High-Res Facial Appearance Capture from Polarized Smartphone Images

Dejan Azinović[1] Olivier Maury[2] Christophe Hery[2] Matthias Nießner[1] Justus Thies[3]

[1]Technical University of Munich [2]Meta Reality Labs [3]Max Planck Institute for Intelligent Systems

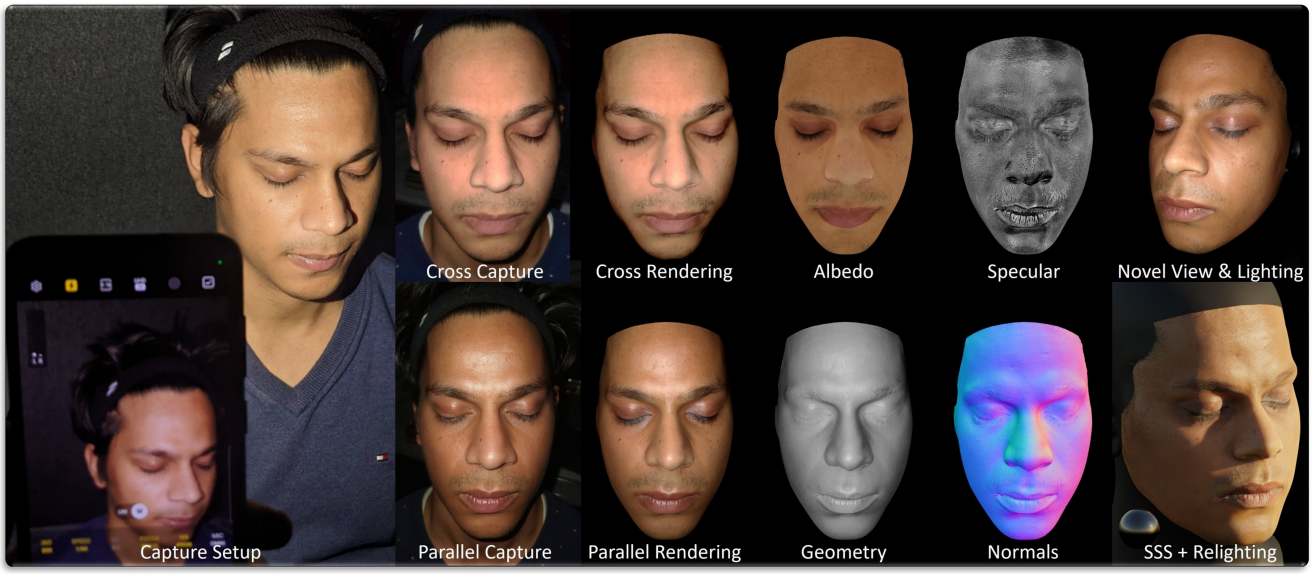

Figure 1. Our method obtains high-resolution skin textures from two RGB input sequences captured with polarization foils attached to a smartphone. The core idea is to separate the skin's diffuse and specular response by capturing one cross-polarized and one parallel-polarized sequence. We recover an accurate geometry with multi-view stereo, fit a parametric head model, and employ a differentiable rendering strategy to recover 4K diffuse albedo, specular gain and normal maps. These can be used with off-the-shelf rendering software, such as Blender, to produce photo-realistic images from novel views, under novel illumination and with subsurface scattering (SSS).

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 colocation 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

All data has been captured at the Technical University of Munich.

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 precomputation 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, 27, 57]; 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 [38,42,55]. These techniques rely on the fact that single bounce specular reflection does not alter the polarization state of incoming light. Riviere et al. [44] propose an approach to reconstruct reflectance in uncontrolled lighting, using the inherent polarization of natural illumination. Nogue et al. [41] 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 [52], 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 [56], compensates for motion during the capture [54] 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. [25] demonstrate that binary gradients are sufficient for separating diffuse from specular without polarization. Lattas et al. [29] 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. [43] 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 [58]. The

