







Improved Lighting Models for Facial Appearance Capture

Yingyan Xu^{1,2}  Jérémy Riviere²  Gaspard Zoss²  Prashanth Chandran^{1,2}  Derek Bradley²  Paulo Gotardo² 

¹ETH Zurich

²DisneyResearch|Studios

Abstract

Facial appearance capture techniques estimate geometry and reflectance properties of facial skin by performing a computationally intensive inverse rendering optimization in which one or more images are re-rendered a large number of times and compared to real images coming from multiple cameras. Due to the high computational burden, these techniques often make several simplifying assumptions to tame complexity and make the problem more tractable. For example, it is common to assume that the scene consists of only distant light sources, and ignore indirect bounces of light (on the surface and within the surface). Also, methods based on polarized lighting often simplify the light interaction with the surface and assume perfect separation of diffuse and specular reflectance. In this paper, we move in the opposite direction and demonstrate the impact on facial appearance capture quality when departing from these idealized conditions towards models that seek to more accurately represent the lighting, while at the same time minimally increasing computational burden. We compare the results obtained with a state-of-the-art appearance capture method [RGB*20], with and without our proposed improvements to the lighting model.

CCS Concepts

• **Computing methodologies** → **Reflectance modeling; Reconstruction; Appearance and texture representations; 3D imaging;**

1. Introduction

An essential aspect of rendering realistic human faces is to accurately model the appearance of the skin, which has a complex multi-layer structure with combined surface and subsurface reflectance properties. Appearance capture techniques estimate geometry and reflectance by performing a computationally intensive inverse rendering optimization (*i.e.*, analysis by synthesis) in which one or more images are re-rendered a large number of times and compared to real images from multiple cameras [RGB*20]. While state-of-the-art facial appearance capture techniques have achieved impressive results, they often make simplifying assumptions that do not hold in reality, but help tame high computational complexity. Lighting is often modeled as a collection of point sources (environment map) that are very far from the face, making its distribution spatially invariant within a small capture volume and, thus, easier to calibrate by taking directional samples from a single 3D location. However, face capture studios often have near light sources and, thus, incident lighting angles from a single source indeed vary across the face. In addition, appearance capture techniques often only account for lighting coming directly from the sources, neglecting the indirect light bouncing off of other parts of the face. Similarly, the effect of multiple subsurface bounces of light is also often ignored. As a result, the free parameters in the appearance model tend to overshoot to account for the missing indirect lighting, leading to inaccurate estimates. Polarization is also often used for diffuse-specular separation in face capture systems. This separation is often assumed to be perfect, such that one can directly fit a diffuse albedo to the cross-polarized imagery and estimate specular reflectance once the diffuse component is known. However, under

linearly polarized illumination, this separation is view-dependent because the optimal polarizer orientation for each light source is different for each camera view, and thus the perfect separation assumption does not hold in general multi-view setups. In practice, we observe degradation in diffuse-specular separation for cameras that view the face from below, as discussed in [GFT*11].

This paper demonstrates the impact on inverse rendering quality when departing from these idealized conditions towards more realistic lighting models. Our departure point is the state-of-the-art face capture system of [RGB*20], and we improve the lighting model in several ways. First, we present a practical method to calibrate a near field of light sources, going beyond the estimation of distant light directions to actually compute their 3D positions in the scene (Section 4). This method relies on the usual protocol of capturing a mirror sphere from multiple views, without requiring additional images. Second, following [GCP*10], we more accurately account for polarized light transport using Stokes vectors and Mueller matrices (Section 5), rather than simply assuming the specular component is zero for cross-polarized views. And third, to account for indirect bounces of light on the skin surface, we propose an efficient texture-space technique that is well suited for inverse rendering optimization and can be combined with efficient texture-space subsurface reflectance models [RGB*20] (Section 6). We then show how our enhanced lighting model leads to more accurate appearance optimization and higher visual fidelity, generalizing better to different lighting conditions (Section 7). Although the importance of some of these individual rendering components have been highlighted before, individually, another main contribution of our work is to demonstrate their combined effect on today's state of the art.

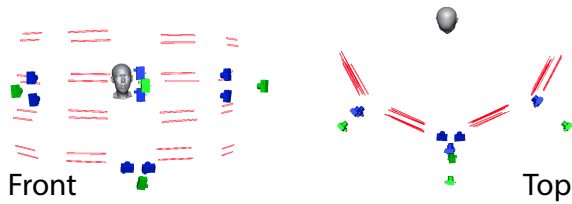


Figure 1: Our setup of 4 cross-polarized (green) and 8 unpolarized (blue) cameras, and positional light reconstruction result (red).

