properties of the face, making the face more or less shinier, as shown in Figure 9.

8. Conclusion and Discussions

In this paper, we propose SunStage, a lightweight and practical facial capture, rendering, and editing system that can serve as a minimalist light stage. With a video of an individual rotating in-place under the sun, our system reconstructs a physical model of the subject and the scene lighting, which enables us to relight the subject with realistic reflections and cast shadows. Our system allows arbitrary lighting and reflectance control in the reconstructed physical space, which can be rendered to produce photo-realistic results. We demonstrate several applications such as editing skin reflectance, relighting, and view synthesis.

**Limitations.** Our system inherits the limitations of morphable face models and is unable to model hair, teeth, or clothing geometry, beyond slight deformations. Additionally, certain regions which are seen under constant shadow or specular reflection (and therefore have no cues on reflectance or albedo) are sometimes unable to be decomposed accurately into separate reflectance and lighting components. Visualization and further discussions of the system’s limitations are provided in the supplemental material.
References


[26] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with tempo-


