Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

Volume 42 (2023), Number 7

Neural Shading Fields for Efficient Facial Inverse Rendering

G. Rainer[1,2] L. Bridgeman[1] A. Ghosh[1,2]

[1]Lumirithmic Ltd, UK [2]Imperial College London, UK

Figure 1: *From casual smartphone frames (left, top) in arbitrary unknown lighting, our inverse optimization uses tiny shading networks to efficiently factorize light transport (left, middle) and disentangle shading from explicit reflectance maps (left, bottom). The resulting asset (mesh & textures) can be used in common pathtracers to render novel viewpoints and lighting (right).*

Abstract

Given a set of unstructured photographs of a subject under unknown lighting, 3D geometry reconstruction is relatively easy, but reflectance estimation remains a challenge. This is because it requires disentangling lighting from reflectance in the ambiguous observations. Solutions exist leveraging statistical, data-driven priors to output plausible reflectance maps even in the under-constrained single-view, unknown lighting setting. We propose a very low-cost inverse optimization method that does not rely on data-driven priors, to obtain high-quality diffuse and specular, albedo and normal maps in the setting of multi-view unknown lighting. We introduce compact neural networks that learn the shading of a given scene by efficiently finding correlations in the appearance across the face. We jointly optimize the implicit global illumination of the scene in the networks with explicit diffuse and specular reflectance maps that can subsequently be used for physically-based rendering. We analyze the veracity of results on ground truth data, and demonstrate that our reflectance maps maintain more detail and greater personal identity than state-of-the-art deep learning and differentiable rendering methods.

CCS Concepts

• ***Computing methodologies*** → ***Reflectance modeling; Neural networks;***

1. Introduction

Rendering faces realistically is a challenging Computer Graphics problem. The interactions between light and skin are very complex due to the translucency of skin layers [DJ05]: Light rays penetrate through the outer layers and bounce around multiple times under the surface before exiting towards the viewer's eye. While the ge-

ometry of the face can be accurately modeled by a surface representation such as a triangle mesh, the appearance is in essence volumetric. Without such subsurface scattering, skin would look too flat and opaque which would break the realism. Humans are very accustomed to looking at faces, meaning the threshold for realism is much higher than for less common objects and materials. Humans are experts at discerning artifacts on skin and faces and accurate reflectance models are essential for rendering.

In this paper, we tackle the problem of reflectance estimation for human faces. In particular, we operate in a multi-view, in-the-wild acquisition setting: Our method takes as input multiple images of a face in unknown, arbitrary and static lighting. Geometry is obtained by *Multi-view Stereo* (MVS), and our technique outputs appearance maps in texture space.

Given a specific appearance model and sufficient appearance data, the standard approach would be inverse rendering. The parameters of the model are tuned until the data can be reconstructed accurately enough. With the development of computing power and GPUs, differentiable rendering [NDVZJ19, JSRV22] has become practicable. Complex pathtracing of a scene can be differentiated and chosen parameters are optimized until the renderings match the input data. However, inverting complex reflectance models to realistically render materials like skin requires a tremendous computational budget and state-of-the-art hardware.

Instead of explicitly simulating the full light transport ray by ray, we take inspiration from *Precomputed Radiance Transfer* (PRT) [SKS02] and rather model the aggregate results. Since lighting is fully unknown, and to avoid making assumptions, we adopt an implicit, learnt representation. Popularized in Computer Graphics by *Neural Radiance Fields* (NeRF) [MST*20], *Multi-Layer Perceptrons* (MLPs) have proven an efficient tool to approximate and learn spatial and directional functions in rendering. We use such MLPs to learn implicit lighting codes and shading kernels in a given scene. We parametrize them carefully and keep the dimensionality low so that the networks cannot overfit and extract other unwanted, coincidental correlations from the data than the real, scene-specific lighting conditions.

Our contributions are as follows:

- A lightweight, implicit light transport model using compact spatial and directional MLPs, used to isolate shading from high-quality facial reflectance maps.
- A low-cost optimization formulation for inverse rendering from multiple views in unknown lighting using explicit texture maps and neural networks rasterized on a static mesh.

2. Related Work

Inverse rendering, light transport modeling and neural rendering have been extensively researched in Computer Graphics. In the following, we focus on face scanning methods that produce reflectance maps of skin, aiming to be used in a physically-based renderer. We refer to Tewari et al. [TTM*22] for a survey on neural rendering including neural face rendering and relighting.

2.1. Multiview Face Scanning

Active Illumination Researchers have traditionally employed controlled illumination from a *Light Stage* setup for estimating facial reflectance from dense lighting measurements coupled with multiview capture [DHT*00, WMP*06]. Polarized spherical gradient illumination has been employed for efficient multiview capture of facial reflectance and photometric normals using a LED sphere with multiple banks of polarization [GFT*11]. Recently, binary illumination has been employed for multiview facial capture with reflectance separation using a practical desktop setup [LLK*22]. However, reflectance separation in these approaches does not rely on multiview observations but on controlled, known illumination.

Passive Illumination Gotardo et al. [GRB*18] employ a multiview passive facial appearance capture setup to estimate dynamic facial reflectance including time varying changes in diffuse albedo and changes in specular reflectance and mesostructure due to skin deformation during facial performance. Riviere et al. [RGB*20a] have proposed a single-shot passive facial appearance capture method that employs polarized illumination panels and a combination of cross-polarized and unpolarized cameras for obtaining high-quality diffuse-specular separation for reflectance estimation with view-multiplexing under passive illumination. These approaches don't vary the lighting but assume a known illumination condition which is exploited in the inverse rendering process.

Phone-based Scanning Recently, Azinovic et al. [AMH*22] introduced a method for facial reflectance scanning using a smartphone with flash illumination in the dark. The flash is used as active illumination and the camera-collocated lighting is exploited in the optimization, in conjunction with polarization filters that provide physical diffuse-specular separation of the appearance. Bao et al. [BLC*21] also use smartphone videos to perform multiview scanning of a subject, but operate in unknown lighting. However, they also use the depth sensor to get additional geometry measurements. Their reflectance decomposition is based on a parametric fit of textures from a PCA-based model, and a Convolutional Neural Network is used to synthesize details on the fitted low-resolution maps. [CSK*22] also exploit the ease of capture of smartphone videos to reconstruct accurate head models through the use of a universal avatar prior that has been trained on high-quality multiview video captures of facial performances of hundreds of human subjects. However, they use a volumetric representation which requires a dedicated renderer and cannot easily be relit.

2.2. Monocular In-the-wild Face Acquisition

With the ease of access to smartphones and the abundance of image datasets, more recent work has focused on removing the two main constraints of controlled acquisition setups – multiple viewpoints and known lighting. Several methods [SWH*17, YSN*18, CCZ*19, BRTO*21] have demonstrated the capability to infer both geometry and relightable reflectance maps from a single image *in-the-wild*, with potentially partial face coverage. Due to the limited and incomplete information, methods such as *Avatarme* [LMG*20] and *Avatarme++* [LMP*21] rely on priors learnt from data: The geometry estimate is based on a morphable model (3dMM), and

a Generative Adversarial Network (GAN) is used to produce the reflectance maps needed for realistic skin renderings: diffuse and specular albedos and normals.

