DOI: 10.1111/cgf.142622
EUROGRAPHICS 2021 / N. Mitra and I. Viola
(Guest Editors)

Volume 40 (2021), Number 2

Practical Face Reconstruction via Differentiable Ray Tracing

A. Dib[1] and G. Bharaj[2,3] and J. Ahn[1] and C. Thébault[1] and P. Gosselin[1] and M. Romeo[3] and L. Chevallier[1]

[1]InterDigital R&I [2]AI Foundation [3]Technicolor Inc.

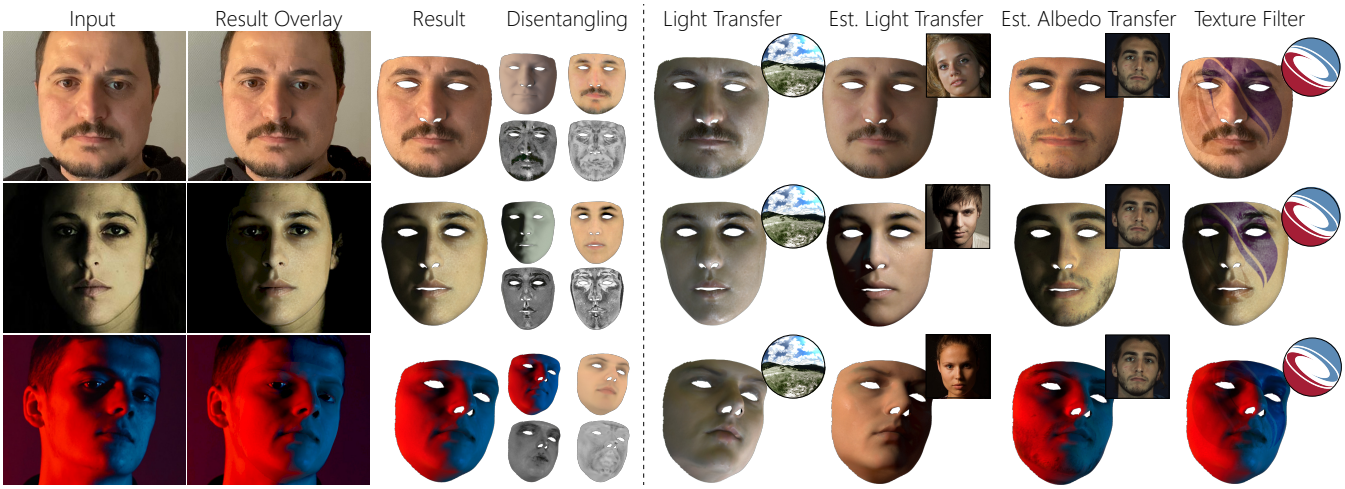

Figure 1: *Our method takes as input an unconstrained monocular face image and estimates face attributes – 3D pose, geometry, diffuse, specular, roughness and illumination (left). The estimation is self-shadow aware and handles varied illumination conditions. We show several resulting style transfer applications: albedos, illumination and textures transfers from and into face portrait images (right).*

Abstract

We present a differentiable ray-tracing based novel face reconstruction approach where scene attributes – 3D geometry, reflectance (diffuse, specular and roughness), pose, camera parameters, and scene illumination – are estimated from unconstrained monocular images. The proposed method models scene illumination via a novel, parameterized virtual light stage, which in-conjunction with differentiable ray-tracing, introduces a coarse-to-fine optimization formulation for face reconstruction. Our method can not only handle unconstrained illumination and self-shadows conditions, but also estimates diffuse and specular albedos. To estimate the face attributes consistently and with practical semantics, a two-stage optimization strategy systematically uses a subset of parametric attributes, where subsequent attribute estimations factor those previously estimated. For example, self-shadows estimated during the first stage, later prevent its baking into the personalized diffuse and specular albedos in the second stage. We show the efficacy of our approach in several real-world scenarios, where face attributes can be estimated even under extreme illumination conditions. Ablation studies, analyses and comparisons against several recent state-of-the-art methods show improved accuracy and versatility of our approach. With consistent face attributes reconstruction, our method leads to several style – illumination, albedo, self-shadow – edit and transfer applications, as discussed in the paper.

CCS Concepts

• ***Computing methodologies*** → *Mesh geometry models;* ***Reflectance modeling;*** *Ray tracing;*

1 Introduction

Photorealistic *avatarized* telecommunication, interactive AR/VR experiences and unobtrusive special effects for professional and consumer applications (e.g. *selfie* filters) require accurate face reconstruction without specialized scene capture and subject/actor constraints. In several such *in-the-wild* scenarios, users lack access to high quality and expensive camera and lighting hardware, or spe-

cialized personnel. For example, while interacting at-home through a monocular front facing camera, the user may encounter harsh self-shadows (for example, shadows cast by the nose or by the superciliary arch on the cheek), multicolored illumination or highly reflective skin conditions. Under varied conditions, consistent reconstruction of face attributes, while avoiding self-shadows biases, etc. is required. The method should work without manual intervention due to consumer constraints, while the reconstruction quality is on par with professional face motion capture systems.

Monocular image-based face reconstruction with meaningful attributes estimation is hard due to its under-constrained nature. Given a face image, its pixel's final color values can be explained by several factors – face shape, skin reflectance, camera position, or light color(s). This ambiguity makes it difficult to consistently estimate attributes. Unknown and unconstrained illumination conditions and consequent face self-shadows further add to the complexity. Our aim is to handle such scenarios using only monocular face images, while maintaining face reconstruction quality. This setup alleviates the need for specialized hardware and light requirements, that opens up avenues for movie production and VFX industry scenarios.

Face reconstruction methods [ZTB*18, TL18, TBG*19, SBFB19] estimate geometry based on parametric face models – 3D morphable model (3DMM) [EST*19]. Such methods assume Lambertian skin reflectance [AS*12] with distant light illumination, where the incoming radiance is a function of direction. Under this assumption, *spherical harmonics* [RH01] have been widely used to model scene illumination. These methods do not model self-shadows. The projected face shape's *geometry-patch* corresponding to color saturated (due to shadows, albedos, illumination) pixel patches can lead to unnatural geometric deformations and inconsistent attribute estimation. More recently, [SYH*17, YS*18, SSD*20, LMG*20] introduce specular reflectance modeling based data-driven priors, however, they do not explicitly handle self-shadows. While, more complete controlled face reconstruction methods [DHT*00, GCP*09, GRB*18] exist, such methods are not applicable for at-home consumer, unobstructed and live performance capture scenarios, due to extensive hardware requirements, and set pre-conditions.

Our objective is 3D face reconstruction with explicit separation of face attributes – skin reflectance (diffuse, specular and roughness), 3D geometry (identity and expression), pose and illumination – from input images. To this end, we use statistical 3DMM to model base face geometry, diffuse and specular albedos priors, along with Cook-Torrance bidirectional scattering distribution function (BRDF) [Sch94] to model skin reflectance. Each vertex on the geometry is characterized by diffuse, specular and roughness parameters; illumination is modeled via a novel virtual light stage with parameterized lights. We also obtain personalized albedos, that refine the statistical 3DMM-based initial estimates. Modeling parameters are used to synthesize an image using differentiable ray tracing, that also obtains self-shadows. Input and synthesized images are used to minimize a photo-consistency loss in two stages, where each stage minimizes a subset of the parameters. We note that although more accurate and complete reflectance modeling approaches [WMP*06] exist, given the quality and nature of input images, the Cook-Torrance reflectance model suffices for our reconstruction needs.

Face attribute reconstruction from monocular images is highly non-linear, our experiments show that naively optimizing all the parameters jointly can lead to poor results. Optimized jointly, specular albedo may get *baked* into diffuse albedo, shadows, etc. Thus, a better strategy for attributes reconstruction is required. We introduce a two stage optimization (Figure 2), where in first stage, similar to [GZC*16, SSD*20] we optimize the pose, illumination, geometry, diffuse and specular albedos, statistically regularized by the 3DMM, while specular roughness remains fixed. Due to ray tracing, the interplay between estimated geometry and illumination helps extract self-shadows. At this stage, person specific (from input image) face attributes such as facial hair, moles, etc. are not estimated. In second stage, we extract unconstrained diffuse, specular and roughness that captures person specific facial details not modeled via statistical diffuse or specular albedos. This staged optimization strategy adds structure and makes the under-constrained optimization problem tractable, leading to superior reconstruction vs. the naive approach. To summarize, the main contributions of our work include:

- A novel virtual light stage formulation, which in-conjunction with differentiable ray tracing, obtains more accurate scene illumination and reflectance, implicitly modeling self-shadows. The virtual light stage models, the switch from point to directional area lights and vice-versa, Sec. 3.
- Face reflectance – diffuse, specular and roughness reconstruction that is scene illumination and self-shadows aware.
- A robust optimization strategy that extracts semantically meaningful personalized face attributes, from unconstrained images, Sec. 4.

To demonstrate the efficacy of our approach we provide several results (Sec 5), ablation (Sec 6) and extensive comparisons (Sec 7) against state-of-the-art methods, where geometric, diffuse and specular albedo estimates are compared. We also compare the proposed light-stage formulation against high-order spherical-harmonic light modeling. Since our method provides fine control over the face attributes, it leads to several style edition and transfer applications (Sec 8) such as face portrait relighting, illumination transfer, specular reflections and self-shadow editing, etc. Scenarios such as changing face pose with accurate resultant self-shadows, or changing illumination, or addition of face *texture* filters, while maintaining original specular albedo (Figure 1), are possible. Finally, in Sec 10 we conclude with limitations and future works.