2. Related Work

The seminal work by Debevec *et al.* [DHT*00] employed a specialized light stage to acquire a dense reflectance field of a human face and a few view-dependent reflectance maps. Since then, many following techniques employed LED spheres for reflectance capture [WMP*06, GFT*11], usually requiring multiple active lighting patterns and relying on polarization filters for diffuse-specular reflectance separation. More recently, passive polarized light methods have been proposed, allowing to estimate facial appearance parameters in dynamic videos [GRB*18] and single-shot face capture systems [RGB*20]. All of these methods employ one or more of the lighting model simplifications that we discuss in this paper, either assuming distant lighting, perfect separation of diffuse and specular reflection through polarization, or ignoring indirect bounces of light. In this work, we demonstrate the impact of these lighting-related assumptions on the recent method of [RGB*20].

3. Capture Setup and Baseline Method

As a baseline, we use the light-weight face capture system of Riviere *et al.* [RGB*20], which employs polarized, passive illumination to provide high-quality single-shot geometry and appearance. The capture setup consists of 32 LED bars (as 16 pairs) that are placed approximately 1 meter away from the face. Each bar contains a horizontal, linear polarization filter. Twelve video cameras are arranged into 4 triplets (Fig. 1), each consisting of (i) a narrow baseline stereo pair that captures full facial reflectance (also used for stereo reconstruction), and (ii) a central camera that is cross-polarized with respect to the lights and mainly captures diffuse reflection. The illumination is modeled as a collection of distant directional sources in a usual latitude-longitude environment map, measured using the traditional approach of capturing multiple exposures of a mirror sphere from one camera view. In an effort to improve the speed of appearance capture, a sparse representation of the lighting is computed with a resolution of only 450 lighting directions uniformly distributed over the frontal hemisphere. Forward rendering follows a traditional ray-tracing approach, and to optimize appearance parameters, the baseline algorithm uses an auto-differentiable package (*ceres-solver.org*) to implement a custom renderer that operates in UV texture-space, and reconstructs a diffuse albedo, specular albedo (intensity), and high resolution normal map [RGB*20]. Below, we introduce improvements to this system by reconstructing the near field light sources explicitly, modeling the light polarization, and accounting for indirect illumination.

4. Near Field Light Reconstruction

Rather than a distant environment map, we propose to reconstruct the 3D geometry and intensities of near field lights explicitly. Ban-

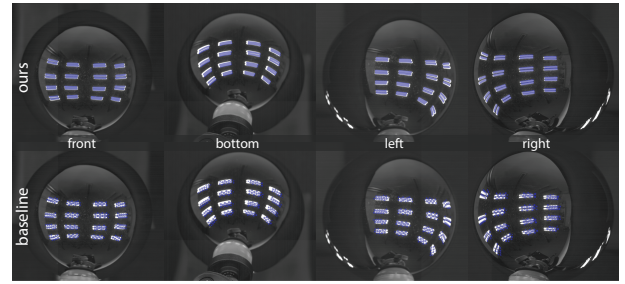


Figure 2: Real mirror sphere images with LED bar reflections: (top) overlaid, purple reprojections of our positional lights; and (bottom) baseline environment map. Notice the better alignment of the positional light representation with the real-world lights.

terle *et al.* [BCD*13] show that even rough 3D depth of light sources adds plausible near lighting effects that are otherwise missing in an environment map. Here, we push this idea further using multi-view images and triangulation to precisely measure the 3D locations of light sources for high-fidelity face capture. For our setup, we represent the geometry of each LED bar as a rectangular 3D surface with known width and height (from manual measurements), whose position and orientation in 3D space must be recovered from multi-view images of the mirror sphere. In addition, we model a fixed number of point lights that are uniformly placed inside each light rectangle, and we calibrate the intensities of these sources using the captured HDR image, as described below. Our method is naturally applicable to other area light shapes.