In addition to Deep Learning, many techniques are also based on inverse or differentiable rendering: the unknown parameters are estimated by differentiating the rendering process and iteratively optimizing. Renderers that support both differentiable rasterization and path-tracing are now available [NDVZJ19, JSRV22, MHS*21]. However, a full physically-based differentiable rendering of skin is not yet tractable due to the complex light-material interactions in human skin such as subsurface scattering. Approximations such as direct lighting and opaque BRDFs have to be used to solve the problem in practice.

Dib et al. [DBA*21] propose a method for inverse raytracing from one-or-more images, employing a morphable model with associated PCA diffuse and specular maps. Following a parameteric model fit, they apply a two-stage inverse rendering approach, first estimating the scene shading and a parameterised reflectance estimation, followed by a free-form refinement of the albedos and a roughness map. However, the method does not estimate normals, and struggles to capture a likeness of the subject. Differentiable ray-tracing has enabled the self-supervised training of a neural-network for reflectance, shape and lighting estimation from a single image [DTA*21, DAT*22]. A network regresses the parameters for a morphable model, lighting, and statistical reflectance model from a single image, and a secondary pair of networks predict a detailed offset to the reflectance, capturing personal details.

Caselles et al. [CRG*23] also operate on monocular data by differentiably rendering, but their textures are represented by latents of a model rather than explicitly, to constrain the optimization. Additionally, their renderer uses explicit lights and a standard BRDF model which is not complex enough to accurately represent skin appearance, producing maps that are somewhat lacking in quality.

2.3. MLPs for Scene Decomposition

With the advent of Neural Radiance Fields [MST*20], the use of *Multi-Layer Perceptrons* (MLPs) in scene-specific optimizations has become commonplace to learn / overfit to certain components of the rendering process. NeRFactor [ZSD*21] and NeRD [BBJ*21] extend the original model to include reflectance components and explicit lighting, to output SVBRDFs to relight a given scene. Neural Precomputed Radiance Transfer [RBRD22] and Neural Radiance Transfer Fields [LTL*22] also use MLPs to model the way objects interact with lighting (radiance transfer).

We propose to use MLPs to both learn implicit lighting as well as material shading kernels to convolve this incoming light and create shading layers, circumventing the differentiation of full complex skin appearance models. The output of our method is the other part of our rendering model, represented by explicit texture maps. Although this paper focuses on face scanning, which drives some of the specific decisions, the essence of the approach can be transferred to any scenes and materials.

3. Approach

Matching the standard of realism required in facial skin rendering to look plausible to the human eye requires complex BRDF models which are hard and expensive to invert. Similarly, simple lighting models like *Spherical Harmonics* (SH) will not excite accurate highlights due to the low-frequency nature of the basis. Instead of using an explicit BRDF kernel and explicit lighting model, we propose to use an implicit learnt representation for both.

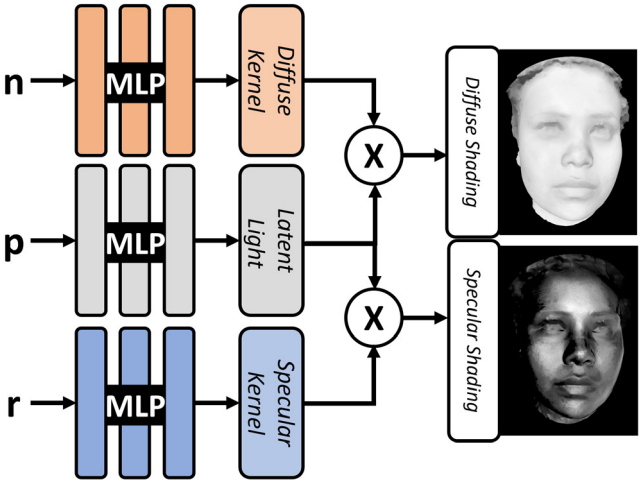

Figure 2: *Three separate MLPs respectively learn spatially-varying lighting, a diffuse and a specular reflectance kernel, parameterized on world-space position* ***p****, normal* ***n*** *and reflected direction* ***r****. View-dependent and -independent shading components are obtained via dot-product of kernels and lighting.*

Tiny MLPs have proven to be a powerful tool to find correlations in light transport data [RBRD22, MRNK21]. In this spirit, we propose to use three MLPs to approximate lighting and shading in the scene. We carefully split the networks and their inputs (see Figure 2) so that they cannot extract other unwanted, coincidental correlations from the data: The two directional MLPs are parametrized on normal direction **n** (for the view-independent component) and on reflected direction **r**, i.e. reflection of the view direction around the surface normal (for the view-dependent component), a parametrization introduced in *Ref-NeRF* [VHM*22], shown to simplify the signal that needs to be regressed.

These networks essentially learn low-dimensional shading kernels, akin to basis projections in PRT. The MLP parameterized on world-space position **p** learns a spatial lighting vector, similar to the transferred incoming radiance vector in PRT. The diffuse and specular shading is then obtained by dot-product of this light vector with the shading kernel vectors, again like in PRT. This dot-product operates in a basis learnt by the networks in a scene-specific manner, thereby encoding information the most efficiently.

While in the following we often refer to these shading components as diffuse and specular, the kernels are more expressive than a choice of explicit BRDF model. The learnt diffuse shading includes some subsurface scattering effects, and the specular kernel can learn a narrower or wider lobe, or even multiple lobes, whichever best explains the data.