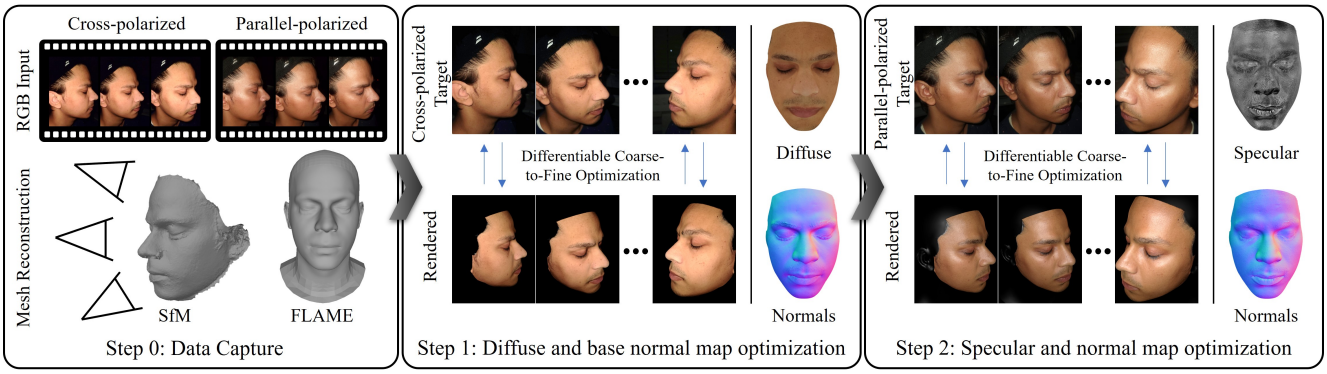

Figure 2. Our optimization has three steps: In step 0, we capture data with a handheld smartphone which is equipped with polarization foils (on the camera, as well as on the flashlight; see Figure 3). We reconstruct the facial geometry and estimate camera poses based on all captured images using structure-from-motion and multi-view stereo. To ensure consistent texture parameterization across different subjects, we non-rigidly fit a FLAME mesh to the scan. In a subsequent photometric optimization step (step 1), we estimate a high-resolution diffuse texture of the skin from the cross-polarized data, as well as an initial normal map. The reconstructed geometry, diffuse and normal map are used as input for step 2 of the optimization. Using the parallel-polarized sequence, we estimate the specular gain and final normal map in a second photometric optimization. In addition, a global skin roughness value is optimized in this step.

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, 51, 60, 61] has led to the development of mature frameworks [23, 28, 40] and a number of methods that try to jointly estimate appearance and lighting [39]. For an overview of differentiable rendering techniques, see [26]. Luan et al. [34] use a co-located camera and light setup to reconstruct shape and material, relying on a differentiable renderer [61] to produce unbiased gradients for shape estimation on an explicit mesh. With the same co-located setup, Zhang et al. [62] 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 [33]. 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. [53] propose a capture setup with the sun as the main light source. A FLAME [32] 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, 30, 31, 36, 45, 59]. Zhang et al. [63] 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, 47, 64] or multi-view [48, 50] 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 [32] 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

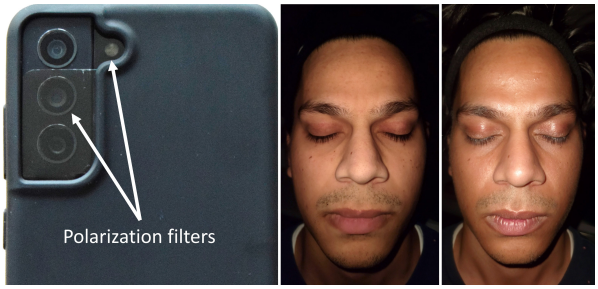

Figure 3. Left to right: smartphone equipped with polarization filters, cross-polarized image (perpendicular filter orientation) and parallel-polarized image (parallel filter orientation).

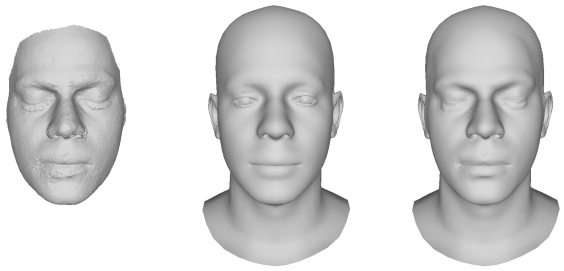

Figure 4. Geometry reconstruction for subject from Figure 3. From left to right: reconstruction via structure from motion, fitted FLAME [32] mesh, ICP-based refinement of the mesh.

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 [65], to make the reconstruction more robust to rigid motion of the head. We then fit the FLAME model [32] 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 $\mathbf{l}$ in a dark environment, the rendering equation that defines the outgoing radiance $L_o(\mathbf{x}, \omega)$, at point $\mathbf{x}$ with normal direction $\mathbf{n}^\top$ in direction ω, has the following simplified form:

$$L_o(\mathbf{x}, \omega) = \frac{f(\mathbf{x}, \omega)(\mathbf{n}^\top \omega) L_i(\mathbf{x}, \omega)}{|\mathbf{x} - \mathbf{l}|_2^2}. \quad (1)$$

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

$$f_s(\mathbf{x}, \omega) = k_s(\mathbf{x}) \frac{D(\omega, \mathbf{n}^\top, \alpha) G(\mathbf{n}, \omega) F(\mathbf{n}, \omega)}{4(\mathbf{n}^\top \omega)(\mathbf{n}^\top \omega)}, \quad (2)$$

with k_s being the spatially-varying specular gain and α 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 [43]. G denotes the geometry term of the Cook-Torrance BRDF model. We use Shlick's approximation [46] for the Fresnel term F:

$$F(\mathbf{n}, \omega) = F_0 + (1 - F_0)(1 - \mathbf{n}^\top \omega)^5. \quad (3)$$