2 Related Works

Face reconstruction from single, multi-camera images, videos or time-of-flight depth data, is a classic computer vision problem, where the goal is accurate geometry and reflectance reconstruction. With rapid progress in mobile camera technologies, *selfie*-photography, social media, and telecommunication applications, *single camera* face reconstruction approaches has gained special attention. Camera depth ambiguity, capture conditions, non-convexity of face shapes, reflectance properties of human skin, shadows, and illumination conditions make monocular face reconstruction extremely challenging. Several methods have been proposed, that solve for a subset of the face attributes – 3D geometry (neural shape and expressions), pose, diffuse, specular, roughness and illumination (including self-shadows).

Geometry and Reflectance Modeling. [BHB*11] presents multi-view camera and controlled illumination based photogrammetric method

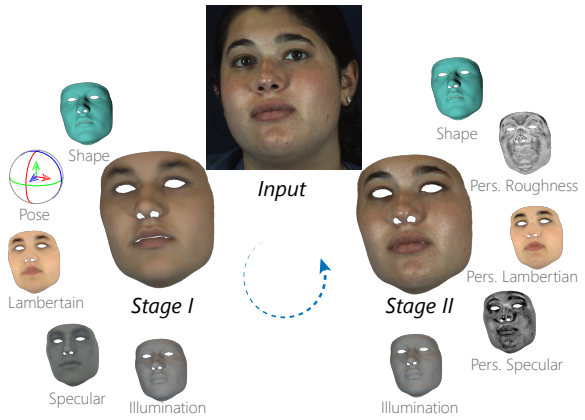

Figure 2: *System Overview: Our method is divided into two stages. In Stage I, for an input image, geometry (pose, identity and expression), statistical diffuse* C *and specular* S *albedos and light stage illumination attributes are optimized. During this stage, the self-shadows are estimated as well. In Stage II, personalized diffuse* $\hat{\mathsf{C}}$*, specular* $\hat{\mathcal{S}}$ *and roughness* $\hat{\mathcal{R}}$ *attributes are estimated. Stage II takes into consideration attributes estimated in the previous stage.*

that produces high-quality (includes mesoscopic face details) temporally stable face geometries. [WVL*11, VWB*12] propose a stereo-image methods for face reconstruction and shape-from-shading based geometry refinement. [GHP*08] captures high-fidelity and multilayered face reflectance using (single camera) multiple images without other external hardware. [GFT*15] captures high quality face geometry and reflectance (diffuse, specular) via a multiview camera setup. More recently, [RGB*20] present a lightweight low-cost rig for high-quality acquisition of facial geometry and appearance with fine-scale pore details.

Photogrammetric and external hardware based approaches provide extremely accurate results, but add constraints on the capture scenarios: multi-view cameras, extensive illumination setups, or lighting conditions (e.g. no self-shadows) for optimal capture. With such approaches if a single camera is used, the reconstruction formulation has infinite deformation degrees-of-freedom, making the problem infeasible. Thus, such methods are not applicable for *in-the-wild* monocular images. Most of these methods do not model specular reflectance and assume a diffuse skin reflectance model.

In-order to use unconstrained monocular images, statistical priors have been introduced [ZTB*18]. Such priors add structure to the reconstruction formulation. 3D Morphable Models (3DMMs) [BV99, LBB*17, EST*19] use facial scanning hardware to capture ground-truth geometry and (diffuse) reflectance. Later, dimensionality reduction method such as principles component analysis (PCA) is used to create linear parametric models. [GVWT13, SKSS14, GZC*16] introduce optimization formulation for geometry (and diffuse) reflectance reconstruction, where 3DMM based priors act as optimization regularizer. They estimate camera parameters and minimize photo-consistency losses based on input images. Such methods also use sparse face image features such as landmarks [SLC11], that *regularize* the optimization against local minima. In-order to separate neutral face shape from expression, FACS [Ekm97] based blendshapes PCA models are used. These methods work well for controlled scene conditions, and often do no generalize well for *in-the-wild* images scenarios. Where they can bake shadows, specularity into diffuse albedo and vice-versa.

[LZL14] extracts diffuse and specular albedos from a single image using Spherical Harmonics (SH) illumination, however, they do not model explicit self-shadows. [TZK*17, TBG*19] use self-supervised autoencoders and inverse-rendering architectures to *infer* 3DMM's linearized semantic attributes. Nonlinear face geometry models such as mesh autoencoders [RBSB18] and CNN encoder [TL18] have also been proposed. Using high quality face datasets and novel deep learning algorithms, [SWH*17, LKA*17, BWS*18] show vast improvements in geometry reconstruction. [HCS*18] shows further improvements by inferring mesoscopic facial attributes given monocular images, an attribute we do not model in our reconstruction approach.

True human skin reflectance capture and reconstruction is a hard problem and several BRDF-based [NRH*92] formulations have been proposed. [T*98, DHT*00, WMP*06, ARL*10, GFT*11] propose extensive measurement systems, structured light setups and data-driven methods. While such approaches lead to highly accurate skin (diffuse and specular) reflectance modeling, they require controlled capture conditions and extensive calibration. Our aim, instead, is to robustly extract face attributes from unconstrained images, where a highly accurate skin reflectance models may not be applicable due to the low quality of input images. [GRB*18] provides a more practical approach to model skin reflectance and ambient occlusions-based shading. Although their setup is less extensive than other approaches, it still requires a controlled multi-view and multi-light illumination setup for reflectance modeling.

Most face reconstruction approaches rely on a lightweight parametric skin reflectance model using linear Lambertian models, where it is assumed that skin *does not* have specular attributes. This simplification has shown great success for face reconstruction [GZC*16, TZK*17, SKCJ18]. Recently, [YS*18, SSD*20, LMG*20] add specular (without roughness) reflectance modeling from unconstrained images, as a result the extracted face models have better attribute disentangling. These methods are more robust against strong self-shadows and specular reflections in input images. However, as discussed in Section 7, they do not *fully* estimate face attributes under several illumination scenarios and bake these attributes in diffuse and specular albedos. While [YS*18, LMG*20] infer geometry and reflectance, but not the illumination. Self-shadows baked into the albedos can be observed, whereas we model self-shadows implicitly.

Illumination modeling. Scene illumination can be modeled via light probes [RHD*10, LYL*16], environment maps [HSL01], sparse mixture of spherical gaussians [KSES14], and illumination model relying on Spherical Harmonics [RH01] that assume Lambertian reflectance. While illumination capture requires specialized hardware, having a linear illumination model limits attributes separation such as self-shadows. Most approaches assume that illumination is mostly uniform resulting in self-shadow being baked into albedo attribute. One way to approach this limitation is to *mask* shadowed patched via occlusion maps, and use GANs [NSX*18] to fill-in the albedos. We approach this problem from a different perspective, similar to initial experiments by [DBA*19] a novel parameterized virtual area light stage is introduced that simulates real world illumination conditions. This illumination model is used together with ray tracing, that implicitly models self-shadow attributes. Consequently, it reconstructs geometric patch's reflectance separating incurred shadows (Sec 3.3). To the best of our knowledge, the proposed method is the first to

estimate reflectance (diffuse, specular, roughness), illumination, and self-shadows robustly from monocular images.

Applications. High-quality face reconstruction leads to several use cases for consumer and movie production scenarios. While quality face tracking has several advantages, such as reenactment, realistic virtual avatars [SSKS17, KGT*18], attributes separation opens up new possibilities. Photoshop-like applications for face portrait touch-up have been proposed. For example, [SPB*14] shows how *style* from one image can be transferred to another employing image-based methods for style transfer. [SHS*17] proposes a method for illumination transfer from source to target images, while [SBT*19] describes a method for portrait relighting. More recently, [ZBT*20] proposes a method for foreign shadow removal from images. Since our method can separates several face attributes, it makes many such applications feasible, as discussed in the paper.

3 Face Modeling Formulation

Overview. We propose a practical formulation to model and reconstruct face attributes. Sec 3.1 describes geometry modeling, and Sec 3.2 describes parameterized reflectance model for diffuse and specular albedo modeling using statistical priors and Cook-Torrance model for personalization. Sec 3.3 introduces our novel parameterized virtual light stage for scene illumination modeling with differentiable ray-tracing. These parametric attributes are then formulated in Sec 4 into an optimization, and solved with a new two-stage optimization strategy (Fig 2).

3.1 Geometry Modeling

Similar to [GZC*16], geometry is modeled via 3DMM and photo-consistency loss. This loss is regularized via a sparse set of face landmarks, where we employ state-of-the-art 2D landmarks estimation [BT17]. This sparse landmark loss (Section 4), helps regularize against local minima where photo-consistency loss is under-constrained, especially under low light, heavy specular or self-shadow conditions. We use [BV99, GMFB*18]'s statistical face model, where identity is given by $\mathsf{e} = \mathsf{a}_s + \Sigma_s \alpha$. e a vector of face geometry vertices with $|\mathsf{e}| = N$. The identity space is spanned by $\Sigma_s \in \mathbb{R}^{3N \times K_s}$ composed of $K_s = 80$ principal components of the identity shape-space. $\alpha \in \mathbb{R}^{K_s}$ describes weights for each coefficient of the 3DMM and $\mathsf{a}_s \in \mathbb{R}^{3N}$ is the average face mesh. We model face expressions over the neutral identity by e via linearized blendshapes $\mathsf{v} = \mathsf{e} + \Sigma_e \delta$, where v is the final vertex position displaced from e by weight vector $\delta \in \mathbb{R}^{K_e}$ and $\Sigma_e \in \mathbb{R}^{3N \times K_e}$ containing $K_e = 75$ principal components of the expression space.