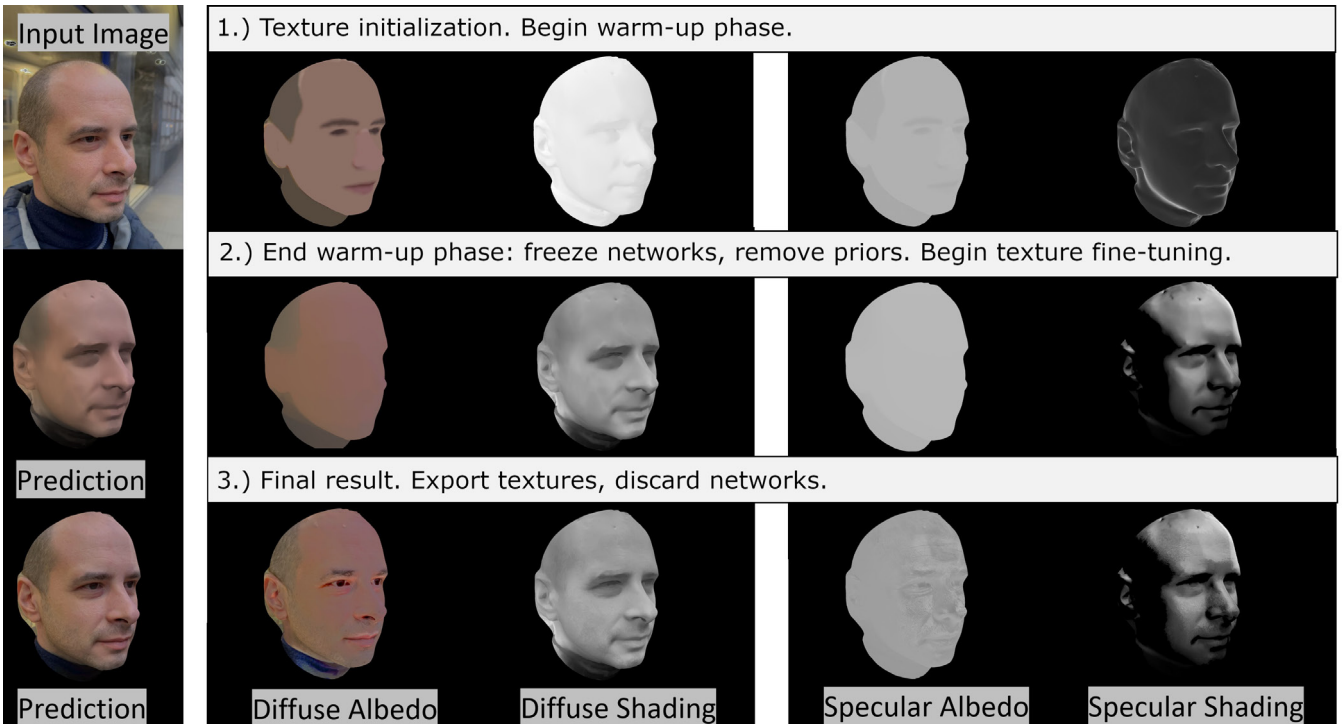

Figure 3: *Evolution of neural networks outputs and explicit albedo textures through different stages of the optimization.*

4. Method

Our method is based on inverse rendering: We jointly optimize shading networks and reflectance maps to best reproduce the observed data using our rendering model. This is performed from scratch for each captured scene and the outputs of the method are reflectance maps, namely diffuse albedo, specular albedo, diffuse normal and specular normal. The shading networks are discarded after optimization as their purpose is solely to disentangle shading from reflectance by leveraging correlations across space and directions.

For any subsequent physically-based rendering using the maps, the implicit roughness is set to a standard skin roughness value, which is common in skin appearance capture [GFT*11, GRB*18]: solutions usually estimate either roughness or detail normal and fix the other, to constrain the optimization.

4.1. Input Data

Our method operates on multiple views of a subject in unknown, arbitrary lighting. We assume geometry has been reconstructed using a prior method, and that all camera poses have been estimated. Our contribution lies in the reflectance decomposition, hence the method takes multiple posed images and a mesh as input.

4.2. Rendering Model

Although the inputs to the networks are three-dimensional in space and direction, the optimization operates in pixel-space and uses a pixel-shading model. The networks inherently regress the result of light transport in the scene so no pathtracing is required. We rasterize the mesh and G-Buffer properties in every input camera and supervise the rendering using the input image.

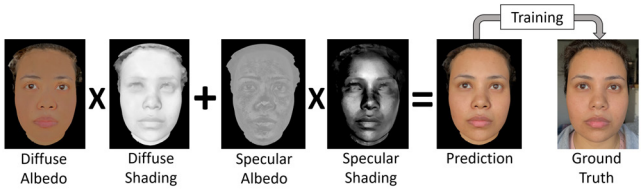

Figure 4: ***Rendering model****: During training, the predicted diffuse and specular shading components are respectively multiplied by the albedos, then summed to produce the predicted rendering, which is supervised by the input images. The specular shading is additionally multiplied by a Fresnel term.*

The rendering is obtained through the simple shading model depicted in Figure 4. The predicted diffuse and specular shading are multiplied by the respective albedos and summed to form the final rendering. The specular shading is multiplied by an additional view-dependent Fresnel term, computed via Schlick's approximation [Sch94]:

$$F(\mathbf{n},\mathbf{v}) = F_0 + (1-F_0)(1-\mathbf{n}\cdot\mathbf{v})^5 \quad (1)$$

where $F_0 = 0.04$ as used by Azinovic et al. [AMH*22]. We found that explicitly extracting the Fresnel term led to better results as the specular kernel MLP learns a simpler function.

4.3. Optimization Strategy

Shading-albedo decomposition from multiple views is an under-constrained problem. In order to encourage the optimization to converge to the preferred solution, we perform it in several stages (as described in Fig. 3) with alternating frozen components, a common strategy to optimize facial reflectance maps ([RGB*20b]).

4.3.1. Warm-up Phase

In the warm-up phase, we jointly train albedos and MLPs. We add a prior on the diffuse albedo to encourage piece-wise constancy. In practice, we formulate this as an additional loss which is calculated across the texels A as the average of

$$L_{alb} = (\sqrt{1+|\nabla_u A|}-1)+(\sqrt{1+|\nabla_v A|}-1) \quad (2)$$

where ∇_u and ∇_v are the first order derivatives in texture space. Without this prior, the optimization could simply put all view-independent appearance into the diffuse albedo (see ablation in supplemental material). This loss forces the MLPs to find as many correlations as possible in the shading across space and directions, and remove it from the textures.

4.3.2. Fine-tuning

In the fine-tuning phase, we freeze the shading MLPs and only optimize textures. We introduce a new texture which defines a deviation vector from the mesh normal. The specular normal is obtained by normalizing this new vector and used to compute the reflection direction used by the specular kernel MLP. The diffuse normal (used by the diffuse kernel MLP) is obtained by first blurring the perturbation, then adding it to the mesh normal and normalizing the resulting vector.

The textures are optimized in two alternating phases: First we optimize specular albedo and normal perturbation, while freezing diffuse albedo. Then normal and specular albedo are frozen and we only optimize diffuse albedo. The full procedure is repeated twice to produce the final reflectance maps.

5. Implementation Details

5.1. Data Capture

The acquisition process involves capturing a 10-20 second video of the subject, with a range of viewpoints extending to roughly 45°from the centre. We assume that the subject is standing still, and that the scene lighting is static. The datasets showcased in this paper are captured with the back camera of the iPhone 14 Pro. The capture comprises pairs of low and high exposure frames, which are aligned and combined into HDR images in a pre-processing step. The mesh reconstruction pipeline takes all these HDR video frames as input, although the geometry is out of scope for this work, and can be substituted with any equivalent method. We simply ensure that there is sufficient data to obtain accurate geometry such that it has no impact on our method.

Our contribution lies in the reflectance estimation for which we use 10 randomly chosen viewpoints. We refer to the supplemental for an ablation on the number of views; we found that more views increase blurring through fine-scale reprojection errors, while fewer views result in limited coverage of the face. All results shown are obtained from 10 views except the comparison to Lattas et al. [LLK*22] where 5 views are used, and the Digital Emily data where 9 views are provided. We demonstrate in an ablation study in the supplementary material that the method is quite robust to the exact number of views, which could be reduced. In practice, we simply choose 10 random views to ensure there is sufficient coverage of the face.