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:

$$f_d(\mathbf{x}, \omega) = \frac{28 k_d(\mathbf{x})}{23\pi}(1 - F_0)(1 - (1 - \frac{\mathbf{n}^\top \omega}{2})^5)^2, \quad (4)$$

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:

$$f_a(\mathbf{x}, \omega) = k_a(\mathbf{x})(1 - (1 - F_0)(1 - (1 - \frac{\mathbf{n}^\top \omega}{2})^5)^2), \quad (5)$$

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, 57], 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:

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

with $\hat{I} = \mathcal{M} \cdot L_o$, where $\mathcal{M}$ 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 $\mathbf{n}^\top \omega$ 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 $(\mathbf{n}^\top \omega)(1 - l)$ 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 $k_d(\mathbf{x})$ and an initial tangent-space normal map $n(\mathbf{x})$, while assuming $f_s(\cdot) = 0$ for the specular term. In the second step, we fix the diffuse texture and optimize for specular gain $k_s(\mathbf{x})$, specular roughness α, and the final normal map $n(\mathbf{x})$. 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 512×512, and increasing the size by a factor of two after convergence of each level, up to the final resolution of 4096×4096.

We implement our optimization framework in PyTorch, using nvdiffrast [28] as our differentiable renderer. We optimize on batches of 4 images, using Adam with an initial learning rate $lr_0 = 10^{-3}$ for all parameters at the beginning of every coarse-to-fine step, and updating it to $lr = lr_0 \cdot 10^{-0.001t}$ 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.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
NLT [63]	31.51	**0.96**	0.11
NextFace [11]	22.85	0.89	0.31
Ours	**32.37**	**0.96**	**0.10**

Table 1. We compare our method to NLT and NextFace on validation frames over 10 different subjects.

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 [63] 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