To locate each rectangle in 3D space, we manually annotate its four corner vertices on the multi-view images of the mirror sphere, Fig. 2. For each vertex, we shoot rays from each camera through the annotated pixels and reflect them on the sphere. The desired 3D vertex position is the one closest to all reflected rays, in a least-squares sense. We thus optimize for the 3D rectangle positions that minimize the four corner ray intersection distances, while also weakly enforcing a regularization constraint that aims to orient the rectangle's normal towards the center of the sphere, accounting for inaccuracies in annotations. The 3D sphere position is initialized from multi-view annotations and thus may be inaccurate, so we also jointly optimize for the sphere position during light reconstruction. The result is a set of 3D oriented rectangles representing the geometry of the LED bars, as shown in Fig. 1. As a second step, we sample L point light sources inside each rectangle and calibrate their intensities ($L = 10$ in our experiments). This is done using the HDR image of the mirror sphere, captured by a frontal unpolarized camera. We loop over the pixels in this image, again tracing rays and assigning the observed pixel intensity to the point source nearest to the reflected ray. Each point source accumulates the contribution of several pixels. During accumulation, the intensity of each pixel is adjusted to factor out the attenuation of light along the ray from the light source to the mirror sphere, using the inverse-square law. This allows for more accurate, spatially-varying light source intensities during rendering, which is not possible with a traditional environment map. Fig. 2 shows captured multi-view images of the mirror sphere with overlaid renderings of the light model using our reconstructed field of near positional lights, compared to a traditional distant environment map. As our positional

lights are computed using all viewpoints, they more faithfully represent the true lights in the scene. In practice, we found that the few pixels from the HDR image of the mirror sphere do not provide accurate *intensities* for the point sources. Alternatively, we obtained better solutions via inverse rendering, by imaging a diffuse object (e.g., using cross-polarized views of a face itself, without manual intervention) while turning on each light independently and optimizing for the intensities of the point lights to best fit the images.

5. Modeling Polarization

We simulate polarized light transport [GCP*10], where the incident light is horizontally (linearly) polarized and modeled as the Stokes vector $\mathbf{s} = (1, 1, 0, 0)$. At each point on the face, given the incident and outgoing directions, we compute the Mueller matrix \mathbf{M}_1 and the transformed $\mathbf{s}' = \mathbf{M}_1\mathbf{s}$, after specular reflection off a dielectric material. From the resulting $\mathbf{s}' = (s_0, s_1, s_2, s_3)$, we take s_0 as the Fresnel gain for the specular component and modulate the diffuse component with $1 - s_0$. For cross-polarized views, a second Mueller matrix \mathbf{M}_2 models the camera's vertical linear polarizer, and we instead compute $\mathbf{s}' = \mathbf{M}_2\mathbf{M}_1\mathbf{s}$. Thus, we do model the (weak) specular component in cross-polarized views. In contrast, the baseline model of [RGB*20] assumes purely diffuse reflectance for these views and, for cameras without a polarizer, simply uses the Fresnel curve for p-polarized light with the usual equations for unpolarized light transport. The added complexity in our model is very small, and runtime increases only slightly by the evaluation of surface reflection for cross-polarized views, but Section 7 shows that we obtain improvement in both the appearance maps and re-rendering error when modeling the polarization. Note that we do not yet model depolarization in indirect lighting (described next), which is left for future work.

6. Texture Space Indirect Illumination

Our baseline [RGB*20] is one of the few methods to efficiently model subsurface scattering, using a texture space technique with precomputed visibility and shadow maps. However, indirect bounces of light *on the surface* remain ignored (as with most facial appearance capture methods). We propose an efficient texture-space indirect illumination method for inverse rendering based on Monte Carlo sampling. This new method samples UV positions of neighboring vertices that contribute to indirect lighting (e.g., additional point lights) and fixes the directions of these rays during optimization. To optimize the appearance parameters at texel u_0 under indirect lighting, our algorithm works as follows:

1. From the 3D position of u_0 , shoot rays along random directions over the visible hemisphere, intersect the geometry, and store the UV coordinates u_i of these intersections (neighboring texels).
2. For each intersection, use the current appearance parameters at the neighboring texel u_i to render it from the viewpoint of u_0 and take the result as the intensity of indirect lighting arriving from u_i to constrain inverse rendering at u_0 .
3. Perform one optimization step to update the appearance parameters at u_0 , rendering it with both direct and indirect lighting.
4. Iterate over texels and repeat steps 2-3 until convergence.