5.2. Geometry Reconstruction

We recover the geometry of the subject using Agisoft Metashape [Agi21], which applies a standard structure-from-motion and multi-view stereo pipeline to the images. However, the geometry reconstruction may be substituted with any viable method, including morphable model fitting, or neural reconstruction. The mesh is automatically trimmed and unwrapped into a continuous UV layout. The aim of these steps is to reconstruct accurate, smooth geometry, with a continuous UV mapping since our albedo prior loss computes gradients in the UV texture space.

5.3. Training Details

The diffuse albedo texture is initialized semantically (see Figure 3). We use the face parser of Zheng et al. [ZYZ*21] to segment the input views into various face components (eyes, mouth, skin, hair etc.) and reproject those labels to the uv-domain. This allows us to segment the diffuse albedo texture into different regions, which are all initialized using the average of reprojected pixel colors. The specular albedo is then initialised as a grayscale version of the diffuse albedo, re-scaled such that it fits in the $0-0.05$ range. The specular normal deviation is initialized as zero, and activated with a tanh, then multiplied by 0.5 to restrict the range of deviation. In the warm-up phase, the texture resolution is 128×128 pixels, which is then upsampled to 2048×2048 in the fine-tuning phase. The final amount of detail visible in the textures strongly depends on the input photo resolution.

Tiny MLPs. The neural networks all have 2 hidden layers of 16 neurons with ReLU activations, and their output (latent vectors used in the shading dot-product) is 8-dimensional. We provide a study of the influence of the dimensionality of the MLPs in the supplemental: it is empirically balanced such that the networks have sufficient expressive power to learn detailed shading in the scene, without overfitting to texture details. The final activation of the spatial network is an exponential, which ensures non-zero lighting everywhere, and allows easier encoding of high dynamic range lighting. The final activation of the directional MLPs is Softplus, which we empirically found more stable than ReLU, which would not converge once in a while, and softer (no boundary artifacts).

The networks and textures are trained under the L1-loss between renderings and input photographs, weighted by a form-factor term inspired by *Unstructured Lumigraph Rendering* [BBM*01] weights: the product of the cosine between view and normal with the inverse of the blurred depth derivative. This allows use to give

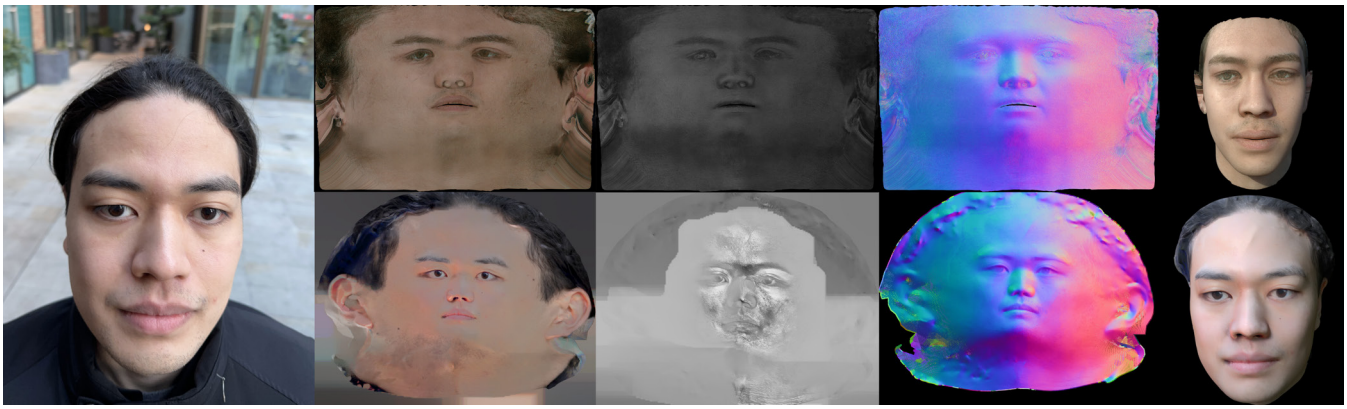

Figure 5: *Left to right: Input image, diffuse albedo, specular albedo, specular normal textures, Blender rendering in novel lighting. Top: Avatarme++ [LMP*21]. Bottom: Ours. Although more complete, the AvatarMe++ textures contain some inpainting artifacts (merged eyebrows) and missing details (cheek moles). Our method also preserves the skin tone and eye color more accurately.*

less relative importance to observation that either have a small footprint due to grazing view angles, as well as less importance to pixels close to depth discontinuities (which are more prone to reprojection errors).

We train with a single (virtual) batch containing all images, for 2000 epochs of warm-up, and 250 epochs of fine-tuning for each of the 4 stages, making a total of 3000 iterations. The warm-up phase is performed at half resolution, the finetuning at full resolution. The rasterization uses Pytorch3d [RRN*20] and the texture sampling and neural network training use PyTorch [PGM*19] and their Adam optimizer with learning rate 0.001. We train on an NVidia RTX3080 GPU and the entire process takes under 30 minutes.

5.4. Normalization of Albedo Maps

During the optimization, both MLPs and textures are unconstrained, meaning there is a scale ambiguity between the albedos and the predicted shading. The albedos hence need to be re-scaled at the end of the optimization to absorb the scale learnt by the respective MLP. For the diffuse component, we assume that the input images are well-exposed, such that the diffuse appearance of the input image should be obtained by multiplying the albedo with the integral of the cosine term (π). We hence normalize the diffuse shading prediction such that its maximum is π, which equates to re-scaling the diffuse albedo by $\frac{max(S_d)}{\pi}$, where S_d is the predicted diffuse shading. Independently, we normalize the specular albedo to the range of human skin reflectance, that is between $0-0.05$ as per Ghosh et al. [GHP*08].

6. Results and Evaluation

The shading MLPs allow us to learn the shading in the given training views. Since they are continuous functions of direction and space, we can trivially render novel viewpoints. Please refer to the supplemental video for a side-by-side comparison of *withheld* ground truth photos and a continuous path through the scene rendered using the MLPs. However, the MLPs encode the integration of shading kernels against lighting, meaning only the viewpoint can be controlled freely, while the illumination is baked in.

6.1. Relighting

For free viewpoint relighting, we discard the MLPs and simply use the mesh along with our predicted maps to relight the faces in Blender Cycles [Ble18], using a custom BRDF model with subsurface scattering and a microfacet BRDF with two-lobe GGX for specular reflectance as suggested by Graham et al. [GTB*13]. Maps and renderings for various subjects are displayed in Fig. 6. Due to the map normalization step, the intensities of specular and diffuse albedo textures do not need manual scaling and can directly be used with the generic skin BRDF we defined. All Blender re-renderings/ re-lightings showcased in the paper are rendered using this custom model.

6.2. Comparisons

We compare to an array of competing methods that output reflectance maps for physically-based renderers, from simple monocular scans, to our setting, freeform in-the-wild video captures, to fully controlled active illumination with camera arrays, to the most expensive Light Stage.

