High-Res Facial Appearance Capture from Polarized Smartphone Images

Dejan Azinović¹  Olivier Maury²  Christophe Hery²  Matthias Nießner¹  Justus Thies³
¹Technical University of Munich  ²Meta Reality Labs  ³Max Planck Institute for Intelligent Systems

Abstract

We propose a novel method for high-quality facial texture reconstruction from RGB images using a novel capturing routine based on a single smartphone which we equip with an inexpensive polarization foil. Specifically, we turn the flashlight into a polarized light source and add a polarization filter on top of the camera. Leveraging this setup, we capture the face of a subject with cross-polarized and parallel-polarized light. For each subject, we record two short sequences in a dark environment under flash illumination with different light polarization using the modified smartphone. Based on these observations, we reconstruct an explicit surface mesh of the face using structure from motion. We then exploit the camera and light co-location within a differentiable renderer to optimize the facial textures using an analysis-by-synthesis approach. Our method optimizes for high-resolution normal textures, diffuse albedo, and specular albedo using a coarse-to-fine optimization scheme. We show that the optimized textures can be used in a standard rendering pipeline to synthesize high-quality photo-realistic 3D digital humans in novel environments.

1. Introduction

In recent years, we have seen tremendous advances in the development of virtual and mixed reality devices. At the same time, the commercial availability of such hardware has led to a massive interest in the creation of ‘digital human’ assets and photo-realistic renderings of human faces. In particular, the democratization to commodity hardware would open up significant potential for asset creation in video games, other home entertainment applications, or immersive teleconferencing systems. However, rendering a
human face realistically in a virtual environment from arbitrary viewpoints with changing lighting conditions is an extremely difficult problem. It involves an accurate reconstruction of the face geometry and skin textures, such as the diffuse albedo, specular gain, or skin roughness. Traditionally, this problem has been approached by recording data in expensive and carefully calibrated light stage capture setups, under expert supervision. We seek to simplify this capture process to allow individuals to reconstruct their own faces, while keeping the quality degradation compared to a light stage to a minimum.

The disentanglement of geometry and material of human faces is an extremely ill-posed problem. Current solutions involve a capture setup with multiple cameras and light sources, with millimeter-accurate calibration. A common approach to disentangling face skin surface from subsurface response is the use of polarization filters [9] in tandem with such expensive capture setups. Given such a carefully calibrated capture setting, one can use differentiable rendering to estimate the individual skin parameters in an analysis-by-synthesis approach. While these methods do produce visually impressive results, they are limited to high-budget production studios.

In this paper, we propose a capture setup consisting of only a smartphone and inexpensive polarization foils, which can be attached to the camera lens and flashlight. Inspired by light stage capture setups, a user captures two sequences of their face, one with perpendicular filter alignment, and one with parallel alignment. This allows for a two-stage optimization, where we first reconstruct a high-resolution diffuse albedo texture of a user’s face from the cross-polarized capture, followed by recovery of the specular albedo, normal map, and roughness from the parallel-polarized views. Data is captured in a dark room to avoid requiring pre-computation of an environment map. In addition to visually compelling novel view synthesis and relighting results, our method produces editable textures and face geometry.

In summary, the key contributions of our project are:

- We propose a commodity capture setup that combines a smartphone’s camera and flashlight with polarization foils. The polarization allows us to separate diffuse from specular parts, and to reconstruct the user’s face textures, such as diffuse albedo, specular albedo and normal maps.
- Our proposed capture setting with the co-located camera and light enables separation of skin properties from illumination, which is of key importance for realistic rendering of faces.
- We propose a coarse-to-fine optimization strategy with mip-mapping, which increases sharpness of the reconstructed appearance textures.

2. Related Work

High-fidelity face appearance capture and reconstruction has received significant attention in the entertainment industry for creating digital humans and more recently in the AR/VR community for generating realistic avatars. In our context, facial appearance reconstruction means recovering a set of high-resolution albedo, specular (gain and roughness) and normal maps. Over the years, physically-based skin scattering models have become ever more sophisticated [6, 26, 53]; however, their input texture quality remains the single most important factor to photo-realism.