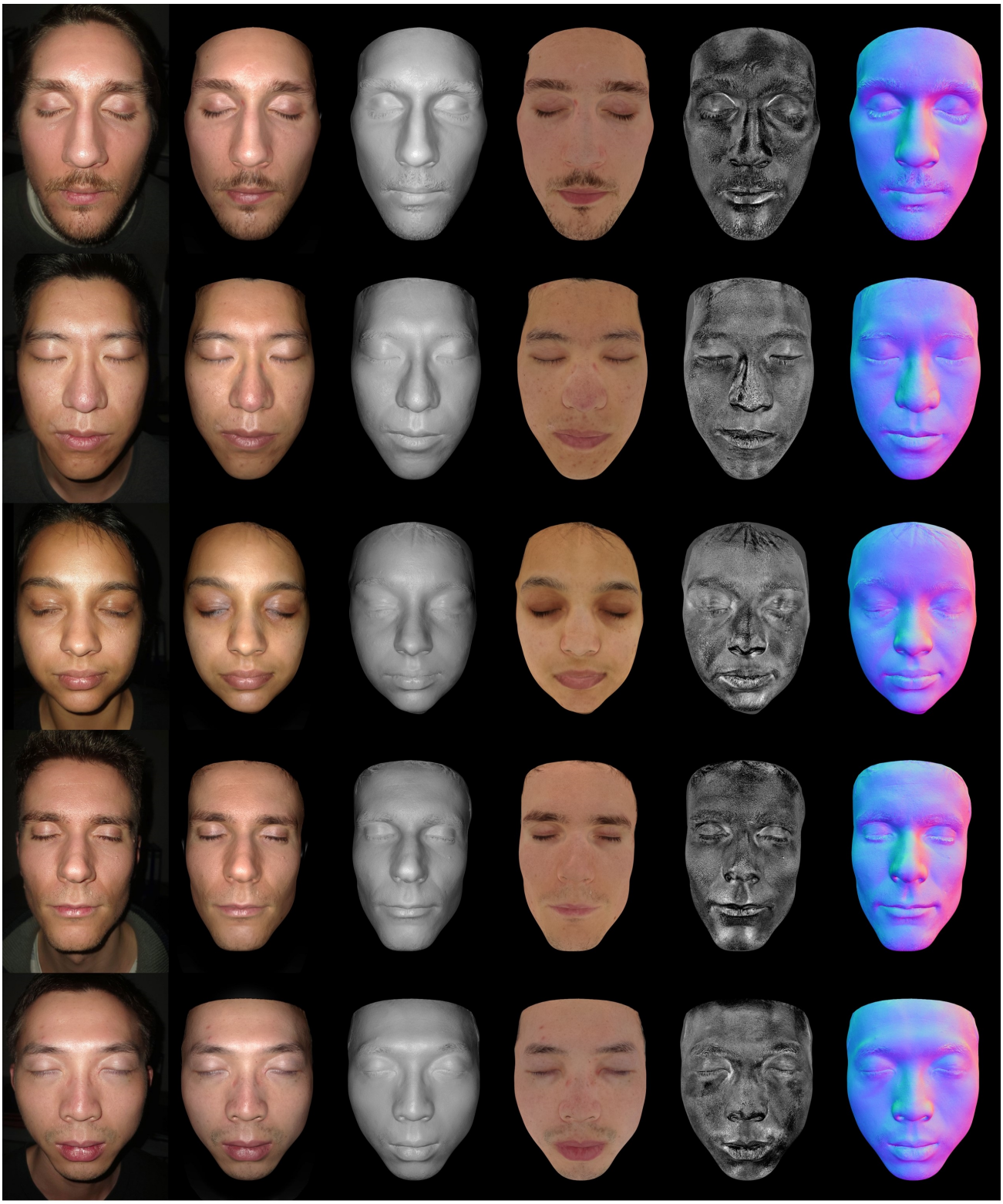

Figure 5. We show skin texture reconstructions of several actors of different skin type. The rendered images closely match the reference target images, and we achieve good separation of diffuse and specular textures.

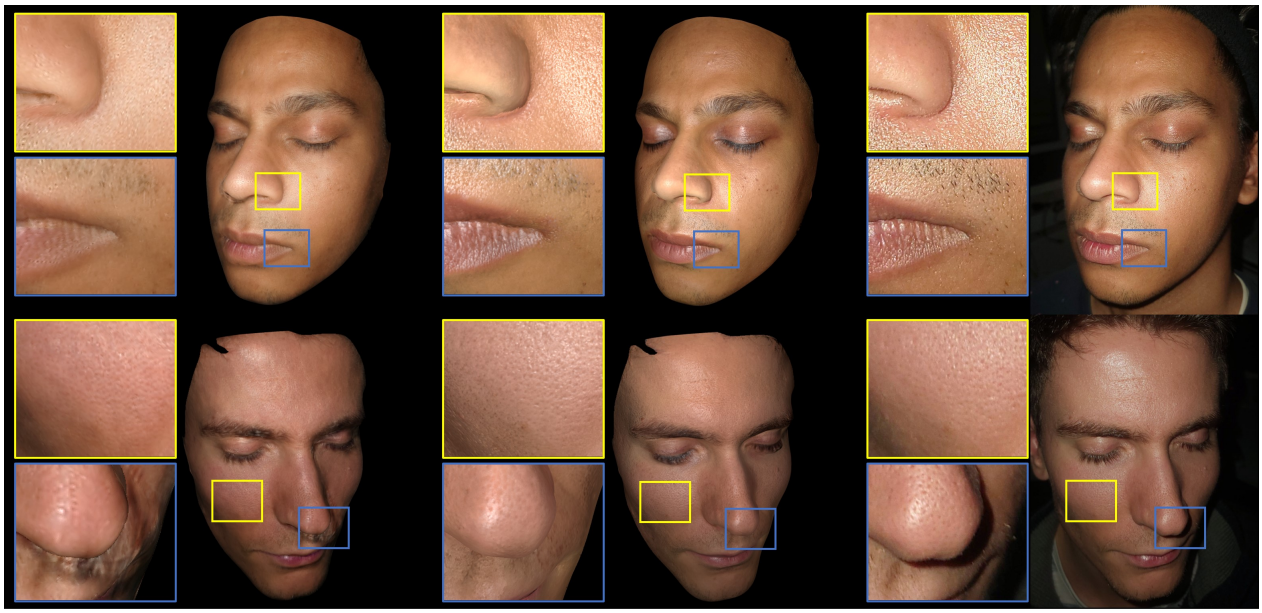

Figure 6. We evaluate on a validation frame from a novel viewpoint and with novel lighting that was held out during the optimization. As visible in the crop regions, our method is able to synthesize sharper texture details and specular highlights compared to NLT [63].

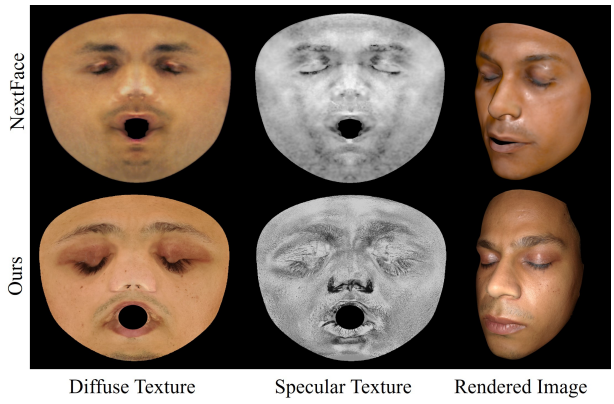

Figure 7. Comparison to NextFace [11] in terms of reconstructed appearance textures.

shaded image. Since the method does not take light intensity and falloff into account, we optimize the rendered validation image's brightness to match the target as closely as possible, before computing the rendering error.