6.2.1. Deep Learning-based Methods

We compare to the deep-learning based, forward technique *Avatarme++* [LMP*21]. It only takes a single image as input and estimates both geometry and reflectance. Due to the limited information in the single view, a big part of the appearance is inpainted and *hallucinated*, leading to plausible textures but loss of identity. Figure 5 shows the results of *Avatarme++* on two different subjects, but the appearance maps and the texture in the re-renderings are quite similar across the two. Since our method is not based on statistical models and completion but purely infers from the multi-view information, it is not influenced by biases of the network or

Figure 6: *Left to right: Example input view, relighting under two different environment maps (Uffizi and Pisa) and under point light illumination in Blender. The top 3 subjects were captured with static cameras (data kindly provided by Lattas et al. [LLK*22], the bottom 3 were captured via casual phone-based videos. Our MLPs plausibly remove baked shading from the albedos across a range of challenging lighting conditions, from controlled uniform (top) to unknown, arbitrary lighting (bottom). Zooming is recommended for full details. Respective appearance maps for these subjects and additional datasets can be found in the supplemental.*

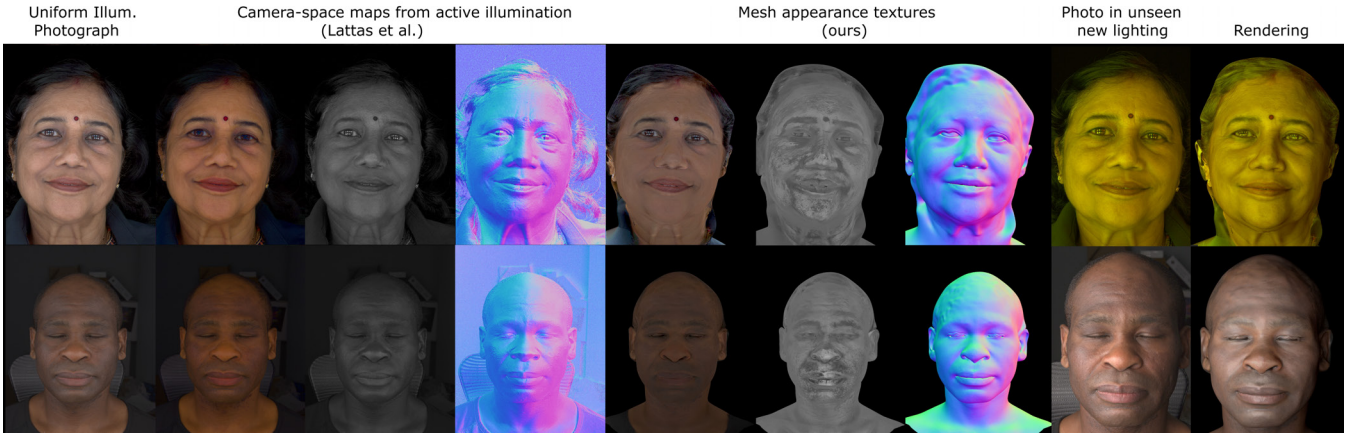

Figure 7: *Comparison of our mesh textures to the camera-space maps [LLK*22] obtained from active illumination. The last frame shows a rendering of our result in Blender, approximately simulating one of the active illumination conditions which are unseen by our method.*

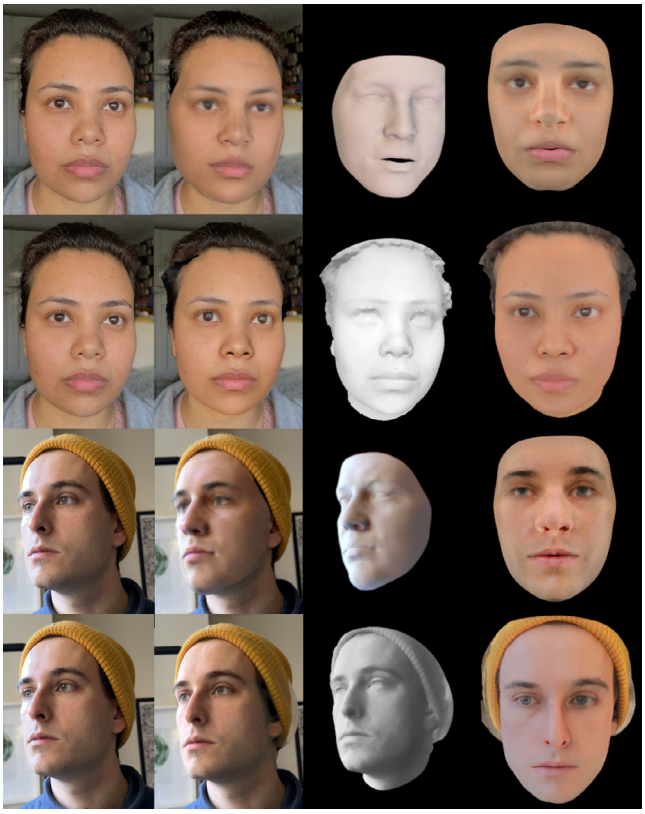

Figure 8: *Top: NextFace [DBA*21]. Bottom: Ours. Left to right: input view, re-rendering (overlaid), estimated diffuse shading, Blender relighting. Dib et al.'s use of statistical models for both geometry and texture biases the estimation towards a more average face, while our results preserve better gender, ethnicity and identity details.*

dataset and better preserves gender, ethnicity and appearance details. For instance, *Avatarme++* produces the same eye color for both subjects, which reflects a bias in their dataset and model.

6.2.2. Differentiable Rendering-based Methods

We also compare to the differentiable rendering method proposed by Dib et al. [DBA*21]. We apply both methods to 10 views of a casually-captured subject under indoor illumination. We compare reconstruction, estimated shading and a re-lit novel view in Figure 8. While the shading estimation is quite accurate for the method of Dib et al., the computation and memory cost of explicit differentiable pathtracing is tremendous, so the optimization is slow (>1 hour) and can only be run on downsized images or it will run out of GPU memory. The resolution (1024×1024) and detail of the albedo maps is hence low and a simplified shape/texture model such as a 3D Morphable Model is used to reduce complexity, resulting in a decrease in likeness and skin tone accuracy. Finally, their method does not estimate specular normals, and the difference in detail in the renderings highlights their importance for realism. Our method comfortably runs with native camera resolution images and 2048×2048 textures and does not use any statistical prior such as the morphable model: both geometry and textures are purely inferred from the captured data, leading to better identity and detail preservation overall.

6.2.3. Analytic Methods

We also compare our estimated reflectance components to the outputs of the desktop-based capture setup of Lattas et al. [LLK*22] who use binary illumination from 8 IPad screens. We apply our method on the same 5 views they captured, under full frontal illumination (sum of the binary patterns), which effectively represents unknown area lighting for our method. The data was kindly provided by the authors.