**Polarization.** For some time, polarization has been used to separate specular from diffuse [35,39,51]. These techniques rely on the fact that single bounce specular reflection does not alter the polarization state of incoming light. Riviere et al. [41] propose an approach to reconstruct reflectance in uncontrolled lighting, using the inherent polarization of natural illumination. Nogue et al. [38] recover SVBRDF maps of planar objects with near-field display illumination, exploiting Brewster angle properties. Deschaintre et al. [10] use polarization to estimate the shape and SVBRDF of an object with normal, diffuse, specular, roughness and depth maps from a single view. Dave et al. [8] propose a similar approach for multi-view data. In MoRF [48], a studio setup with polarization is used to reconstruct relightable neural radiance fields of a face.

**Lightstage capture systems.** In their foundational work,Debevec et al. [9] introduced the Lightstage system to capture human face reflectance using a dome equipped with controlled lights, separating the diffuse from the specular component using polarization filters. Follow-up work reconstructs high-resolution normal maps using photometric stereo [52], compensates for motion during the capture [50] and expands the captured area [19].

The proposed capture studios didn’t come without limitations, as the lighting environment needed to be tightly controlled, the lighting patterns involved took a relatively long time, and the polarization filters were challenging to set up for multiple cameras and lights. Fyffe et al. [14–17] proposed the use of color gradients and spectral multiplexing to reduce capture time. With the objective of designing a more practical system, Kampouris et al. [24] demonstrate that binary gradients are sufficient for separating diffuse from specular without polarization. Lattas et al. [28] use an array of monitors or tablets for a practical binary gradients capture studio. In line with this thread of research, Gotardo et al. [20] present a multi-view setup for dynamic facial texture acquisition without the need for polarized illumination. Riviere et al. [40] build a similar lightweight system reintroducing polarization without active illumination, and modeling subsurface scattering. This effort was refined to include global illumination and polarization modeling [54].
proposed solutions deliver impressive visual results, but require expensive and difficult to use hardware. We propose a solution for high-resolution facial texture reconstruction using commodity devices, such as smartphones.

**Differentiable rendering.** Recent progress in differentiable rendering \[4, 47, 56, 57\] has led to the development of mature frameworks \[23, 27, 37\] and a number of methods that try to jointly estimate appearance and lighting \[36\]. For an overview of differentiable rendering techniques, see \[25\]. Luan et al. \[33\] use a co-located camera and light setup to reconstruct shape and material, relying on a differentiable renderer \[57\] to produce unbiased gradients for shape estimation on an explicit mesh. With the same co-located setup, Zhang et al. \[58\] improve results by using a hybrid volume radiance field and neural SDFs for the shape estimation. While using a similar capture configuration to our work, the previous techniques focus on shape reconstruction, while we can lean on an accurate prior for the basis of our face shape. Furthermore, by using polarization, we can properly decouple diffuse from specular textures.

Dib et al. \[11–13\] propose the estimation of face skin textures by modelling the illumination with a virtual light stage, and using a differentiable ray tracer \[32\]. The method fits a parametric face mask to the observed images and is able to handle self-shadowing, but complex lighting environments can have an impact on separation of lighting and material. Wang et al. \[49\] propose a capture setup with the sun as the main light source. A FLAME \[31\] model is fit to the observed data, after which geometry and material are jointly refined using an analysis-by-synthesis approach. As with other methods in uncontrolled lighting, separation of individual textures remains a challenge.

**Deep learning-based approaches.** A wide range of work proposes learning a neural network from large collections of high-quality light stage data, and subsequently applying the model to new data \[5, 22, 29, 30, 34, 42, 55\]. Zhang et al. \[59\] propose learning a neural light transport model from uv-space light and view direction information. At test time, the model generalizes to novel views and lighting. Several other works propose learning neural rendering models, either from single-view \[18, 21, 44, 60\] or multi-view \[45, 46\] data, for a range of different applications.

### 3. Method

We propose a two-step analysis-by-synthesis approach for the estimation of high-resolution face textures, as depicted in Figure 2. The user captures two video sequences and a series of photographs of their face under linear-polarized point light illumination using a smartphone. The first sequence has the polarization filters oriented in a perpendicular fashion, i.e., the filter covering the camera lens is perpendicular to the filter covering the smartphone’s flashlight. In accordance with existing literature, we denote this sequence as the cross-polarized sequence. The second video sequence has parallel oriented filters and will be referred to as the parallel-polarized sequence.

We use structure-from-motion and multiview-stereo on all captured frames jointly to compute the camera alignment and reconstruct coarse geometry in form of a triangle mesh. We then non-rigidly fit the FLAME model \[31\] to the scan and use it as our base geometry model. This fitting helps us avoid noise from the multiview-stereo and provides a consistent UV-parameterization for all subjects. Based on this geometry, we recover the diffuse albedo texture of the
subject using the cross-polarized data and photo-metric optimization. While keeping the diffuse albedo fixed, we estimate the remaining textures based on the parallel-polarized data. Note that we reconstruct textures using only the photographs, as these capture more detail than the video frames. For the geometry reconstruction, we use all captured data, as we found that this leads to more robust results compared to only using a small set of photographs.