Camera model. We use a pinhole camera model with rotation $\mathsf{R} \in \mathsf{SO}(3)$ and translation $\mathsf{T} \in \mathbb{R}^3$. We assume the camera is always centered at the origin and $\Gamma(\mathsf{v}_i) = \mathsf{R}^{-1}(\mathsf{v}_i - \mathsf{T})$ is the transformation that maps a vertex $\mathsf{v}_i \in \mathbb{R}^3$ to the camera coordinate frame. Π is the perspective camera matrix that maps a 3D vertex to a 2D pixel.

3.2 Reflectance Modeling

We use Cook-Torrance BRDF [CT82, WMLT07] to model face skin reflectance, that defines for each geometry vertex v_i: a diffuse (color) $\mathsf{c}_i \in \mathbb{R}^3$, specular $\mathsf{s}_i \in \mathbb{R}^3$ and roughness $r_i \in \mathbb{R}$ albedos. The BRDF model that defines how the incoming light is reflected on the surface geometry is given by:

$$f_r(\mathsf{s}_i, r_i, \mathsf{c}_i, \mathsf{n}_i, \mathsf{l}, \mathsf{o}) = f_d(\mathsf{c}_i) + f_s(\mathsf{s}_i, r_i, \mathsf{n}_i, \mathsf{l}, \mathsf{o}) \quad (1)$$

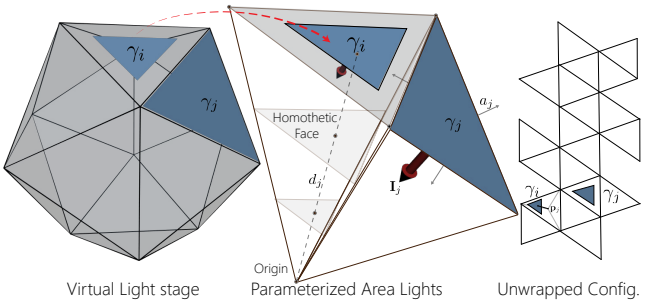

Figure 3: *Left: Our virtual light stage has an icosahedronic geometric construction. Middle: From each of the twenty faces of the icosahedron, we create: parameterized area lights γ_j with intensity i_j, surface area a_j, position d_j and distance to the origin d_j. Right: Unwrapped representation of the icosahedron.*

f_d is the material term for diffused light in all directions. f_s is the specular term for light reflected for a given viewing direction. In contrast to Lambertian BRDF model, the Cook-Torrance BRDF allows us to model specular highlights on the skin's surface. $\mathsf{n}_i \in \mathbb{R}^3$ is the normal at vertex v_i and $\mathsf{l} \in \mathbb{R}^3$ is the incident area light direction (Section 3.3). $\mathsf{o} \in \mathbb{R}^3$ the view direction pointing to the origin of the pinhole camera. For a quick refresher on f_d and f_s reflectance modeling, we refer interested reader to the supplementary material (section A).

The statistical diffuse albedo $\mathsf{c} \in \mathbb{R}^{3N}$ is derived from 3DMM as $\mathsf{c} = \mathsf{a}_r + \Sigma_r \beta$, where $\Sigma_r \in \mathbb{R}^{3N \times K_r}$ defines the PCA diffuse reflectance with $K_r = 80$ and $\beta \in \mathbb{R}^{K_r}$ the coefficients. a_r is the average skin diffuse reflectance. Similarly, we employ the statistical specular prior introduced by [SSD*20] to model the specular reflectance: $\mathsf{s} = \mathsf{a}_b + \Sigma_b \gamma$ where $\Sigma_b \in \mathbb{R}^{3N \times K_b}$ defines the PCA specular reflectance with $K_b = 80$ and $\gamma \in \mathbb{R}^{K_b}$ as the coefficients. a_b is the average specular reflectance. Note that, [SSD*20] recommends using $\gamma = \beta$, however, we use separate parameterization with regularization that leads to similar results with more flexibility.

In unwrapped (UV) image texture space, $\mathcal{C} \in \mathbb{R}^{M \times M \times 3}$ and $\mathcal{S} \in \mathbb{R}^{M \times M \times 3}$ are the statistical diffuse and specular albedos, respectively. $\hat{\mathcal{R}} \in \mathbb{R}^{M \times M}$ defines roughness (no given statistical prior), with $M \times M$ texture resolution. For each projected vertex onto the texture, $\mathcal{C}, \mathcal{S}$ and $\hat{\mathcal{R}}$ describes the interpolated (r, g, b) color, specularity and roughness factors for vertex v_i, where, statistical diffuse albedo $\mathsf{c}_i = \mathcal{C}(\mathfrak{u}_i, \mathfrak{v}_i)$, statistical specular albedo $\mathsf{s}_i = \mathcal{S}(\mathfrak{u}_i, \mathfrak{v}_i)$, roughness $r_i = \hat{\mathcal{R}}(\mathfrak{u}_i, \mathfrak{v}_i)$. $\{\mathfrak{u}_i, \mathfrak{v}_i\} \in [0, 1]$ is projection of vertex v_i onto UV space.

Image-based Personalized Albedo. In Stage I (Section 4), statistical diffuse $\mathcal{C}$ and specular $\mathcal{S}$ albedos are constrained by 3DMM. In Stage II, we personalize albedos using the input image to capture person specific details – facial hair, moles, coloration, and oiliness. Thus, Stage II refines the initially estimated (Stage I) albedo for unconstrained diffuse $\hat{\mathcal{C}}$, specular $\hat{\mathcal{S}}$, and additionally roughness $\hat{\mathcal{R}}$.

3.3 Illumination Modeling

Introduced by [RH01], spherical harmonics (SH), is a method for illumination modeling (assumes light at infinity) with Lambertian reflectance. [DHT*00] introduces a method to capture scene light, that can be used as an environment maps for image-based lighting. [GGSC96] introduced Lumigraph, to model a complex 4D

plenoptic function that describes the flow of light at all positions in all directions for a given scene. Some of these methods require physical apparatus, some are parametrically complex, while others introduce material modeling limitations. In our initial experiments, we formulated illumination modeling using both higher-order SH and environment maps. However, these methods result in sub-optimal self-shadows modeling, and attribute disentangling (see Section 6). For our problem, we need a lightweight yet flexible, parametric scene illumination approach that can not only approximate incoming light, but also model bright, dim, non-uniform, multi-color illumination over non-convex face geometry. Moreover, unlike SH and environment maps, we want to model semantically meaningful light configurations such as point, area, and directional. Thus, we introduce the *virtual light stage* illumination model. For physical face geometry capture, structured light approaches [GCP*09] exist, such methods build physical rigs, known as light stages, with programmable lights and cameras. Inspired from light rigs, we form our virtual light stage that loosely simulates these physical structures to model scene illumination.

To model incoming light on face geometry, we explore various geometric configurations such as a tetrahedron, octahedron, icosahedron and spherical – convex 3D manifolds. Such configurations' triangles can be thought of as area lights, directed towards the manifold's origin. In our experiments, we observe that these light stage configurations practically satisfy the requirements for inctoming light needed for face modeling. During our nascent explorations, we tried very simple structures such as a tetrahedron with four area lights, and more complex geometries like discrete sphere with eighty area lights. Along the various geometric structures, *icosahedron* provides optimal complexity for illumination modeling. See Section 6 for comparisons and Supplementary material for various configurations and resultant face reconstructions.

Virtual Light Stage. A virtual light stage with area lights γ_j, $j \in \{1,...,20\}$, an icosahedron is shown in Fig 3. The shape, size and position of the area lights are derived from the face triangles of the icosahedron. Each area light, modeled independently, has the following parameters: distance $\mathsf{d}_j \in \mathbb{R}$ from the face geometry (at the origin), relative surface area $\mathsf{a}_j \in \mathbb{R}$, local position $\mathsf{p}_j \in \mathbb{R}^2$ of the light center in barycentric coordinates within the face triangle, and perceived intensity $\mathsf{i}_j \in \mathbb{R}^3$. We define $\gamma_j = \{\mathsf{d}_j, \mathsf{a}_j, \mathsf{p}_j, \mathsf{i}_j\}$ as the set of parameters for an area light. Each light can be *switched-off* by setting the perceived intensity parameter i_j to zero. The physical intensity $\mathsf{I}_j \in \mathbb{R}^3$ used for illumination is given by:

$$\mathsf{I}_j = \frac{\mathsf{d}_j^2}{\mathsf{a}_j}\mathsf{i}_j \tag{2}$$

Here, the surface area a_j of the light is relative to the face triangle's area. a_j is bound between 0 – corresponding to a point light, and 1 – maximum surface area of the face triangle. This parameter set has been chosen to better decouple the light parameters. With the standard illumination equation, the light influx reaching an object depends on the physical intensity, distance and size of the light. But, our formulation decouples these parameters and makes it possible to operate only on a single light parameter without effecting other parameters. These variables are orthogonal, and ease the optimization. Without this orthogonal representation, if the effect of a light is too strong, the optimization would have several degrees-of-freedom to change intensity, such as position, size of the light, etc., while, in our formulation, only parameter i_j is needed to modify intensity.

During the initialization, an area light is positioned at center of each triangle of the light stage icosahedron. Each light γ_j can move according to its distance d_j from the geometry center – its size remaining proportional to d_j. a_j and p_j are used to control position and size of each light γ_j within the surface defined by the homothetic face – the icosahedron face scaled by d_j. Thus, the area light remains parallel to original icosahedron's face. A soft box constraint ensures the area lights stay within these homothetic faces (see Section 4). The position and size of the area light control incident light beams, and thus determine the position and the appearance of self-shadows – soft or hard, and specular reflections. When the lights share identical parameters, they are uniformly distributed over 3D angular space; in this case, the model can approximate uniform illumination. The surface of an area light can also become small enough to approximate point light sources.