We show diffuse, specular albedos and specular normals in Figure 7. Lattas et al. reconstruct the SVBRDF components per pixel, in camera space, from the active illumination information, then reproject these camera-space maps onto the mesh to texture it. For

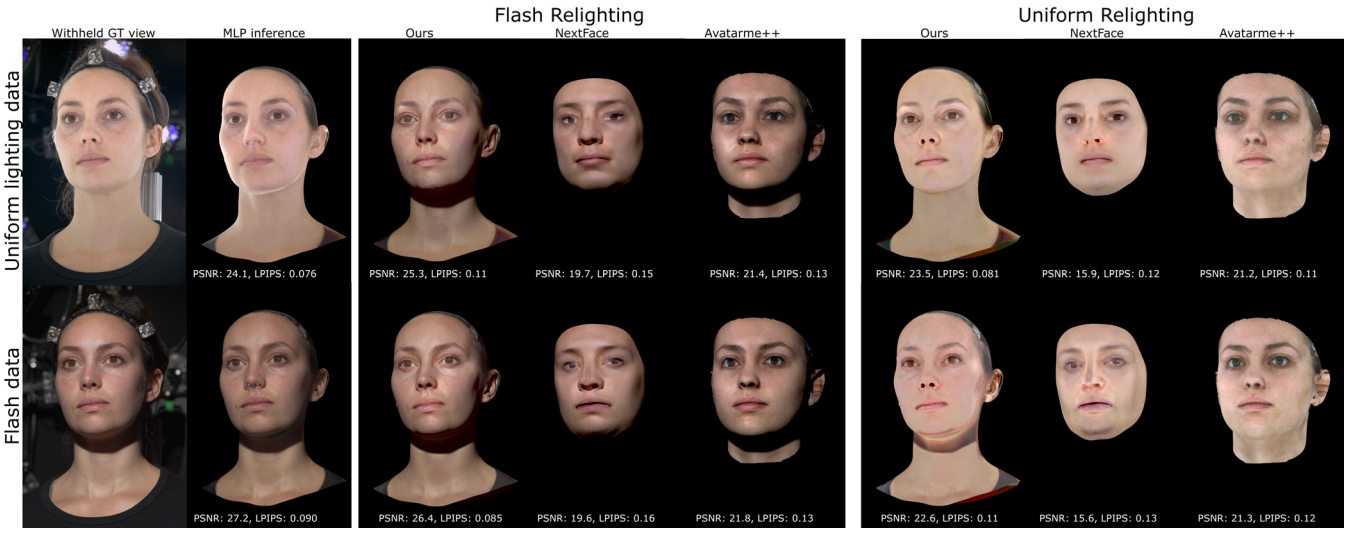

Figure 9: *Quantitative comparison on Digital Emily [The15]. Each row showcases renderings created using results from the respective input capture (left) and the errors are computed against the respective ground truth observation, across the face area.*

the sake of comparison, we rasterize our estimated UV-space textures into the same camera. While slightly lower resolution than the native camera maps, our maps are free of baked shading, which is still present in the outputs of Lattas et al. who do not compensate for the ambient occlusion and limited spatial extent of the lighting.

6.2.4. Light Stage

At the highest complexity and control of scanning stand Light Stage captures, which have traditionally been considered to produce the closest thing to ground truth reflectance maps for faces. The LEDs approximate uniform illumination as closely as possible, and ambient occlusion can be corrected for. This assumed uniformity of the illumination means that shading is considered inherently removed in the lighting itself. The subsequent separation of diffuse and specular albedo is obtained via polarization. Both the illumination as well as the camera lenses are polarized – in parallel position, the cameras capture the entire appearance, while in cross-polarized position, the lenses only permit the diffuse component to pass through.

We run our method on the Digital Emily 2 [The15, ARL*09] dataset, which contains both a flash capture as well as a uniform lighting capture, which allows us to probe the result of our decomposition in the two extremes of illumination conditions (see Figure 10). Under flash illumination, the high dynamic range and narrow field of illumination create very dark shadows in occluded areas like the underside of the chin, which are difficult to remove from the albedos. Nevertheless, the shadows on the cheeks, where there are no occlusions, are adequately removed. Under uniform illumination, the shading networks almost only regress ambient occlusion, and in turn the specular signal is very low. Despite the two extreme cases of illumination, the method still outputs meaningful reflectance maps, which we show in full resolution, compared to the manually curated original Digital Emily 2 maps, in the supplemental, along with a visualization of the learnt diffuse and specular shading modulated by the albedos, compared to the original polarization-based separation.

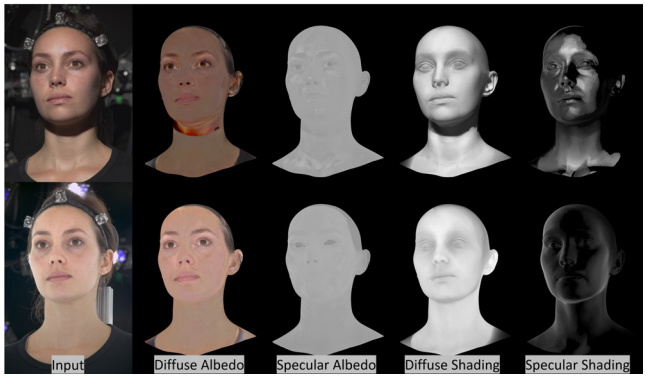

Figure 10: *Reflectance decomposition achieved by our method. Top: Flash lighting. Bottom: Uniform Light Stage lighting.*

Finally, the Digital Emily data also allows to perform a quantitative comparison against competing methods, shown in Figure 9. We compute appearance maps on one capture (e.g. flash) and show view synthesis compared to a withheld view in the same (flash) illumination conditions, and show relighting in the other captures lighting (e.g. uniform), and vice-versa. The image metrics suggest that, as expected, the view synthesis which directly uses the MLPs in the original scan lighting achieves the lowest error, followed only closely by a Blender relighting of our maps in the simulated original lighting. This very small loss in quality when switching from implicit, baked neural shading, to pathtracing in Blender with an explicit SVBRDF model indicates there is a negligible gap between what the MLPs learn and a physically-based skin rendering. Lastly, Blender relighting of our result in drastically opposite lighting con-

ditions (uniform to flash, or vice-versa) still produces plausible results and better fidelity than competing methods.

6.3. Limitations and Discussion

Our contribution lies in the reflectance estimation, so standard *SfM* and *MVS* are used for camera registration and geometry reconstruction. Some of the artifacts in the reflectance decomposition are inherited from this pre-processing. In particular, hair, eye brows, eye lashes etc., will be explained poorly by the reconstructed mesh. Reprojection errors will lead to blurriness in the textures, and reconstruction artifacts (bumps in the geometry) will lead to incorrect albedo estimation as the inputs to the networks will be erroneous in those regions.

Additionally, the images from phone-based scanning are generally lower-resolution and contain slight motion blur or subject movement, which can lead to blur in the estimated textures compared to the amount of detail achievable by the method (see Figure. 6: the top subjects, scanned with the static camera array of Lattas et al. [LLK*22] exhibits higher detail in the maps and renderings than the phone-based captures). Additional examples of results can be found in the supplemental.

The reflectance decomposition is most reliable in areas that are well-observed and well-reconstructed. At the edges of the face / the start of the hair, geometry is not well reconstructed and the reflectance estimation cannot hide the artifacts. Also, since the method uses no statistical priors on the textures, regions that are not sufficiently observed (either in terms of views, or of illumination) will also lack detail. This can be seen in parts of the specular map that will be empty if the subject's face is only illuminated on one side: the specular albedo will not be estimated accurately when there is no specular signal. Casual phone-based captures can also suffer from exposure problems, typically clipping of the specular highlights (see Fig. 8, bottom left, input image), which can induce errors in the estimated maps.