NextFace. NextFace [11–13] first fits a morphable face model to the input frames, then estimates the face shape, pose, lighting, statistical diffuse and specular albedos by minimizing a photo-consistency loss between the target image and a ray traced estimate. In a final step, the statistical albedos are refined on a per-texel basis. We conducted several experiments with different illumination conditions and number of frames, including an experiment on our data for which we replaced the spherical harmonics lighting representation with a small area light, modelling our flashlight.

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.

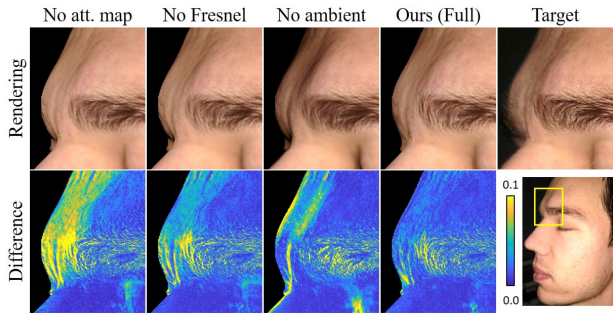

Figure 8. Ignoring the angle-dependent flashlight attenuation, the Fresnel effect, or the ambient light leads to an incorrect reconstruction, that can no longer reproduce the shading from all views. We account for these effects to closely match the target data.

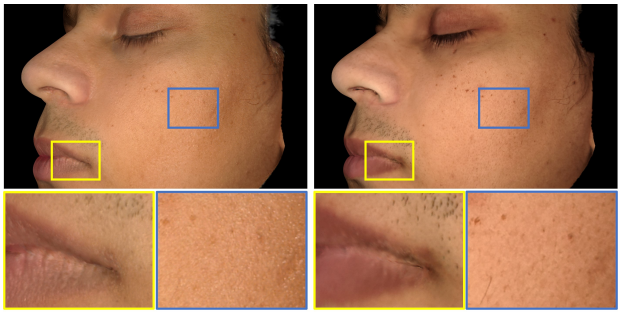

Figure 9. We compare joint optimization of all textures to our full approach on a purely diffuse render. Optimizing jointly leaks specular and normal map information into the diffuse texture.

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 4096×4096 and target image resolution of 3840×2160, 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.

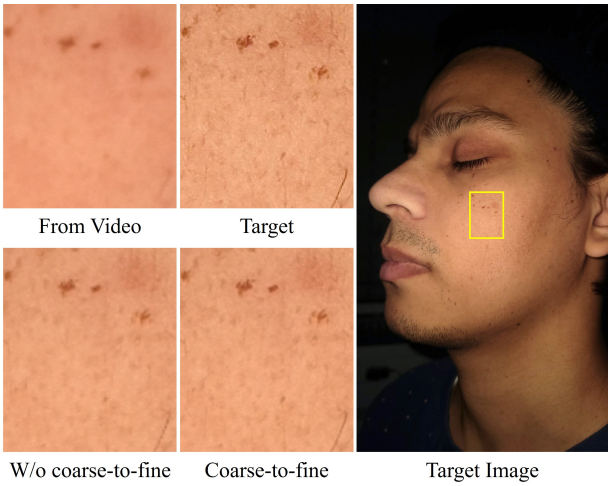

Figure 10. To increase sharpness, we optimize from photographs, instead of video frames. Using only pixels of the appropriate mip level in a coarse-to-fine approach further enhances results.

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

[1] Agisoft. *Agisoft Metashape Professional (Version 1.8.4)*. Agisoft, 2022. 4, 12

[2] Michael Ashikhmin and Peter Shirley. An anisotropic phong light reflection model. *University of Utah Computer Science Technical Report*, 2000. 5

[3] Michael Ashikhmin and Peter Shirley. An anisotropic phong brdf model. *Journal of Graphics Tools 5 (2), 25–32*, 2002. 5

[4] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Niessner. Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[5] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Trans. Graph.*, 40(4), jul 2021. 3, 13

[6] Matt Jen-Yuan Chiang, Peter Kutz, and Brent Burley. Practical and controllable subsurface scattering for production path tracing. In *ACM SIGGRAPH 2016 Talks*, pages 1–2. 2016. 2, 5

[7] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Trans. Graph.*, 1(1):7–24, jan 1982. 4

[8] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. Pandora: Polarization-aided neural decomposition of radiance. *arXiv preprint arXiv:2203.13458*, 2022. 2

[9] 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. 2

[10] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[11] Abdallah Dib, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. *arXiv preprint arXiv:2203.07732*, 2022. 3, 5, 7

[12] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe Gosselin, Marco Romeo, and Louis Chevallier. Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum*, volume 40, pages 153–164. Wiley Online Library, 2021. 3, 7

[13] 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. 3, 7

[14] Graham Fyffe. Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH'09: Posters*, pages 1–1. 2009. 2

[15] Graham Fyffe. Single-shot photometric stereo by spectral multiplexing. In *ACM SIGGRAPH ASIA 2010 Sketches*, pages 1–2. 2010. 2

[16] Graham Fyffe and Paul Debevec. Single-shot reflectance measurement from polarized color gradient illumination. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2015. 2

[17] Graham Fyffe, Paul Graham, Borom Tunwattanapong, Abhijeet Ghosh, and Paul Debevec. Near-instant capture of high-resolution facial geometry and reflectance. In *Computer Graphics Forum*, volume 35, pages 353–363. Wiley Online Library, 2016. 2, 13

[18] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, June 2021. 3

[19] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.*, 30(6):1–10, dec 2011. 2

[20] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Trans. Graph.*, 37(6), dec 2018. 2

[21] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. *arXiv preprint arXiv:2112.01554*, 2021. 3

[22] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018. 3

[23] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. Dr. jit: a just-in-time compiler for differentiable rendering. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 3

[24] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, Sep 2021. 12

[25] Christos Kampouris, Stefanos Zafeiriou, and Abhijeet Ghosh. Diffuse-specular separation using binary spherical gradient illumination. In *EGSR (EI&I)*, pages 1–10, 2018. 2

[26] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*, 2020. 3

[27] Oliver Klehm, Fabrice Rousselle, Marios Papas, Derek Bradley, Christophe Hery, Bernd Bickel, Wojciech Jarosz, and Thabo Beeler. Recent advances in facial appearance capture. In *Computer Graphics Forum*, volume 34, pages 709–733. Wiley Online Library, 2015. 2

[28] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 3, 5

[29] Alexandros Lattas, Yiming Lin, Jayanth Kannan, Ekin Ozturk, Luca Filipi, Giuseppe Claudio Guarnera, Gaurav Chawla, and Abhijeet Ghosh. Practical and scalable desktop-based high-quality facial capture. 2022. 2, 13

[30] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020. 3

[31] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph.*, 39(6), nov 2020. 3

[32] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3, 4, 12

[33] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):222:1–222:11, 2018. 3

[34] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, volume 40, pages 101–113. Wiley Online Library, 2021. 3

[35] Nejc Maček, Baran Usta, Elmar Eisemann, and Ricardo Marroquim. Real-time relighting of human faces with a low-cost setup. *Proc. ACM Comput. Graph. Interact. Tech.*, 5(1), may 2022. 13

[36] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[37] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 13

[38] Volker Müller. Polarization-based separation of diffuse and specular surface-reflection. In *Mustererkennung 1995*, pages 202–209. Springer, 1995. 2

[39] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 3

[40] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 38(6):1–17, 2019. 3

[41] Emilie Nogue, Yiming Lin, and Abhijeet Ghosh. Polarization-imaging Surface Reflectometry using Near-field Display. In Abhijeet Ghosh and Li-Yi Wei, editors, *Eurographics Symposium on Rendering*. The Eurographics Association, 2022. 2

[42] Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 2

[43] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph.*, 39(4), jul 2020. 2, 5, 13

[44] Jérémy Riviere, Ilya Reshetouski, Luka Filipi, and Abhijeet Ghosh. Polarization imaging reflectometry in the wild. *ACM Transactions on Graphics (TOG)*, 36(6):1–14, 2017. 2

[45] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[46] Christophe Schlick. An inexpensive brdf model for physically-based rendering. *Computer Graphics Forum*, 13(3):233–246, 1994. 5

[47] Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. A light stage on every desk. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2420–2429, October 2021. 3, 13

[48] Artem Sevastopolsky, Savva Ignatiev, Gonzalo Ferrer, Evgeny Burnaev, and Victor Lempitsky. Relightable 3d head portraits from a smartphone video. *arXiv preprint arXiv:2012.09963*, 2020. 3

[49] Olga Sorkine and Marc Alexa. As-Rigid-As-Possible Surface Modeling. In Alexander Belyaev and Michael Garland, editors, *Geometry Processing*. The Eurographics Association, 2007. 12

[50] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021. 3

[51] Delio Vicini, Sébastien Speierer, and Wenzel Jakob. Path replay backpropagation: differentiating light paths using constant memory and linear time. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3

[52] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. 2, 13

[53] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Cecilia Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. *arXiv preprint arXiv:2204.03648*, 2022. 3

[54] Cyrus A Wilson, Abhijeet Ghosh, Pieter Peers, Jen-Yuan Chiang, Jay Busch, and Paul Debevec. Temporal upsampling of performance geometry using photometric alignment. *ACM Transactions on Graphics (TOG)*, 29(2):1–11, 2010. 2