3.1. Capturing Polarized Data with a Smartphone

We capture one cross-polarized and one parallel-polarized video sequence with a smartphone in a dark room, with the smartphone’s flashlight as the only source of illumination. Such a capture setup has the advantage of not requiring optimization of the scene lighting, leading to better separation of appearance and shading. We assume that the flashlight is co-located with the camera lens and that its color is white. We capture a color-checker under both filter orientations to color-calibrate both sequences. This is important, since the filters introduce wavelength-dependent attenuation which tints the color of the light. We use an affine color calibration scheme to compute the corresponding color correction matrix only once, and apply it to all subsequent sequences. Furthermore, since an arbitrary smartphone’s flashlight does not behave like an ideal point light (e.g., due to occlusion by the phone’s cover along grazing directions), we pre-compute a per-pixel light attenuation map, that is multiplied with the final rendered images during optimization. To this end, we put markers on a flat white surface and record a cross-polarized sequence of the surface. We form an optimization problem with the unknowns being the surface’s diffuse texture and the per-pixel light attenuation map. The map is then kept fixed for all future face texture optimizations. We refer to the supplemental material for more detail on this calibration step.

We ensure that all captures have consistent and fixed camera settings: focal length, exposure time and white balance. We capture at 4K resolution and 30fps and select the sharpest frame from every 10-frames window, using variance of the Laplacian as the sharpness metric. In addition to the video data, we capture a set of cross-polarized and a set of parallel-polarized photographs to obtain higher-quality data. Since the flash is much brighter for photographs than for videos, we capture the photographs with shorter exposure and lower ISO to roughly match the brightness of the video frames. The entire capture takes about five minutes.

3.2. Geometry Reconstruction

We use Agisoft Metashape [1] on all frames jointly to obtain an initial mesh reconstruction. We provide Metashape with face masks estimated by [61], to make the reconstruction more robust to rigid motion of the head. We then fit the FLAME model [31] to the scanned geometry, first by optimizing the shape parameters of the FLAME face space, and then by an ICP-based as-rigid-as-possible deformation approach (see Figure 4). For the non-rigid deformation, we subdivide the triangles of the face region, to obtain detailed geometry. The resulting mesh is used as the base mesh for the subsequent texture optimizations.

3.3. Rendering Equation & BRDF

We model the skin with a spatially-varying bidirectional reflectance distribution function (SV-BRDF). Assuming a point light source \( l \) in a dark environment, the rendering equation that defines the outgoing radiance \( L_o(x, \omega) \), at point \( x \) with normal direction \( n^\top \) in direction \( \omega \), has the following simplified form:

\[
L_o(x, \omega) = \frac{f(x, \omega)(n^\top \omega)L_i(x, \omega)}{|x - l|^2}.
\]

Here, we make use of the fact that the light direction aligns with the view direction, i.e., \( \omega = \omega_o = \omega \). The BRDF \( f(x, \omega) \) has a diffuse component \( f_d \), and a specular component \( f_s \). We use the Cook-Torrance [7] BRDF for our specular term:

\[
f_s(x, \omega) = k_s(x)\frac{D(\omega, n^\top, \alpha)G(n, \omega)f(n, \omega)}{4(n^\top \omega)(n^\top n)}.
\]

with \( k_s \) being the spatially-varying specular gain and \( \alpha \) a global roughness blend factor for the Blinn-Phong distri-
bution term $D$ of the 2-lobe mix ($D_{12}$ and $D_{48}$) suggested by [40]. $G$ denotes the geometry term of the Cook-Torrance BRDF model. We use Shlick’s approximation [43] for the Fresnel term $F$:

$$F(n, \omega) = F_0 + (1 - F_0)(1 - n^T \omega)^5.$$ (3)

To model the skin’s diffuse response, we implement the BRDF model proposed by Ashikhmin and Shirley [2, 3], that accounts for the fact that a portion of the light has already scattered before penetrating the skin surface:

$$f_d(x, \omega) = \frac{28k_d(x)}{23} (1 - F_0)(1 - (1 - n^T \omega)^3)^2,$$ (4)

where $F_0 = 0.04$ is the reflectance of the skin at normal incidence. Indirect light bouncing from the capture environment and on the captured face itself might have a significant contribution to pixel intensity at grazing angles, so we also add a Fresnel-modulated ambient term to our BRDF $f$:

$$f_a(x, \omega) = k_a(x)(1 - F_0)(1 - (1 - n^T \omega)^5)^2,$$ (5)

with an ambient map $k_a$ which is regularized to be smooth via a total variation loss and close to zero.