Similarly, the diffuse albedo sometimes tends to become too saturated in areas of extreme shadow. The single diffuse shading model solved for results in spectral color-bleed at shadow boundaries in skin being moved to the diffuse albedo, creating a reddish orange tone. Finally, the method is not designed for hard shadows and colored light: Implicitly, shading is assumed to be smooth and grayscale, otherwise disentangling shading from albedo would be too unconstrained. This disambiguation would require additional information that could be extracted from the background, as shown in *TRUST* [FBT*22].

Although the method does not use any statistical priors, there are some conscious simplifications specific to skin reflectance, for example the semantic initialization of the diffuse albedo, or the final normalization step of the specular albedo. The roughness variation in the scene is assumed to be small, and the roughness is implicitly encoded in the MLPs. The kernel networks learn outgoing radiance distributions, which are weighted by the spatial MLP that outputs multipliers akin to visibility/fall-off for each of these virtual light sources. There is no trivial way to recover a roughness value after training since the networks learn a pre-convolved light model, so we cannot disentangle lobe width in the BRDF from blur in the illumination. Retrieving a roughness value would require known controlled lighting or optimizing an explicit lighting representation. Thankfully, the roughness of skin is empirically known and can be set manually in a physically-based renderer to produce plausible results.

7. Conclusion

We present a method for facial reflectance estimation under unknown lighting conditions, from multiview data. Contrary to most in-the-wild facial reflectance estimation methods, our method does not rely on any data-driven priors on geometry or texture. The reconstruction is only driven by the information contained in the captured scene, thereby avoiding any bias in the results.

The decomposition relies on tiny neural networks to regress shading in a scene, driving the reconstruction of albedo and normal maps. Given scene geometry and posed views, the MLPs efficiently find correlations in the global shading data, which locally allows to disentangle reflectance from illumination. This is especially useful for complex materials such as human skin, for which explicit pathtracing is computationally very expensive to fully differentiate. Neural shading fields are light-weight and much more efficient at regressing and approximating the light transport of a given scene. We demonstrate the quality of results across a variety of subjects and data sources, both from high-quality facial capture systems and from casual phone-based free-form in-the-wild scans.

Future Work. To avoid inheriting some artifacts from the geometry reconstruction, a robust template fitting step would be needed to complete the unobserved parts of the head. Similarly, a post-processing step of inpainting could complete the albedos in unobserved regions (for diffuse) or in unlit regions (for specular).

A middle ground could be found between fully basing the results on the input data and leveraging statistical priors. For example, the reflectance maps could be comprised of patches drawn from a known, measured distribution (e.g. data-driven skin patches) to ensure plausibility while maintaining fidelity to the observations.

The compactness of the MLPs and the low hardware requirements of our optimization setup also opens possibilities for speed-ups and potentially achieving interactive viewpoint change or albedo editing from a phone or webcam video.

8. Acknowledgments

This work was partly supported by EPSRC grant EP/X011364/1 GNOMON. We thank Alexandros Lattas for providing data from [LLK*22] and outputs of Avatarme++ [LMP*21]. For the purpose of open access, the author has applied a 'Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[Agi21] AGISOFT LLC: *Agisoft Metashape Professional (version 1.8.4)*, 2021. 5

[AMH*22] AZINOVIĆ D., MAURY O., HERY C., NIESSNER M., THIES J.: High-res facial appearance capture from polarized smartphone images. arxiv. 2, 4

[ARL*09] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG M., DEBEVEC P.: The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses* (New York, NY, USA, 2009), SIGGRAPH '09, Association for Computing Machinery. URL: https://doi.org/10.1145/1667239.1667251, doi:10.1145/1667239.1667251. 9

[BBJ*21] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCH H. P.: Nerd: Neural reflectance decomposition from image collections. In *IEEE International Conference on Computer Vision (ICCV)* (2021). 3

[BBM*01] BUEHLER C., BOSSE M., MCMILLAN L., GORTLER S., COHEN M.: Unstructured lumigraph rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2001), SIGGRAPH '01, Association for Computing Machinery, p. 425–432. URL: https://doi.org/10.1145/383259.383309, doi:10.1145/383259.383309. 5

[BLC*21] BAO L., LIN X., CHEN Y., ZHANG H., WANG S., ZHE X., KANG D., HUANG H., JIANG X., WANG J., YU D., ZHANG Z.: High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics* (2021). 2

[Ble18] BLENDER FOUNDATION: *Blender - a 3D modelling and rendering package*. Stichting Blender Foundation, Amsterdam, 2018. URL: http://www.blender.org. 6

[BRTO*21] B R M., TEWARI A., OH T.-H., WEYRICH T., BICKEL B., SEIDEL H.-P., PFISTER H., MATUSIK W., ELGHARIB M., THEOBALT C.: Monocular reconstruction of neural face reflectance fields. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021). 2

[CCZ*19] CHEN A., CHEN Z., ZHANG G., MITCHELL K., YU J.: Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 9429–9439. 2

[CRG*23] CASELLES P., RAMON E., GARCIA J., GIRO-I NIETO X., MORENO-NOGUER F., TRIGINER G.: Sira: Relightable avatars from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (January 2023), pp. 775–784. 3

[CSK*22] CAO C., SIMON T., KIM J. K., SCHWARTZ G., ZOLLHOEFER M., SAITO S.-S., LOMBARDI S., WEI S.-E., BELKO D., YU S.-I., SHEIKH Y., SARAGIH J.: Authentic volumetric avatars from a phone scan. *ACM Trans. Graph. 41*, 4 (jul 2022). URL: https://doi.org/10.1145/3528223.3530143, doi:10.1145/3528223.3530143. 2

[DAT*22] DIB A., AHN J., THEBAULT C., GOSSELIN P.-H., CHEVALLIER L.: S2f2: Self-supervised high fidelity face reconstruction from monocular image. *arXiv preprint arXiv:2203.07732* (2022). 3

[DBA*21] DIB A., BHARAJ G., AHN J., THÉBAULT C., GOSSELIN P., ROMEO M., CHEVALLIER L.: Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum* (2021), vol. 40, Wiley Online Library, pp. 153–164. 3, 8

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), pp. 145–156. 2

[DJ05] DONNER C., JENSEN H. W.: Light diffusion in multi-layered translucent materials. 1032–1039. URL: https://doi.org/10.1145/1186822.1073308, doi:10.1145/1186822.1073308. 1

[DTA*21] DIB A., THEBAULT C., AHN J., GOSSELIN P.-H., THEOBALT C., CHEVALLIER L.: 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* (2021), pp. 12819–12829. 3

[FBT*22] FENG H., BOLKART T., TESCH J., BLACK M. J., ABREVAYA V.: Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision* (2022). 10

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. *ACM TOG 30*, 6 (2011). 2, 4