Shadows approximation. In Section 4, we introduce our optimization formulation that relies on differentiable ray tracing for image synthesis. By varying the number of ray-*bounces* against scene geometries and subsequent indirect illumination, self-shadows can be modeled. That is, *gradient* of shading for a geometric face is dependent on the ray bounces that contribute to incoming light on a face. In our formulation, since we have no information on scene geometry (other than the human face), we do not model indirect illumination due to lack of geometry to bounce-off from. We avoid self-geometry bounces, as in our experiments, it did not lead to substantial gains in accuracy. By using area lights that can be turned *on* or *off*, and by controlling their intensity, position and surface area, we are capable of modeling several illumination and self-shadow scenarios.

4 Optimization

Our goal is robust face reconstruction via geometry (pose, identity and expression), reflectance (diffuse, specular, roughness) and illumination estimation. With unconstrained illumination the optimization can become under-constrained, we therefore resort to a carefully designed two staged optimization strategy. In each stage, Figure 2, we select a subset of the face attributes. Our *analysis-by-synthesis* approach consists in synthesizing an image using parameters $\chi = \{\omega, \alpha, \delta, \beta, \gamma, \mathsf{R}, \mathsf{T}\}$ (where $\omega = \{\mathsf{d}, \mathsf{a}, \mathsf{p}, \mathsf{i}\}$ are the light stage parameters) using differentiable ray tracing [LADL18]. This minimizes a photo-consistency loss between synthesized $\mathcal{I}^S$ and real $\mathcal{I}^R$ images on per pixel basis:

$$\mathsf{E}_{ph}(\chi) = \sum_{i \in \mathcal{I}} |\mathsf{p}_i^S(\chi) - \mathsf{p}_i^R| \tag{3}$$

Here, $\mathsf{p}_i^S, \mathsf{p}_i^R \in \mathbb{R}^3$ are ray traced and real image pixel colors, respectively. Rendered pixel colors are given by $\mathsf{p}_i^S = \mathcal{F}(\omega, \alpha, \delta, \beta, \gamma, \mathsf{R}, \mathsf{T})$, where $\mathcal{F}$ is the Monte Carlo estimator of the rendering equation [Kaj86]. We also define a sparse landmark loss that measures the distance between the projection of $L = 68$ facial landmarks and their corresponding pixel projections z_l on input image:

$$\mathsf{E}_{land}(\chi) = \sum_{l=1}^{L} ||\Pi \circ \Gamma(\mathsf{v}_i^l) - \mathsf{z}_l||_2^2 \tag{4}$$

The sparse landmark loss regularizes the optimization against local minima occuring when photo-consistency loss is ambiguous.

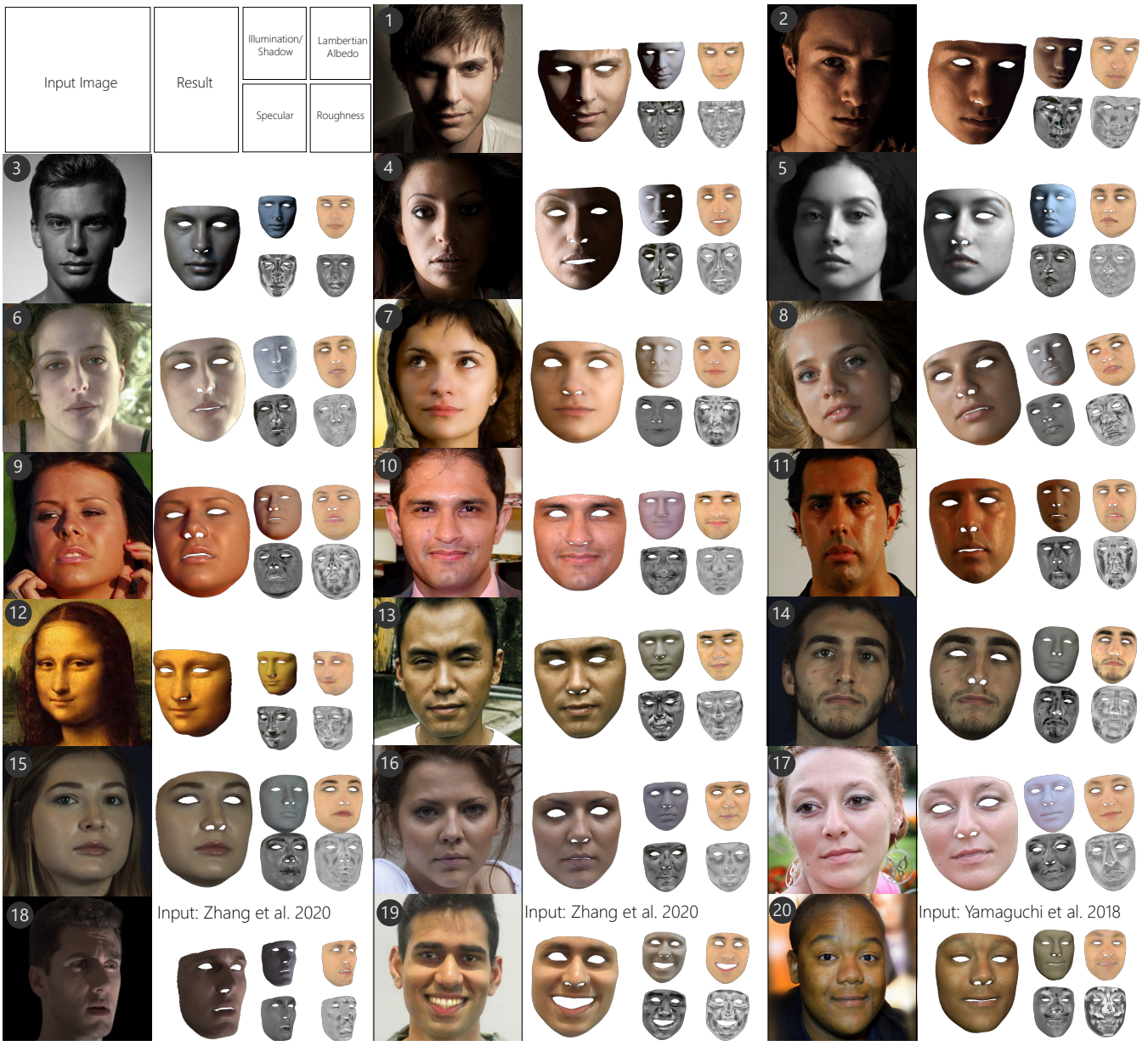

Figure 4: *For each image we show the final optimization result with the estimated parameters: illumination (with estimated self-shadows), diffuse, specular albedo and roughness.*

Optimization strategy. We introduce a two-stage optimization strategy, where Stage I uses statistically regularized albedo priors and Stage II optimizes unconstrained albedos:

Stage I. We optimize camera parameters Γ and blendshape coefficients using the landmark loss (Eq 4). After this pose and expression initialization, we introduce the optimization for statistical albedos (β and γ), face geometry and expression (α, δ), illumination (ω), and camera (R, T), while other parameters – specularity $\hat{\mathcal{S}}$, roughness $\hat{\mathcal{R}}$ and diffuse albedo $\hat{\mathcal{C}}$ – remain fixed. The statistical albedo and *virtual light stage* illumination model guide the optimization and avoid mixing intrinsic albedo and illumination. The loss is:

$$\underset{(\omega,\alpha,\delta,\beta,\gamma,R,T)}{\operatorname{argmin}} \quad E_d(\chi) + E_p(\alpha,\beta,\gamma,\omega) + E_b(\gamma,\delta) \tag{5}$$

With $E_d(\chi) = E_{ph}(\chi) + \alpha_1 E_{land}(\chi)$ and $E_p(\alpha,\beta,\gamma,\omega)$ is a prior that ensures optimization tractability and given by $E_p(\alpha,\beta,\gamma) + w_1 E_p(\omega)$. $E_p(\alpha,\beta,\gamma)$ is the statistical face (shape and albedo) prior that regularizes against implausible face geometry and reflectance deformations, and given by $E_p(\alpha,\beta,\gamma) = w_i \sum_{k=1}^{K_s}(\frac{\alpha_k}{\sigma_{\alpha_k}}) + w_c \sum_{k=1}^{K_r}(\frac{\beta_k}{\sigma_{\beta_k}}) + w_s \sum_{k=1}^{K_s}(\frac{\gamma_k}{\sigma_{\gamma_k}})$. σ_{α_k}, σ_{β_k} and σ_{γ_k} are the standard deviations for shape, diffuse and specular albedo, respectively. Light intensity regularizer $E_p(\omega) = \sum_{j=0}^{M} ||I_j - m_j||_2^2$, where m_j is mean intensity of the j^{th} light. We observe that the final illumination is sensitive to weight w_1, where high value for w_1 leads to monochromatic illumination, while smaller values favor multi-colored illumination. For all our experiments, we use $w_1 = 0.01$, that helps model various

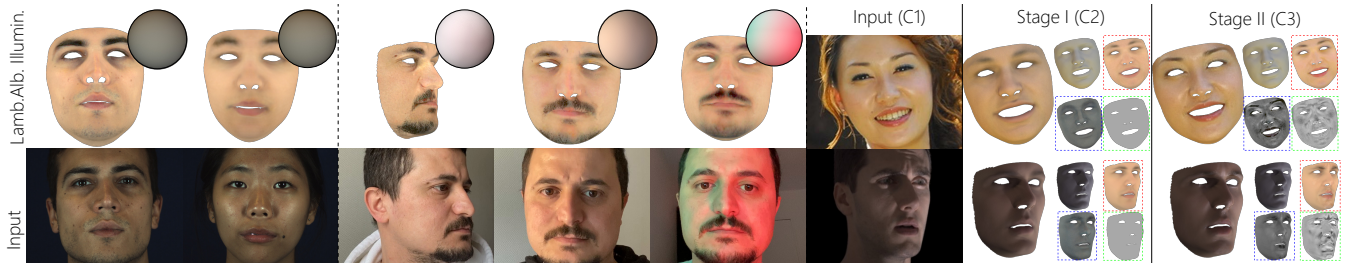

Figure 5: *Left: Consistency of the estimated light for different subjects under the same lighting condition. Right: Consistency of the estimated diffuse albedo for the same subject under different lighting conditions and poses. Right: Stage II, C2 refines the estimated priors of Stage I, C1 and capture person specific facial details in the final diffuse (red), specular (blue), roughness (green) albedos. Even under strong directional light (second row), our method successfully captures the shadows and produces shadows-free personalized albedos.*

illumination scenarios and avoids baking albedos into illumination. Finally, $\mathsf{E}_b(\delta,\omega)$ is a box constraint that restricts δ to range $[0,1]$. $\mathsf{d}_j > 0$, $\mathsf{a}_j > 0$, $\mathsf{i}_j > 0$ and p_j ensure that the area lights stay within the homothetic icosahedron faces.