Note that using a diffuse scattering model for the optimization is compatible with state-of-the-art physically-based subsurface scattering skin shading [6, 53], as shown in Figure 1. Production-ready subsurface scattering models typically include an albedo inversion stage, which takes a diffuse albedo as input, and converts it to extinction coefficients for the volume rendering random walk.

### 3.4. Optimization

The objective of the photometric optimization step is to minimize the difference between rendered images $\hat{I}$ and color-corrected target images $I$:

$$\mathcal{L}(\hat{I}, I) = \left| W \cdot (\hat{I} - I) \right|,$$ (6)

with $\hat{I} = M \cdot L_{\omega o}$, where $M$ is the pre-computed light attenuation map, that accounts for uneven light distribution in different directions. We apply a per-pixel loss weight $W$ based on the respective mip level and the angle between viewing direction and normal $n^T \omega$ to improve sharpness. Specifically, to ensure that distant or grazing angle observations do not blur the resulting textures, for each pixel that is projected from the target image to texture space, we calculate which mip level $l$ would need to be looked up in classical forward rendering. $W$ is set to $(n^T \omega)(1 - l)$ if the pixel corresponds to a mip level below 1, and zero otherwise.

We optimize $\mathcal{L}(\hat{I}, I)$ in two steps, using a coarse-to-fine optimization strategy in each. In the first step, we only use the cross-polarized images to optimize the spatially-varying diffuse albedo texture $k_d(x)$ and an initial tangent-space normal map $n(x)$, while assuming $f_s(\cdot) = 0$ for the specular term. In the second step, we fix the diffuse texture and optimize for specular gain $k_s(x)$, specular roughness $\alpha$, and the final normal map $n(x)$. To account for potentially different light attenuation in the cross and parallel-polarized filter settings, we also optimize per-channel scaling factors for the diffuse texture. The optimization is performed entirely in texture space. In each step, we employ a four-level coarse-to-fine optimization strategy, starting with a texture resolution of $512 \times 512$, and increasing the size by a factor of two after convergence of each level, up to the final resolution of $4096 \times 4096$.

We implement our optimization framework in PyTorch, using nvdiffrast [27] as our differentiable renderer. We optimize on batches of 4 images, using Adam with an initial learning rate $l_{r_0} = 10^{-3}$ for all parameters at the beginning of every coarse-to-fine step, and updating it to $l_{r} = l_{r_0} \cdot 10^{-0.0011}$ in every iteration $t$. We scale the FLAME mesh to unit size and set the light intensity to 10. The total optimization time is about 90 minutes.

### 4. Results

In this section, we present texture reconstruction and rendering results on several subjects. Figure 5 shows the texture reconstruction on several actors of different ethnicity. Our method is able to reconstruct pore-level detail in the diffuse, specular and normal maps. Further, we evaluate the quality of our reconstructed textures by rendering the mesh from novel views and under novel illumination. Figure 6 shows that our method faithfully reconstructs the skin’s appearance under novel views and lighting.

Comparison to state of the art. We perform both a qualitative and quantitative evaluation of our method and compare to state-of-the-art methods for relighting and texture reconstruction. During optimization, we hold out a validation frame on which we compute image metrics.

**Neural Light Transport.** Neural Light Transport [59] is a deep learning-based method that takes as input pre-computed diffuse base, light-cosine and view-cosine uv-space maps. The diffuse base is computed as the average of all observations. The cosine maps contain per-texel cosines of the angles between the normal vector and the light or view vector. Based on these inputs, as well as nearest neighbor observations, a neural network learns to predict the final

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>LPIPS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLT [59]</td>
<td>31.51</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>NextFace [11]</td>
<td>22.85</td>
<td>0.89</td>
<td>0.31</td>
</tr>
<tr>
<td>Ours</td>
<td>32.37</td>
<td>0.96</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Table 1. We compare our method to NLT and NextFace on validation frames over 10 different subjects.
Figure 5. We show skin texture reconstructions of several actors of different skin type. The rendered images closely match the reference target images, and we achieve good separation of diffuse and specular textures.
Figure 6. We evaluate on a validation frame from a novel viewpoint and with novel lighting that was held out during the optimization. As visible in the crop regions, our method is able to synthesize sharper texture details and specular highlights compared to NLT [59].