[55] Lawrence B Wolff and Terrance E Boult. Constraining object features using a polarization reflectance model. *Phys. Based Vis. Princ. Pract. Radiom*, 1:167, 1993. 2

[56] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 2

[57] Magnus Wrenninge, Ryusuke Villemin, and Christophe Hery. Path traced subsurface scattering using anisotropic phase functions and non-exponential free flights. Technical report, Tech. Rep. 17-07, Pixar. https://graphics. pixar. com/library . . . , 2017. 2, 5

[58] Yingyan Xu, Jérémy Riviere, Gaspard Zoss, Prashanth Chandran, Derek Bradley, and Paulo Gotardo. Improved lighting models for facial appearance capture. 2022. 3

[59] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[60] Tizian Zeltner, Sébastien Speierer, Iliyan Georgiev, and Wenzel Jakob. Monte carlo estimators for differential light transport. *ACM Transactions on Graphics (TOG)*, 40(4):1–16, 2021. 3

[61] Cheng Zhang, Zihan Yu, and Shuang Zhao. Path-space differentiable rendering of participating media. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 3

[62] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022. 3

[63] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. Neural light transport for relighting and view synthesis. *ACM Trans. Graph.*, 40(1), jan 2021. 3, 5, 7

[64] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13545–13555, June 2022. 3

[65] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. *arXiv preprint arXiv:2112.03109*, 2021. 4

APPENDIX

In this appendix, we describe in more detail the preprocessing steps that are necessary to run our method. Specifically, in Section A we explain in detail how to calibrate the camera light and sensor and in Section B we give more detail on fitting the FLAME mesh to the Structure-from-Motion scan. In addition to that, we discuss differences to prior work in Section C.

A. Calibration

To use a smartphone as a tool to capture high-quality textures of human faces, we apply a calibration step related to the flashlight and camera sensor. Specifically, we compute a light attenuation map to take into account vignetting effects and the fact that the flashlight is not an ideal point light source, and we color-calibrate the cross-polarized and parallel-polarized images.

Light attenuation map. In the general case, a smartphone's flashlight does not behave like an ideal point light. We observed a significant decrease of light intensity towards grazing angles. To account for this effect, we compute a per-pixel attenuation map that we multiply with our rendered images to match the observations. To this end, we put calibration markers on a white wall and recorded a cross-polarized sequence (see Figure 11). The markers allow us to estimate camera poses for the sequence and provide us a sparse point cloud to which we fit a plane. Finally, we pose an optimization problem:

$$\underset{\mathcal{M},k_d}{\operatorname{argmin}} \left| \left(\hat{I} - I \right) \right|, \tag{7}$$

with $\hat{I} = \mathcal{M} \cdot L_o$, where $\mathcal{M}$ is the light attenuation map, and k_d the diffuse texture. Once optimized, we keep $\mathcal{M}$ fixed for all subsequent face texture optimizations.

Color correction. We color-calibrate both the cross-polarized images and the parallel-polarized images using pre-recorded images of a Macbeth colorchecker board. We compute an affine color transformation matrix to match these calibration images to a reference color chart. This calibration step is done once for the smartphone and then used for all recorded sequences. The effect of this calibration step is shown in Figure 12.

Camera settings. We record our data using a Samsung Galaxy S21 FE 5G. For the video sequences, we use an ISO of 800 and exposure time of 1/60s. The photographs were shot with an ISO of 200 and exposure time of 1/90s. The smartphone's white balance was set to 4900K.

B. Geometry Estimation

To estimate the geometry of a subject, we use the Structure-from-Motion method from MetaShape [1] on the captured data (see Figure 13 for a camera pose visualization). The resulting geometry is noisy and might contain holes, so we fit a 3DMM-based face model to the reconstruction. Specifically, we use PIPNet [24] to detect landmarks on a front-facing image of the face. These are then projected to 3D using the known camera extrinsic and intrinsic matrices. Using Procrustes's algorithm, we get a coarse alignment between the FLAME face model [32] and the 3D landmarks. We further improve the alignment by optimizing for both a rigid transform between FLAME and nearby scan vertices, as well as the FLAME shape vector to non-rigidly fit the scan. The resulting mesh is subdivided in the face region by a factor of 16, and the eyes are removed from the mesh. Finally, we employ an As-Rigid-As-Possible (ARAP) [49] non-rigid deformation strategy to refine the face mesh, to better align with the reconstruction of MetaShape.

Calibration Markers | Plane Fitting | Reconstructed Light Attenuation Map

Figure 11. To calibrate the light of the smartphone, we record a cross-polarized sequence of a white planar surface with markers for tracking. We fit a UV-parameterized plane to the data and optimize for a light attenuation map which we use for all experiments.

Colorchecker | Original Image | Color-corrected Image

Figure 12. We found that the polarization filters introduce a color shift depending on the polarization direction. To this end, we perform a color calibration with a Macbeth colorchecker board which we capture in both scenarios (cross-, and parallel-polarized). We use an affine color correction to match both captures, and apply this transformation to recordings of all subjects.

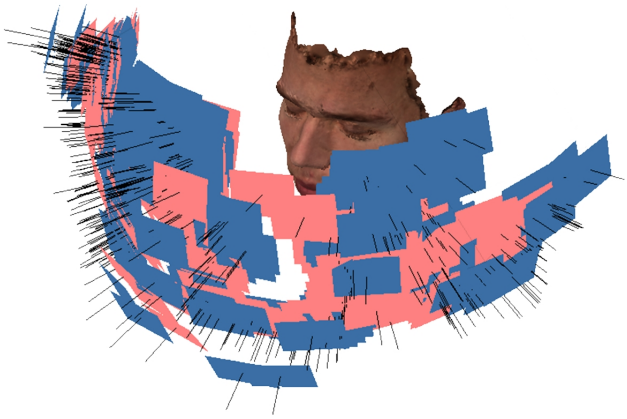

Figure 13. Distribution of cross-polarized (red) and parallel-polarized (blue) views.

C. Comparison to Prior Work

In this section, we explain in more detail the differences between our proposed method and results, and some of the existing solutions for light stage data to which we could not compare directly. Furthermore, we discuss potential benefits of capture setups with independent view and light directions.

MoRF [52] is a generative model trained on a high-quality image database with polarization-based separation of diffuse and specular reflectance. It can generate a volumetric representation of a face based on latent ID codes, which can be optimized to fit new subjects. The database itself is created using the capture setup from [43]. Images of a subject can be rendered by first feeding the subject-specific ID code into a deformation and a canonical MLP. The canonical MLP is composed of a density, diffuse and specular branch, and the output of these branches is used in a volumetric rendering formulation, similar to [37], to render the final image. This is in contrast to our approach, which uses a triangle mesh to represent geometry, and which defines the SVBRDF on the surface of the mesh. The major advantage of MoRF is the fewer number of images it requires at test time and better facial hair and eye handling. This is, however, offset by its limited performance in accurately fitting to faces of new subjects. Furthermore, the material is not separated from lighting and the results are over-smoothed due to the low-order spherical harmonics lighting approximation.

Deep Relightable Appearance Models for Animatable Faces [5] proposes a conditional variational auto-encoder (CVAE) architecture to predict mesh vertices, a corresponding texture warp field and light-dependent textures. A late-conditioned model is first trained on light stage OLAT (one light at a time) data to predict a lit texture map of a subject's face from its average texture (nearest fully-lit frame averaged across all cameras) and an initial estimate of the mesh vertices (provided by an off-the-shelf face tracker). This model has good generalization ability, but is not suitable for real-time rendering. Making use of the good generalization ability of the trained model, a large dataset of synthetic images is generated and used to train an early-conditioned model which can render faces under complex lighting in real-time. The biggest advantage compared to our approach is the capture and rendering of dynamic sequences. Some of the drawbacks include the necessity of a light stage capture setup and the long training time. Futhermore, the model does not separate lighting from material, so its output can not be used in a standard rendering pipeline, or for the creation of virtual assets.

Near-Instant Capture of High-Resolution Facial Geometry and Reflectance [17] performs multi-view color-space analysis to separate diffuse from specular reflectance. Photometric estimation of specular normals further refines geometry compared to the reconstructed base mesh. Similar to our method, and in contrast to the previously described deep learning-based methods, the output is a set of textures that can used in a standard rendering pipeline to render photo-realistic images of a person's face. The carefully calibrated high-cost capture setup, consisting of 24 DSLR cameras, enables reconstruction of fine-scale detail and cannot be matched by current smartphone camera technology. Nevertheless, we see potential benefit of our method's flexibility to capture specular highlights from arbitrary viewpoints, compared to a predetermined set of fixed viewpoints. Another drawback is the necessity of a manual cleanup of the reconstructed multi-view stereo mesh, which is avoided by our method's automated FLAME fitting.

Several prior works [29, 35, 47] on face reconstruction and relighting use a capture setup, in which the light direction is independent from the view direction. While we see potential benefit for convergence speed from the additional constraints provided by such capture setups, given multiple views, our co-located data also provides enough constraints for successful convergence. The shadowing-masking term G is the only term that is directly linked to both the view and light vector. However, by reciprocity of the BRDF, the dependence on view and light direction is the same. Instead of having independent view and light vectors, we found it more important to have a good distribution of the angles between surface normal and view (or light) vector to recover a complete specular and normal map. This is in contrast to [29] and [47] where both camera and light are mostly front-facing.