Stage II. Albedos obtained in Stage I captures the base diffuse and specular statistical albedos. In this stage, we capture personalized face skin attributes – diffuse $\hat{\mathcal{C}}$, specular $\hat{\mathcal{S}}$ and roughness $\hat{\mathcal{R}}$. We use optimized $\mathcal{C}$ and $\mathcal{S}$ to initialize personalized albedos $\hat{\mathcal{C}}$, $\hat{\mathcal{S}}$, and uniform initial roughness $\hat{\mathcal{R}}$ with loss:

$$\underset{(\hat{\mathcal{C}},\hat{\mathcal{S}},\hat{\mathcal{R}})}{\operatorname{argmin}} \mathsf{E}_d(\hat{\chi}) + w_2(\mathsf{E}_s(\hat{\mathcal{C}}) + \mathsf{E}_s(\hat{\mathcal{S}})) + w_3(\mathsf{E}_c(\hat{\mathcal{C}},\mathcal{C}) + \mathsf{E}_c(\hat{\mathcal{S}},\mathcal{S})) + w_4(\mathsf{E}_m(\hat{\mathcal{C}}) + \mathsf{E}_m(\hat{\mathcal{S}}) + \mathsf{E}_m(\hat{\mathcal{R}})) + (\mathsf{E}_b(\hat{\mathcal{S}}) + \mathsf{E}_b(\hat{\mathcal{R}})) \quad (6)$$

Here, $\hat{\chi} = \{\omega, \alpha, \delta, \hat{\mathcal{C}}, \hat{\mathcal{S}}, \hat{\mathcal{R}}, \mathsf{R}, \mathsf{T}\}$ is new parameters set and $\mathsf{E}_b(\hat{\mathcal{S}})$ (resp. $\mathsf{E}_b(\mathcal{R})$) is the soft box constraints that restrict the specular (resp. roughness) to remain in an acceptable range $[0,1]$. $\mathsf{E}_m(\hat{\mathcal{C}})$ (resp. $\mathsf{E}_m(\hat{\mathcal{S}})$ and $\mathsf{E}_m(\mathcal{R})$) is a constraint term that ensures local smoothness of each vertex, with respect to its first ring neighbors in the UV space, and given by $\mathsf{E}_m(\hat{\mathcal{C}}) = \sum_{\mathsf{x}_j \in \mathcal{N}_{\mathsf{x}_i}} ||(\hat{\mathcal{C}}(\mathsf{x}_j) - \hat{\mathcal{C}}(\mathsf{x}_i)||_2^2$, where $\mathcal{N}_{\mathsf{x}_i}$ is 4-pixel neighborhood of pixel x_i.
$\mathsf{E}_s(\hat{\mathcal{C}}) = \sum_{i \in M} |\hat{\mathcal{C}}(\mathsf{x}_i)) - \texttt{flip}(\hat{\mathcal{C}}(\mathsf{x}_i)))|_1$ is a symmetry constraint, where `flip()` is the *horizontal flip* operator, similar to [TL18]. $\mathsf{E}_c(\hat{\mathcal{C}},\mathcal{C})$ is a consistency regularizer that weakly regularizes the optimized $\hat{\mathcal{C}}$ with respect to the previously optimized statistical albedo $\mathcal{C}$ based on the chromaticity κ of each pixel in the texture, given by, $\mathsf{E}_c(\hat{\mathcal{C}},\mathcal{C}) = \sum_{i \in M} |\kappa(\hat{\mathcal{C}}(\mathsf{x}_i)) - \kappa(\mathcal{C}(\mathsf{x}_i))|_1$. $\mathsf{E}_s(\hat{\mathcal{C}})$ and $\mathsf{E}_c(\hat{\mathcal{C}},\mathcal{C})$ help prevent residual self-shadows or specular reflections to bake into the diffuse albedo (same reasoning applies for $\mathsf{E}_s(\hat{\mathcal{S}})$ and $\mathsf{E}_c(\hat{\mathcal{S}},\mathcal{S})$).

Intuitively, when the side of the face is under a shadow, the estimated shadow due to illumination approximation (Stage I), may not fully estimate the real shadow in the input image, while Equation 6 tries to extract meaningful information from the image. Thus, a residual shadow, not fully estimated due to illumination approximation, can get baked into $\hat{\mathcal{C}}$. $E_s(\hat{\mathcal{C}})$ is a symmetric regularizer that prevents baking of the residual shadow into $\hat{\mathcal{C}}$, penalizing for a image-based imbalance between the two sides of the face. $E_c(\hat{\mathcal{C}},\mathcal{C})$ the consistency regularizer, makes sure that diffuse albedo is closer to the statistical diffuse albedo, than the self-shadow's chromaticity.
We note that although the method can be iterated over the Stage I and II, this iteration did not provide substantial improvements in the final results or refinements in disentangling.

Edge Sampling. An important limitation of differentiable ray tracing is the discontinuities present around geometric edges. That is, when solving for the rendering equation [Kaj86] via Monte Carlo ray tracing, very few points on the edge of the geometric shape are sampled, causing a discontinuity along the edges. As a result, back-propagation based gradients calculation fails to take into account sensitive information along the geometric edges. Consequently, the gradients on the edges remain *noisy*, and optimization does not use the *true* gradient during an iteration, especially while optimizing for affine transformations and geometric shape change.
One solution is to use high number of sample points for sampling along edges. However, this is computationally infeasible. Several techniques [LHJ19, LADL18] have been proposed to overcome this limitation. In our work, we rely on [LADL18]'s technique to explicitly sample the geometry edges – a costly yet mandatory operation needed for correct geometric shape estimation.

Variance Reduction. Another aspect when using differentiable ray tracing is image variance due to Monte-Carlo random sampling. Choosing an appropriate sampling strategy can drastically reduce this variance. While, a naive increase in the number of samples can reduce the variance, it is computationally expensive. We use importance sampling [PJH16, LADL18] with 16 samples/pixel and then apply Gaussian smoothing over the synthesized image with a kernel of size 3×3 and $\sigma = 1$. Due to this smoothing operation, variance is considerably attenuated and optimization converges faster.

5 Results and Implementation

We created a dataset of images with various illuminations, self-shadows (hard and soft), ethnicity, facial hair, skin types, expressions and poses to assess the robustness and quality of the reconstruction, Figure 4. For each subject, we show the final reconstruction, along with the estimated reflectance (diffuse, specular and roughness), estimated illumination and self-shadows. Subjects 1-5 (Fig 4) and 2^{nd} and 3^{rd} subjects in Fig 1 shows disentangled attributes of neutral face shapes, expressions, shadow-free albedos and light directions, under challenging lighting conditions.
For Subject 1, the optimized light produces sharp shadows, true to the input image. Subjects 3 and 5, show reconstruction from gray scale input images. Here, a blueish light estimate compensates for the red and yellow components and produces a final gray-scale result similar to the input image, and a meaningful diffuse albedo is also reconstructed (similarly for Subject 12). In addition to handling hard shadows, we show in Subjects 6-8, the ability to produce *soft* shadows. For Subject 6, we get a fair reconstruction under a

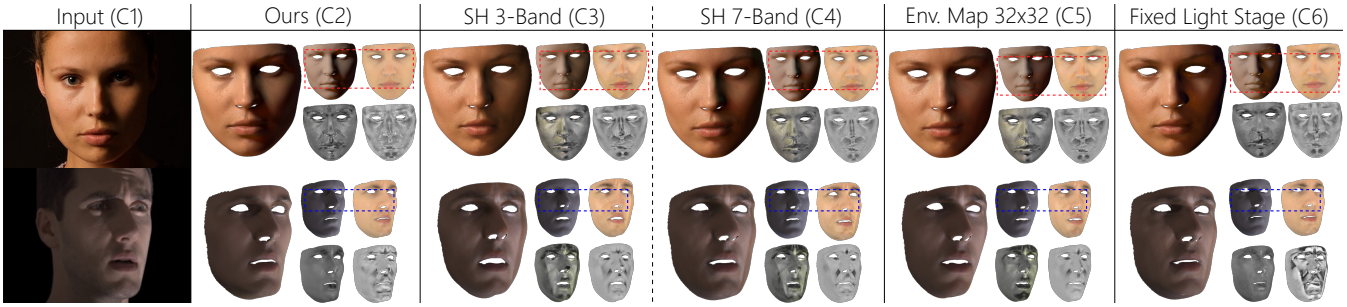

Figure 6: *From left to right. C1: Input image C2: Results obtained by our method C3: Results obtained by using spherical harmonics (SH) 3-bands C4: Results obtained with SH 7-bands C5: Reconstruction using an environment map C6 : Reconstruction using a fixed light stage*

directional light. Subjects 9-11, 15 have visible specular areas on their faces. Our method successfully extracts specular highlights seen in specular and roughness reconstructions.
Subjects 12-17 show reconstructions for people with various skin pigmentations, colorations, facial hair and ethnicities. Our method captures person specific details in the optimized diffuse albedo. Subject 18 (from [ZBT*20]), with challenging lighting conditions is shown, where the face is lit by incoming light from the bottom right † and a hard shadow on the subject's nose. The estimated light captures this shadow and produces shadow-free albedos. Subject 19 is a failure case from [ZBT*20], our method provides a good estimate of self-shadows (especially under the eyes).