As shown in Table 1, our approach achieves favorable image metrics. Figure 6 compares our method to NLT on novel lighting and viewpoint. NLT closely matches the target by using nearby camera views, but specular highlights are often blurry, and the low number of training views results in the model producing artifacts in shadowed areas. We obtained the best NextFace results in an experiment with uniform illumination using three frames that cover the whole face region. As shown in Figure 7, inaccuracies in the face model fitting lead to somewhat blurry textures. This issue is exacerbated by adding more frames. Using fewer frames degraded the separation of the diffuse and specular textures. Our method is able to overcome these issues by accurately fitting a geometric model to the input data and by using polarization to separate the individual textures.

Ablation Studies. We conduct ablation studies to justify our choice of capture setup and training parameters. In Figure 8, we show that accounting for the direction-dependent light attenuation of a smartphone’s flashlight leads to an overall lower error in the re-rendered images. In the same figure, we also show the importance of accounting for the Fresnel effect when reconstructing the diffuse texture. A purely Lambertian BRDF will not be able to model the skin’s diffuse response at all angles. In Figure 9, we show that optimizing textures without cross-polarization will leak specular information into the diffuse texture.

Coarse-to-fine optimization and mipmapping. Pixels of the target images have different footprints in uv-space, depending on distance and angle between camera and surface.
Figure 8. Ignoring the angle-dependent flashlight attenuation, the Fresnel effect, or the ambient light leads to an incorrect reconstruction, that can no longer reproduce the shading from all views. We account for these effects to closely match the target data.

Figure 9. We compare joint optimization of all textures to our full approach on a purely diffuse render. Optimizing jointly leaks specular and normal map information into the diffuse texture. Weighting the loss of each pixel equally leads to blur in the reconstruction. Optimizing coarse-to-fine, where at each resolution we use only pixels with the corresponding uv-space footprint, helps us reconstruct additional detail in the textures. Figure 10 shows a comparison between our full approach and a direct optimization of the highest resolution texture. We additionally show the decrease in quality when optimizing only on video frames (w/o photographs).

Runtime and memory consumption. Including 1 hour spent on MVS, our method needs about 2.5h to reconstruct a face. Photo-metric skin texture reconstruction takes about 90 minutes on an Nvidia RTX A6000. We reconstruct facial geometry with Metashape using an average of 420 video frames and 70 photographs. At a texture resolution of $4096 \times 4096$ and target image resolution of $3840 \times 2160$, the photo-metric optimization requires 30GB of GPU memory. In comparison, NLT takes about 10h and NextFace about 6h given the same number of frames.

Discussion & Limitations. Our method reconstructs high-quality face textures with a low-cost capture routine. However, it is restricted to static expressions, i.e., it does not handle dynamically changing face geometry and textures. An avenue for future research is the reconstruction of dynamic expressions by fitting a parametric model with consistent mesh topology to each frame, and optimizing over the entire non-rigid sequence. Our method does not explicitly handle global illumination. A differentiable path tracer could potentially improve results in the concavities of the eye region. As we assume a static face with closed mouth and closed eyes, we only recover the skin area of a face. Eyes, mouth interior and hair are a subject of future work.

5. Conclusion

We have presented a practical and inexpensive method of capturing high-resolution textures of a person’s face by coupling commodity smartphones and polarization foils. The co-location of the camera lens and light source allows us to reduce the problem complexity and separate material from shading information. As a result, we obtain high-resolution textures of the skin area of the human face. We believe that polarization is a powerful tool for material recovery in the real world, and future smartphones could benefit from including filters directly in the hardware. Overall, we believe that our work is a stepping stone towards democratizing the creation of digital human face assets by making it more accessible to smaller production studios or individual users.

Acknowledgements

Dejan Azinović’s contribution was supported by the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Grant “Making Machine Learning on Static and Dynamic 3D Data Practical”, and the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing”. We would also like to thank Angela Dai for the video voice over and Simon Giebenhain for help with the FLAME fitting.
References

[29] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet


[54] Yingyan Xu, Jérémy Riviere, Gaspard Zoss, Prashanth Chandran, Derek Bradley, and Paulo Gotardo. Improved lighting models for facial appearance capture. 2022. 2