[GHP*08] GHOSH A., HAWKINS T., PEERS P., FREDERIKSEN S., DEBEVEC P.: Practical modeling and acquisition of layered facial reflectance. *ACM Trans. Graph. 27*, 5 (dec 2008). URL: https://doi.org/10.1145/1409060.1409092, doi:10.1145/1409060.1409092. 6

[GRB*18] GOTARDO P., RIVIERE J., BRADLEY D., GHOSH A., BEELER T.: Practical dynamic facial appearance modeling and acquisition. *ACM Trans. Graph. 37*, 6 (Dec. 2018). URL: https://doi.org/10.1145/3272127.3275073, doi:10.1145/3272127.3275073. 2, 4

[GTB*13] GRAHAM P., TUNWATTANAPONG B., BUSCH J., YU X., JONES A., DEBEVEC P., GHOSH A.: Measurement-Based Synthesis of Facial Microgeometry. *Computer Graphics Forum* (2013). doi:10.1111/cgf.12053. 6

[JSRV22] JAKOB W., SPEIERER S., ROUSSEL N., VICINI D.: Dr.jit: A just-in-time compiler for differentiable rendering. *Transactions on Graphics (Proceedings of SIGGRAPH) 41*, 4 (July 2022). doi:10.1145/3528223.3530099. 2, 3

[LLK*22] LATTAS A., LIN Y., KANNAN J., OZTURK E., FILIPI L., GUARNERA G. C., CHAWLA G., GHOSH A.: Practical and scalable desktop-based high-quality facial capture. Springer, pp. 522–537. URL: http://dx.doi.org/10.1007/978-3-031-20068-7_30, doi:10.1007/978-3-031-20068-7_30. 2, 5, 7, 8, 10

[LMG*20] LATTAS A., MOSCHOGLOU S., GECER B., PLOUMPIS S., TRIANTAFYLLOU V., GHOSH A., ZAFEIRIOU S.: Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 2

[LMP*21] LATTAS A., MOSCHOGLOU S., PLOUMPIS S., GECER B., GHOSH A., ZAFEIRIOU S. P.: Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). 2, 6, 10

[LTL*22] LYU L., TEWARI A., LEIMKÜHLER T., HABERMANN M., THEOBALT C.: Neural radiance transfer fields for relightable novel-view synthesis with global illumination. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII* (Berlin, Heidelberg, 2022), Springer-Verlag, p. 153–169. URL: https://doi.org/10.1007/978-3-031-19790-1_10. 3

[MHS*21] MUNKBERG J., HASSELGREN J., SHEN T., GAO J., CHEN W., EVANS A., MUELLER T., FIDLER S.: Extracting Triangular 3D Models, Materials, and Lighting From Images. *arXiv:2111.12503* (2021). 3

[MRNK21] MÜLLER T., ROUSSELLE F., NOVÁK J., KELLER A.: Real-time neural radiance caching for path tracing. *ACM Trans. Graph. 40*, 4 (Aug. 2021), 36:1–36:16. URL: https://doi.org/10.1145/3450626.3459812, doi:10.1145/3450626.3459812. 3

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV* (2020). 2, 3

[NDVZJ19] NIMIER-DAVID M., VICINI D., ZELTNER T., JAKOB W.: Mitsuba 2: A retargetable forward and inverse renderer. *Transactions on Graphics (Proceedings of SIGGRAPH Asia) 38*, 6 (Dec. 2019). doi:10.1145/3355089.3356498. 2, 3

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J., CHINTALA S.: Pytorch: An imperative style, high-performance

deep learning library. In *Advances in Neural Information Processing Systems 32*, Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., (Eds.). Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. 6

[RBRD22] RAINER G., BOUSSEAU A., RITSCHEL T., DRETTAKIS G.: Neural precomputed radiance transfer. *Computer Graphics Forum (Proceedings of the Eurographics conference) 41*, 2 (April 2022). URL: http://www-sop.inria.fr/reves/Basilic/2022/RBRD22. 3

[RGB*20a] RIVIERE J., GOTARDO P., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph. 39*, 4 (July 2020). URL: https://doi.org/10.1145/3386569.3392464, doi:10.1145/3386569.3392464. 2

[RGB*20b] RIVIERE J., GOTARDO P., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph. 39*, 4 (aug 2020). URL: https://doi.org/10.1145/3386569.3392464, doi:10.1145/3386569.3392464. 5

[RRN*20] RAVI N., REIZENSTEIN J., NOVOTNY D., GORDON T., LO W.-Y., JOHNSON J., GKIOXARI G.: Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501* (2020). 6

[Sch94] SCHLICK C.: An inexpensive brdf model for physically-based rendering. *Computer Graphics Forum 13*, 3 (1994), 233–246. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8659.1330233, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8659.1330233, doi:https://doi.org/10.1111/1467-8659.1330233. 4

[SKS02] SLOAN P.-P., KAUTZ J., SNYDER J.: Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *ACM SIGGRAPH* (July 2002), ACM, pp. 527–536. URL: https://www.microsoft.com/en-us/research/publication/precomputed-radiance-transfer-real-time-rendering-dynamic-low-frequency-lighting-environments/. 2

[SWH*17] SAITO S., WEI L., HU L., NAGANO L., LI H.: Photorealistic facial texture inference using deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, jul 2017), IEEE Computer Society, pp. 2326–2335. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.250, doi:10.1109/CVPR.2017.250. 2

[The15] THE WIKIHUMAN PROJECT: Digital emily 2, 2015. URL: https://vgl.ict.usc.edu/Data/DigitalEmily2/. 9

[TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., SIMON T., THEOBALT C., NIESSNER M., BARRON J. T., WETZSTEIN G., ZOLLHÖFER M., GOLYANIK V.: Advances in neural rendering. *Computer Graphics Forum 41*, 2 (2022), 703–735. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14507, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14507, doi:https://doi.org/10.1111/cgf.14507. 2

[VHM*22] VERBIN D., HEDMAN P., MILDENHALL B., ZICKLER T., BARRON J. T., SRINIVASAN P. P.: Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR* (2022). 3

[WMP*06] WEYRICH T., MATUSIK W., PFISTER H., BICKEL B., DONNER C., TU C., MCANDLESS J., LEE J., NGAN A., JENSEN H. W., GROSS M.: Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. Graphics (TOG) 25*, 3 (July 2006), 1013–1024. 2

[YSN*18] YAMAGUCHI S., SAITO S., NAGANO K., ZHAO Y., CHEN W., OLSZEWSKI K., MORISHIMA S., LI H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph. 37*, 4 (jul 2018). URL: https://doi.org/10.1145/3197517.3201364, doi:10.1145/3197517.3201364. 2

[ZSD*21] ZHANG X., SRINIVASAN P. P., DENG B., DEBEVEC P., FREEMAN W. T., BARRON J. T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Trans. Graph. 40*, 6 (dec 2021). URL: https://doi.org/10.1145/3478513.3480496, doi:10.1145/3478513.3480496. 3

[ZYZ*21] ZHENG Y., YANG H., ZHANG T., BAO J., CHEN D., HUANG Y., YUAN L., CHEN D., ZENG M., WEN F.: General facial representation learning in a visual-linguistic manner. *arXiv preprint arXiv:2112.03109* (2021). 5