Implementation Details. Our framework is implemented using PyTorch [PGC*17] with a GPU enabled backend (NVIDIA GeForce RTX 2080 GPU and Intel i7 9800X). Ray tracing is based on the method of [LADL18], and for optimization we use Adam [KB14] with default $\beta_1 = 0.9, \beta_2 = 0.999$ and $\alpha_1 = 1$. In-order to weight all parameters equally during the optimization, we use different learning rates (`lr`) for each parameter. For light stage parameters we use $\texttt{lr} = 0.001$, for statistical albedo $\texttt{lr} = 0.02$ and for shape identity $\texttt{lr} = 0.01$. Camera rotation, translation and blendshapes use $\texttt{lr} = 0.001$. Finally for the diffuse, specular and roughness, we use $\texttt{lr} = 0.005$. For regularization we use $w_i = 0.0025$, $w_c = w_s = 0.0025$, $w_1 = 0.01$, $w_2 = w_3 = 0.3$ and $w_4 = 0.0002$. The processing time of our method depends on input image resolution. An image of resolution 512×512 takes about 6.4 minutes (wall-clock time) for the full optimization, where Stage I takes 5.1 minutes and Stage II takes 1.3 minutes.

6 Ablation Studies

We show ablation studies on comparison against fixed light stage and the importance of the Stage II to capture personalized skin reflectance. We refer the reader to the supplementary material (section B) for additional ablation studies on the choice of geometries for the light stage.

We validate the importance of our parameterized virtual light stage. A fixed light stage is created, where the light intensity l_j is now a parameter – not dependent on d_j or a_j – fully unconstrained. The light surface-area and position are fixed and not optimized and only the light intensity is optimized. We observe that this optimization formulation gives less accurate shadow estimation and leads to suboptimal light-albedo disentangling (Figure 6, C6). Adding structure to l_j parameterization (Equation 2) leads to substantially better results as shown on Figure 6, C2. Figure 5 (left), discusses the effectiveness of Stage II personalization to refine over Stage I's result. Figure 5 shows the consistency of the estimated light and albedos under various input image and subject conditions.

7 Comparisons

Geometry and Albedo. We compared the geometric reconstruction error against state-of-the-art methods, [TZK*17], [TLL19], [CCZ*19], and [LMG*20], where twenty four ground truth geometries from [GZL18,PJY*19] are used. Our method outperforms these methods and the results are available in the supplementary (Section D). We also compare against state-of-the-art methods [YS*18], [SSD*20]‡ and [LMG*20], that extract both diffuse and specular albedos (Figure 7). Note that methods [YS*18] and [LMG*20] does not model scene illumination and directly infer skin reflectance attributes, so we do not have their final image render. For the same reason, without given illumination, their methods can bake some self-shadow information into the estimated diffuse and specular albedos, as highlighted (in blue) in Figure 7.
We note that [YS*18] and [LMG*20] estimates displacement/normal maps while our method does not. This requires high-quality and well lit input images (as reported by authors) for optimal results. Additionally, [LMG*20] estimates reflectance maps for full face head in the UV space, whereas our method restricts reconstruction to frontal face only. [SSD*20] estimates light (three bands spherical harmonics) but, may not correctly estimate personalized reflectance outside the statistical albedo space. A complete catalog of comparisons against these methods is available in the supplementary material (section C). Additionally, we also compare our method with [TZK*17, TLL19, SKCJ18], see supplementary (Section C).

Digital Emily. In Figure 8, we compare our method with the ground truth (GT) data from the Digital Emily [Emi17] project. In addition, we compare quantitatively, our image reconstruction quality against state-of-the-art (see Table 1). For each method, we compute SSIM (max: 1.0) and PSNR (dB) scores for final render, Ground-Truth (GT) diffuse, and GT specular image pairs (GT roughness not compared due to unavailability). Each image is rendered from the GT camera space using a mask depicted in Figure 8 (bottom-left). As shown in Table 1, our method provides images with the highest

† See supplementary video for shadow edition results.

‡ Using https://github.com/waps101/AlbedoMM

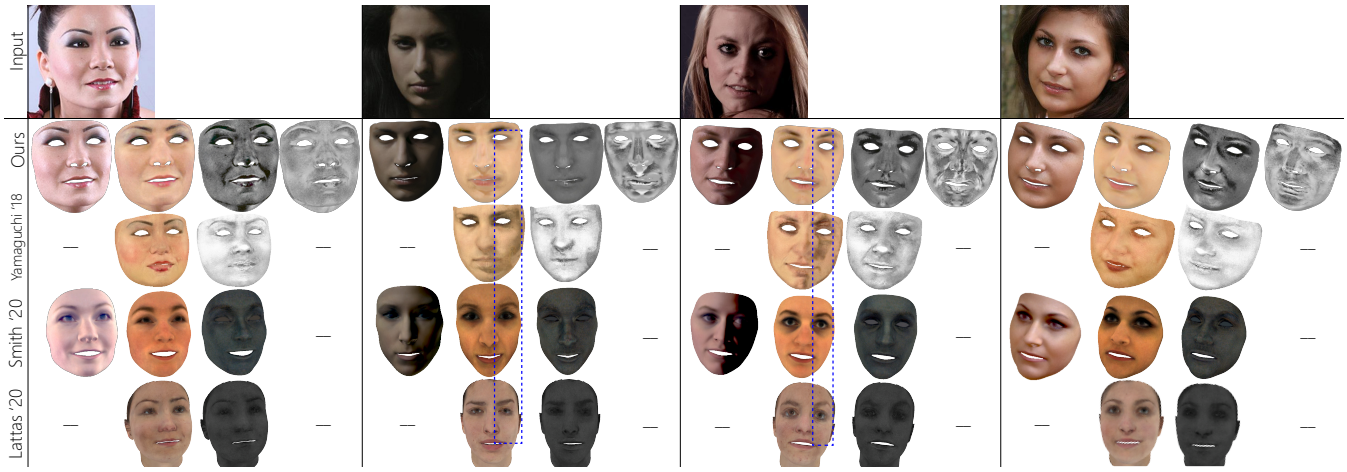

Figure 7: *For each subject (left to right), we compare final reconstruction, diffuse, specular, and roughness albedos with [YS*18, SSD*20, LMG*20]. [YS*18, LMG*20] final reconstruction is not available as their method do not estimate scene light; none of the other methods explicit estimate roughness.*

vs GT Render	Final (SSIM)	Final (PSNR)	Diffuse (SSIM)	Diffuse (PSNR)	Spec. (SSIM)	Spec. (PSNR)
Ours	**0.965**	**36.390**	**0.722**	29.812	0.547	29.670
[YS*18]	-	-	0.679	**30.061**	**0.604**	**30.923**
[SSD*20]	0.906	35.389	0.639	29.006	0.452	28.833
[LMG*20]	-	-	0.540	28.633	0.516	28.926

Table 1: *Final, diffuse and specular albedos in comparison with GT Maya renders for our, [YS*18], [SSD*20] and [LMG*20]. SSIM and PSNR (dB): higher the better.*

similarities in SSIM for diffuse rendered image. For PSNR (diffuse, specular) and SSIM (specular) [YS*18] scores slightly better than our method. Please note that since each method has a different UV map parametrization, we did the comparison on the face mask image renders and not on unwrapped texture space. As [YS*18] and [LMG*20] do not estimate scene light, so we do not have comparison of the final image renders against GT. Finally, we compare rendered GT images (using Autodesk Maya) against input image and obtain $SSIM = 0.973$, $PSNR = 36.526$. We note that our final image render vs. input image have scores $SSIM = 0.982$, $PSNR = 41.475$ that are closer to the input image.

Spherical Harmonics (SH) vs. Light Stage. In this experiment, we use Spherical Harmonics (SH) to model light instead of the light stage (Figure 6). First subject (first row Fig 6), three-bands SH (C3) provides a coarse estimation of the light, and the shadow is barely captured, where estimated albedos get some light and shadows baked into it. Seven-bands SH (C4) captures more shadows but still produces sub-optimal disentangling vs. our light stage (C2). For the second subject (second row), the hard shadow cast by the nose was only captured by our light stage while (3 and 7 bands) SH are visually inaccurate. We also experimented with higher-order SH band (9 and 11) without substantial improvements, especially for subject in row two, Fig 6. These experiments shows that using high-order SH can be used to obtain meaningful shadows estimations, but fails to capture hard shadows produced by point lights in the scene, and leads to sub-optimal disentangling. Finally, our parametric light stage models semantically meaningful light types – point, directional, while basis functions used by SH only model lights at infinity and are harder to manipulate intuitively (e.g. for shadow removal applications).