This technique is easy to implement and can be readily incorporated into different inverse-rendering pipelines. When implemented in the baseline method [RGB*20], we propose to account

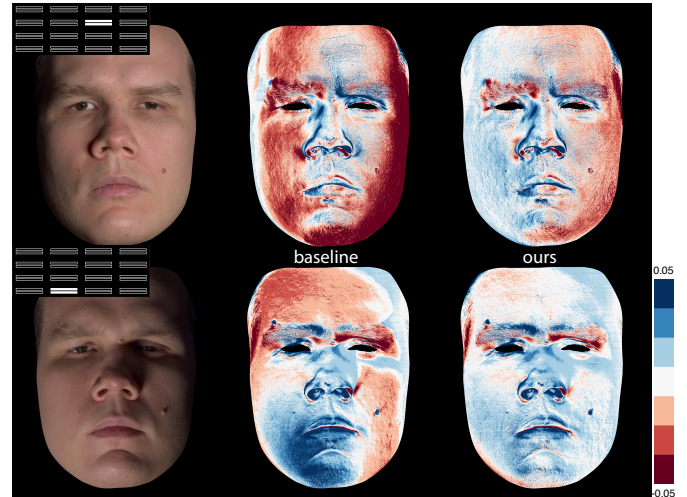


Figure 3: Captured photos under lighting from a single pair of LED bars (left), rendering error with the baseline sparse environment map (mid) and rendering error with our positional lights (right).

for indirect illumination only in the final computation of diffuse and specular albedo, after surface normal optimization, thus minimally increasing computation cost by 15% to 20%. Since we compute shading with samples that remain fixed during optimization, our estimates of the rendering integral are neither unbiased nor consistent. However, in Section 7 we will demonstrate that the subsurface scattering implementation in [RGB*20] is able to attenuate noise artifacts caused by our model of indirect illumination.

7. Results

We now show results of our improved lighting models and compare against the simplifying assumptions used by many other methods. Renderings and errors are computed using appearance maps optimized with the corresponding lighting model, as indicated.

Near Field Positional Lights. Fig. 1 shows the reconstructed LED bars from our positional model (in red). As illustrated in Fig. 2, reprojecting the lights back onto the mirror sphere images shows that our positional light model align much better with the LED bars in the photo, especially on the side views. As a result, we obtain faster and more accurate (inverse) rendering of captured faces. In Fig. 3, negative/positive error means the render is too bright/dark. The improvement is more evident when rendering the face under harsh lighting that produces stronger shadow boundaries (Fig. 4).

Modeling Polarization. By explicitly modeling polarized light transport, our method goes beyond simple separation of diffuse and specular reflectance, to better capture the parameters that determine surface reflection. The improvement is particularly more evident on the quality of recovered specular intensity (specular albedo) map, which shows fewer artifacts, and when rendering (lower) views off of the equator of the sphere containing our light sources, Fig. 5.

Texture Space Indirect Illumination. With indirect lighting, soft shadows under harsh lighting conditions are better modeled during inverse rendering, improving the fidelity and visual quality over the baseline method (Fig. 6). As mentioned earlier, the diffuse albedo estimated by the baseline tends to overshoot near face concavities

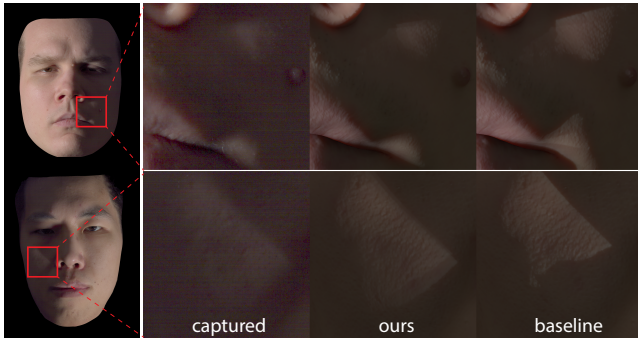


Figure 4: Shadow lines in a real photo (left) are better reproduced by our positional lights (mid), than by the environment map (right).

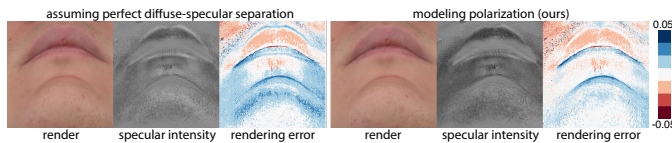


Figure 5: We improve on the baseline result (left) by extending its model to explicitly account for light polarization (right), leading to more homogeneous specular intensity and lower rendering error.

