High-Res Facial Appearance Capture from Polarized Smartphone Images
– Supplemental Document –

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

1. 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 1). 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:

$$\underset{\mathcal{M},k_d}{\operatorname{argmin}} \left| \left(\hat{I} - I \right) \right|, \tag{1}$$

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 2.

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.

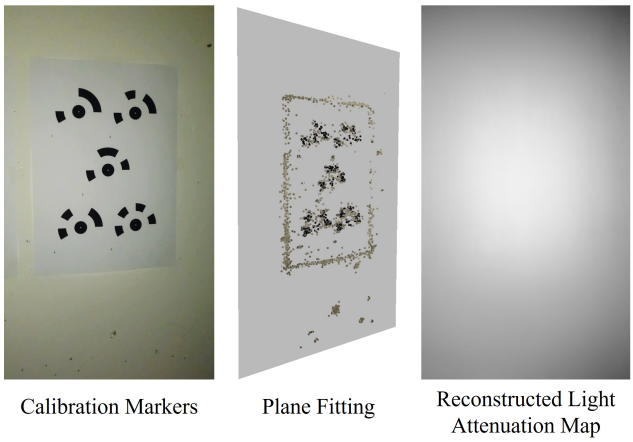

Figure 1. To calibrate the light of the smartphone, we record a cross-polarized sequence of a white planar surface with markers for tracking. We fit a UV-parameterized plane to the data and optimize for a light attenuation map which we use for all experiments.

2. 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 3 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 [4] 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 [6] 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) [11] non-rigid deformation strategy to refine the face mesh, to better align with the reconstruction of MetaShape.

Colorchecker Original Image Color-corrected Image

Figure 2. 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.

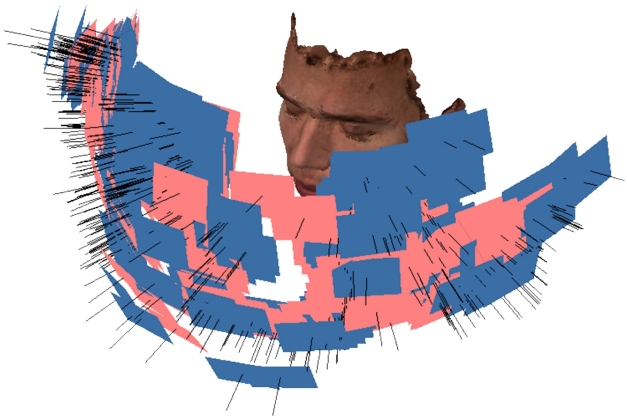

Figure 3. Distribution of cross-polarized (red) and parallel-polarized (blue) views.

3. 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 [12] 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 [9]. 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 [8], 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 [2] 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 [3] 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 [5, 7, 10] 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 [5] and [10] where both camera and light are mostly front-facing.

References

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

[2] 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. 2

[3] 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

[4] 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. 1

[5] 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. 3

[6] 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. 1

[7] 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. 3

[8] 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. 2

[9] 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

[10] 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

[11] Olga Sorkine and Marc Alexa. As-Rigid-As-Possible Surface Modeling. In Alexander Belyaev and Michael Garland, editors, *Geometry Processing*. The Eurographics Association, 2007. 1

[12] 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