Environment Map vs. Light Stage. In this experiment, we replaced the light stage with an environment map to model lighting. Each pixel in the environment map , 32×32 resolution, represents a light source at infinity, where light intensity of each pixel is parameterized. Results for this optimization are shown in Figure 6 (C5). Because environment map can only model lights at infinity, is not flexible enough to model arbitrary (e.g area) lights, opposed to the lightstage, and thus, fails to capture the shadows generated by point lights (for both subjects) and produces sub-optimal disentangling.

8 Applications

Robust estimation of reflectance and illumination provides explicit control over these attributes, with several practical applications: relighting, light transfer, shadow and specularity editing, and image texture filters addition.

Illumination Edition and Transfer. Figure 1 (right) first column shows relighting under novel illumination conditions. Second column, shows results for estimated light transfer, where estimated light from source image is used to illuminate target subject. Source image's self-shadows, due to illumination, are successfully transferred in the target render.

Shadow and Flash Removal. ** Inspired by [ZBT*20], we show self-shadow removal application. While, [ZBT*20]'s method can remove shadows cast from external (foreign) objects; our method handles self-shadow removal, as shown in Figure 9 (left). In the accompanying video, we also show demonstration of camera flash removal for face images, where estimated illumination from first image replaces estimated illuminations in subsequent image frames.

Albedo Edition and Transfer. Third column in Figure 1 (right) shows diffuse and specular albedo transfer applications, from thumbnail source to target image, while the last column shows the result of

** The reader is referred to supplementary video for better visualization

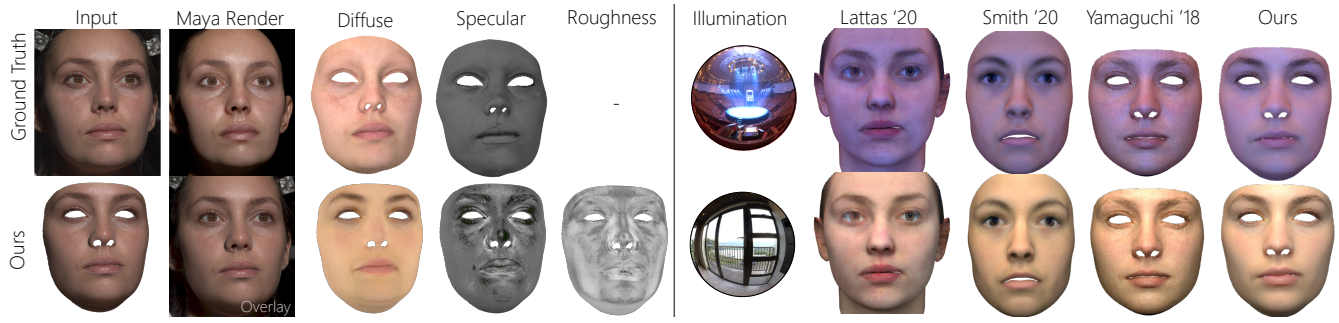

Figure 8: *Left: Comparison of our method with Ground Truth (GT) data of the Digital Emily project. Right: Relighting comparison.*

applying a texture filter (using multiplication operator) on the optimized diffuse albedo in the UV space. Finally, in Figure 9** (right), we show an application where estimated specular albedos can be edited on portrait images. This is done by gradually decreasing the estimated roughness, while using a constant estimated base specular albedo.

9 Limitations and future works

Limitations. Our method relies on sparse landmarks for pose and geometry estimation. While this works well for several illumination scenarios, in extreme partial darkness (Figure 10, left), landmarks estimates and subsequent geometry estimation are less accurate. In several such cases, human landmarks estimation can also be incorrect, thus, a better approach to handle such cases is needed. Our method does not model external shadows (Fig 10, right), in that case our method could benefit from a method such as [ZBT*20]. Another limitation of our method is reliance on statistical albedo priors (Optimization, Stage I) that do not model certain skin tones. As a result, non-Caucasian albedos may not be estimated correctly. The unexplained diffuse albedo can get baked into the illumination, especially for darker skin tones, as shown in Figure 4, Subject 20.

We note that our albedos (esp. roughness) attributes are view and input image illumination condition dependent, however, when available, statistical priors help give meaningful estimates. Here, our method relies on symmetry, consistency and smoothness regularizers (Eq 6) to avoid overfitting. In some cases, due to these regularizers, person specific attributes are not captured. Additionally, while the consistency and symmetry regularizers (Stage-II) help avoid baking shadows in the final albedo, in some cases, when the optimized light and consequent shadows are inaccurate, some light/shadow patches may appear in the estimated albedos. Finally, the proposed light stage may not always recover accurate illumination for certain illumination conditions. For instance, because we model a single area light per icosahedron face, in case there are several light sources in one direction, the light stage may either favor the main light in this direction or an average of these lights.

Future Works. In the future, we want to extend our approach with methods such as [LBZ*20], to model mesoscopic geometric details, [YS*18]. Currently, we use single bounce rays for illumination modeling due to lack of external scene geometries, a natural extension is to model multi-ray bounces for softer shadows. Further, our methods naturally extends to a multi-view face reconstruction formulation that would help improve attribute estimation quality. Finally, we plan to extend our method with more complex skin reflectance models such as BSSRDF/dielectric materials, [WMP*06].

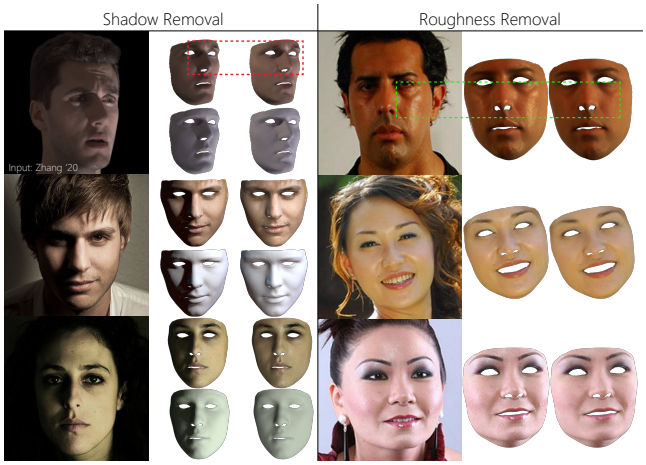

Figure 9: *Left: We show self-shadow editing removing resultant self-shadows (in red) by manipulating optimized illumination to uniform illumination. Right (Input, optimized, and edited specular highlights): By manipulating the optimized roughness map, specular reflections (in green) can be edited.*

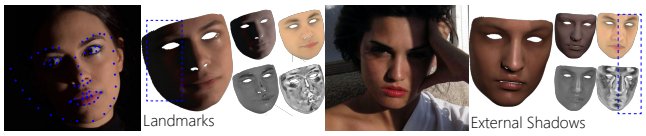

Figure 10: *Limitations – left: Imprecise landmarks under extreme scene illumination produces incorrect geometry reconstruction. Right: External shadows get baked into albedos.*

10 Conclusion

We present a novel and robust face modeling approach, under general illumination conditions. A virtual light stage formulation to model scene illumination is introduced, which, used in-conjunction with a differentiable ray tracing, makes our method self-shadows and specular reflectance aware. We then formulate face modeling as a loss minimization problem, and solve it via a two-stage optimization strategy. This strategy systematically disentangles face attributes, that make the optimization tractable for unconstrained input images. To validate our method, along with several results, we provide ablation studies, analysis of various modeling decisions and limitations. Beyond its accuracy and robustness to light conditions, the rich decomposition resulting from our approach allows for several style – illumination and albedo – transfer and edit applications.

Acknowledgements. We thank the anonymous reviewers for their feedback. A sincere thank you to Christel Chamaret and Lionel Oisel for their support.

References

[ARL*10] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG J.-Y., MA W.-C., WANG C.-C., DEBEVEC P.: The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications 30*, 4 (2010), 20–31.

[AS*12] ANGEL E., SHREINER D., ET AL.: *Interactive computer graphics: a top-down approach with shader-based OpenGL*. Boston: Addison-Wesley,, 2012.

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 75.

[BT17] BULAT A., TZIMIROPOULOS G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision* (2017).

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (1999), ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

[BWS*18] BAGAUTDINOV T., WU C., SARAGIH J., FUA P., SHEIKH Y.: Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3877–3886.

[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.

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG) 1*, 1 (1982), 7–24.

[DBA*19] DIB A., BHARAJ G., AHN J., THEBAULT C., GOSSELIN P.-H., CHEVALLIER L.: Face reflectance and geometry modeling via differentiable ray tracing. *ACM SIGGRAPH European Conference on Visual Media Production (CVMP)* (2019).

[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.

[Ekm97] EKMAN R.: *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[Emi17] EMILY: The wikihuman project. https://vgl.ict.usc.edu/Data/DigitalEmily2/, 2017. Accessed: 2020-05-21.

[EST*19] EGGER B., SMITH W. A. P., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., THEOBALT C., BLANZ V., VETTER T.: 3D Morphable Face Models – Past, Present and Future. *arXiv e-prints* (Sep 2019). arXiv:1909.01815.

[GCP*09] GHOSH A., CHEN T., PEERS P., WILSON C. A., DEBEVEC P.: Estimating specular roughness and anisotropy from second order spherical gradient illumination. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 1161–1170.

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 129.

[GFT*15] GRAHAM P., FYFFE G., TONWATTANAPONG B., GHOSH A., DEBEVEC P.: Near-instant capture of high-resolution facial geometry and reflectance. In *Proceedings of ACM SIGGRAPH 2015 Talks* (Aug. 2015), ACM Press, pp. 1–1. doi:10.1145/2775280.2792561.