(around the eyes and sides of nose) to compensate for the missing indirect lighting. Our new method largely alleviates this issue, as shown in Fig. 7 (left). The modeled indirect illumination also adds new constraints on specular intensity and leads to better estimates (Fig. 7 (right)). Still, some dark regions remain in the specular intensity map, suggesting the need for capturing additional viewpoints (e.g., from above) or adding more complete spherical illumination. Finally, modeling subsurface scattering was shown in [RGB*20] to provide sharper albedo maps. Here, we additionally show that the subsurface scattering model also helps to attenuate noise introduced by our indirect lighting method (Fig. 8).

8. Conclusion and Discussion

In this work we demonstrated the benefits of deviating from traditional idealized assumptions related to the lighting model and

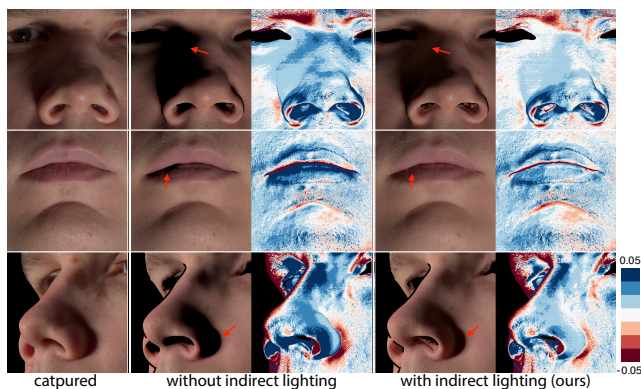


Figure 6: Modeling indirect lighting during inverse rendering improves reconstruction and rendering fidelity relative to a reference image, especially in partly shadowed, concave areas (red arrows).

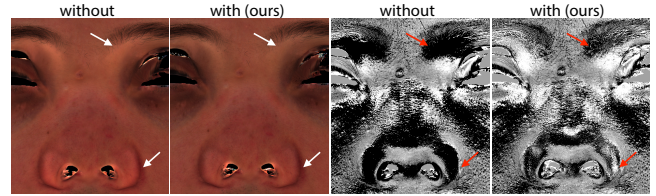


Figure 7: Diffuse albedo and specular intensity maps optimized without and with indirect illumination: over- and under-estimation errors (arrows) are reduced around the face concavities. These maps are shown in linear RGB colors for better visualization.



Figure 8: Modeling subsurface scattering (2nd, 4th columns) helps to reduce noise introduced by indirect illumination samples.

interactions in the context of facial appearance capture. Comparing to a current state-of-the-art approach, improved quality can be achieved by explicitly reconstructing near field lights, modeling polarization, and handling indirect light bounces. Furthermore, we have proposed practical implementations for each of these benefits. With regards to drawbacks and limitations, we note that our positional light reconstruction method requires human annotation and the improvement is not obvious under smooth low-frequency full-on lighting conditions. Accounting for indirect illumination increases the computational cost and memory allocation by a small percentage; however, our implementation is still realizable on modern PCs and we believe the cost can be largely reduced by computing indirect illumination only in face concavities. To summarize, we believe the insights in this paper will be beneficial for both practical facial capture systems as well as future academic research.

References

[BCD*13] BANTERLE F., CALLIERI M., DELLEPIANE M., CORSINI M., PELLACINI F., SCOPIGNO R.: EnvyDepth: An interface for recovering local natural illumination from environment maps. *Computer Graphics Forum* 32, 7 (2013). Proc. Pacific Graphics 2013. 2

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUiker H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *ACM ToG (SIGGRAPH)* (2000). 2

[GCP*10] GHOSH A., CHEN T., PEERS P., WILSON C. A., DEBEVEC P.: Circularly polarized spherical illumination reflectometry. In *ACM ToG (SIGGRAPH Asia)*. 2010. 1, 3

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. In *ACM ToG (SIGGRAPH Asia)* (2011). 1, 2

[GRB*18] GOTARDO P., RIVIERE J., BRADLEY D., GHOSH A., BEELER T.: Practical dynamic facial appearance modeling and acquisition. *ACM ToG (SIGGRAPH Asia)* 37, 6 (2018). 2

[RGB*20] RIVIERE J., GOTARDO P., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *ACM ToG (SIGGRAPH)* 39, 4 (2020). 1, 2, 3, 4

[WMP*06] WEYRICH T., MATUSIK W., PFISTER H., BICKEL B., DONNER C., TU C., MCANDLESS J., LEE J., NGAN A., JENSEN H. W., ET AL.: Analysis of human faces using a measurement-based skin reflectance model. *ACM ToG (SIGGRAPH)* 25, 3 (2006). 2