[GGSC96] GORTLER S. J., GRZESZCZUK R., SZELISKI R., COHEN M. F.: The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (1996), pp. 43–54.

[GHP*08] GHOSH A., HAWKINS T., PEERS P., FREDERIKSEN S., DEBEVEC P.: Practical modeling and acquisition of layered facial reflectance. *ACM Transaction on Graphics 27*, 5 (Dec. 2008).

[GMFB*18] GERIG T., MOREL-FORSTER A., BLUMER C., EGGER B., LUTHI M., SCHÖNBORN S., VETTER T.: Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018), IEEE, pp. 75–82.

[GRB*18] GOTARDO P., RIVIERE J., BRADLEY D., ET AL.: *Practical Dynamic Facial Appearance Modeling and Acquisition*. ACM SIGGRAPH Asia, 2018.

[GVWT13] GARRIDO P., VALGAERT L., WU C., THEOBALT C.: Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics* (2013). doi:10.1145/2508363.2508380.

[GZC*16] GARRIDO P., ZOLLHÖFER M., CASAS D., VALGAERTS L., VARANASI K., PEREZ P., THEOBALT C.: Reconstruction of Personalized 3D Face Rigs from Monocular Video. *{ACM} Trans. Graph. (Presented at SIGGRAPH 2016) 35*, 3 (2016), 28:1–28:15.

[GZL18] GUO J., ZHU X., LEI Z.: 3ddfa. https://github.com/cleardusk/3DDFA, 2018.

[HCS*18] HUYNH L., CHEN W., SAITO S., XING J., NAGANO K., JONES A., DEBEVEC P., LI H.: Mesoscopic Facial Geometry Inference Using Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8407–8416.

[HSL01] HAKURA Z. S., SNYDER J. M., LENGYEL J. E.: Parameterized environment maps. In *Proceedings of the 2001 symposium on Interactive 3D graphics* (2001), pp. 203–208.

[Kaj86] KAJIYA J. T.: The rendering equation. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1986), SIGGRAPH '86, ACM, pp. 143–150. doi:10.1145/15922.15902.

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[KGT*18] KIM H., GARRIDO P., TEWARI A., XU W., THIES J., NIESSNER M., PÉREZ P., RICHARDT C., ZOLLÖFER M., THEOBALT C.: Deep video portraits. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 163.

[KSES14] KHOLGADE N., SIMON T., EFROS A., SHEIKH Y.: 3d object manipulation in a single photograph using stock 3d models. *ACM Transactions on Computer Graphics 33*, 4 (2014).

[LADL18] LI T.-M., AITTALA M., DURAND F., LEHTINEN J.: Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. 37*, 6 (Dec. 2018), 222:1–222:11. URL: http://doi.acm.org/10.1145/3272127.3275109, doi:10.1145/3272127.3275109.

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (ToG) 36*, 6 (2017), 194.

[LBZ*20] LI R., BLADIN K., ZHAO Y., CHINARA C., INGRAHAM O., XIANG P., REN X., PRASAD P., KISHORE B., XING J., ET AL.: Learning formation of physically-based face attributes. *arXiv preprint arXiv:2004.03458* (2020).

[LHJ19] LOUBET G., HOLZSCHUCH N., JAKOB W.: Reparameterizing discontinuous integrands for differentiable rendering. *Transactions on Graphics (Proceedings of SIGGRAPH Asia) 38*, 6 (Dec. 2019). doi:10.1145/3355089.3356510.

[LKA*17] LAINE S., KARRAS T., AILA T., HERVA A., SAITO S., YU R., LI H., LEHTINEN J.: Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2017), pp. 1–10.

[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". *arXiv preprint arXiv:2003.13845* (2020).

[LYL*16] LEGENDRE C., YU X., LIU D., BUSCH J., JONES A., PATTANAIK S., DEBEVEC P.: Practical multispectral lighting reproduction. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 32.

[LZL14] LI C., ZHOU K., LIN S.: Intrinsic face image decomposition with human face priors. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Springer International Publishing, pp. 218–233.

[NRH*92] NICODEMUS F. E., RICHMOND J. C., HSIA J. J., GINSBERG I., LIMPERIS T.: Geometrical considerations and nomenclature for reflectance. *NBS monograph 160* (1992), 4.

[NSX*18] NAGANO K., SEO J., XING J., WEI L., LI Z., SAITO S., AGARWAL A., FURSUND J., LI H.: pagan: real-time avatars using dynamic textures. In *SIGGRAPH Asia 2018 Technical Papers* (2018), ACM, p. 258.

[PGC*17] PASZKE A., GROSS S., CHINTALA S., CHANAN G., YANG E., DEVITO Z., LIN Z., DESMAISON A., ANTIGA L., LERER A.: Automatic differentiation in PyTorch.

[PJH16] PHARR M., JAKOB W., HUMPHREYS G.: *Physically based Rendering: From Theory to Implementation*. Morgan Kaufmann, 2016.

[PJY*19] PILLAI R. K., JENI L. A., YANG H., ZHANG Z., YIN L., COHN J. F.: The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *In Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops* (October 2019).

[RBSB18] RANJAN A., BOLKART T., SANYAL S., BLACK M. J.: Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 704–720.

[RGB*20] RIVIERE J., GOTARDO P. F. U., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *ACM Transactions on Graphics (TOG) 39* (2020), 81:1 – 81:12.

[RH01] RAMAMOORTHI R., HANRAHAN P.: On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. *JOSA A 18*, 10 (2001), 2448–2459.

[RHD*10] REINHARD E., HEIDRICH W., DEBEVEC P., PATTANAIK S., WARD G., MYSZKOWSKI K.: *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

[SBFB19] SANYAL S., BOLKART T., FENG H., BLACK M. J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 7763–7772.

[SBT*19] SUN T., BARRON J. T., TSAI Y.-T., XU Z., YU X., FYFFE G., RHEMANN C., BUSCH J., DEBEVEC P., RAMAMOORTHI R.: Single image portrait relighting. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* (2019).

[Sch94] SCHLICK C.: An inexpensive brdf model for physically-based rendering. *Computer Graphics Forum 13* (1994), 233–246.

[SHS*17] SHU Z., HADAP S., SHECHTMAN E., SUNKAVALLI K., PARIS S., SAMARAS D.: Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 1.

[SKCJ18] SENGUPTA S., KANAZAWA A., CASTILLO C. D., JACOBS D. W.: Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6296–6305.

[SKSS14] SUWAJANAKORN S., KEMELMACHER-SHLIZERMAN I., SEITZ S. M.: Total moving face reconstruction. In *European Conference on Computer Vision* (2014), Springer, pp. 796–812.

[SLC11] SARAGIH J. M., LUCEY S., COHN J. F.: Deformable model fitting by regularized landmark mean-shift. *International journal of computer vision 91*, 2 (2011), 200–215.

[SPB*14] SHIH Y., PARIS S., BARNES C., FREEMAN W. T., DURAND F.: Style transfer for headshot portraits. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 148.

[SSD*20] SMITH W. A., SECK A., DEE H., TIDDEMAN B., TENENBAUM J., EGGER B.: A morphable face albedo model. *arXiv preprint arXiv:2004.02711* (2020).

[SSKS17] SUWAJANAKORN S., SEITZ S. M., KEMELMACHER-SHLIZERMAN I.: Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 95.

[SWH*17] SAITO S., WEI L., HU L., NAGANO K., LI H.: Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5144–5153.

[SYH*17] SHU Z., YUMER E., HADAP S., SUNKAVALLI K., SHECHTMAN E., SAMARAS D.: Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5541–5550.

[T*98] TAKIWAKI H., ET AL.: Measurement of skin color: practical application and theoretical considerations. *Journal of Medical Investigation 44* (1998), 121–126.

[TBG*19] TEWARI A., BERNARD F., GARRIDO P., BHARAJ G., ET AL.: *FML: Face Model Learning from Videos*. CVPR, 2019.

[TL18] TRAN L., LIU X.: Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 7346–7355.

[TLL19] TRAN L., LIU F., LIU X.: Towards high-fidelity nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition* (Long Beach, CA, June 2019).

[TZK*17] TEWARI A., ZOLLÖFER M., KIM H., GARRIDO P., BERNARD F., PEREZ P., CHRISTIAN T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)* (2017).

[VWB*12] VALGAERTS L., WU C., BRUHN A., SEIDEL H.-P., THEOBALT C.: Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph. 31*, 6 (2012), 187–1.

[WMLT07] WALTER B., MARSCHNER S., LI H., TORRANCE K.: Microfacet models for refraction through rough surfaces. pp. 195–206. doi:10.2312/EGWR/EGSR07/195-206.

[WMP*06] WEYRICH T., MATUSIK W., PFISTER H., BICKEL B., DONNER C., TU C., MCANDLESS J., LEE J., NGAN A., JENSEN H. W., OTHERS: Analysis of human faces using a measurement-based skin reflectance model. In *ACM Transactions on Graphics (TOG)* (2006), vol. 25, ACM, pp. 1013–1024.

[WVL*11] WU C., VARANASI K., LIU Y., SEIDEL H.-P., THEOBALT C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In *2011 International Conference on Computer Vision* (2011), IEEE, pp. 1108–1115.

[YS*18] YAMAGUCHI S., SAITO S., ET AL.: *High-fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image*. ACM TOG, 2018.

[ZBT*20] ZHANG X., BARRON J. T., TSAI Y.-T., ZHANG X., NG R., JACOBS D. E.: Portrait shadow manipulation. vol. 39.

[ZTB*18] ZOLLHÖFER M., THIES J., BRADLEY D., GARRIDO P., BEELER T., PÉREZ P., STAMMINGER M., NIESSNER M., THEOBALT C.: State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications.