Single-Shot High-Quality Facial Geometry and Skin Appearance Capture

JÉRÉMY RIVIERE, PAULO GOTARDO, and DEREK BRADLEY, DisneyResearch|Studios
ABHIJEET GHOSH, Imperial College London
THABO BEELER, DisneyResearch|Studios

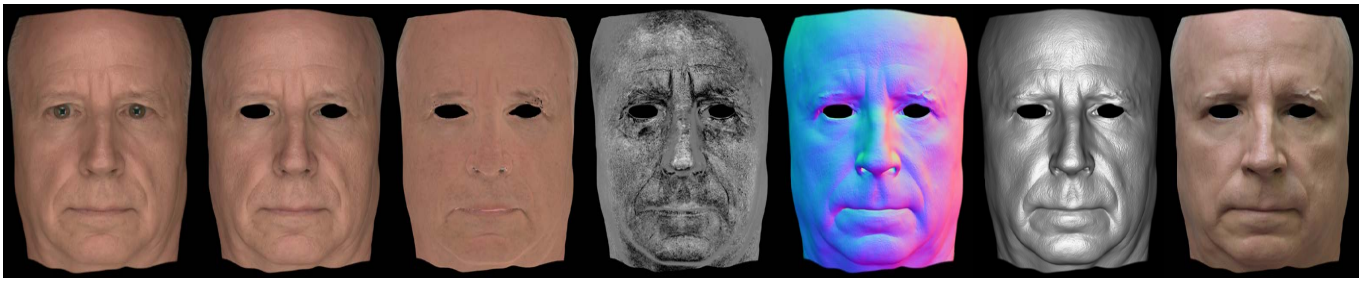


Fig. 1. We present a method to capture complete facial geometry and appearance from a *single* exposure. From left to right: one input image, our matching render, diffuse albedo, specular intensity, normals, high resolution geometry, and a realistic re-render under a different environment map.

We propose a new light-weight face capture system capable of reconstructing both *high-quality geometry and detailed appearance maps* from a *single exposure*. Unlike currently employed appearance acquisition systems, the proposed technology does not require active illumination and hence can readily be integrated with passive photogrammetry solutions. These solutions are in widespread use for 3D scanning humans as they can be assembled from off-the-shelf hardware components, but lack the capability of estimating appearance. This paper proposes a solution to overcome this limitation, by adding appearance capture to photogrammetry systems. The only additional hardware requirement to these solutions is that a subset of the cameras are cross-polarized with respect to the illumination, and the remaining cameras are parallel-polarized. The proposed algorithm leverages the images with the two different polarization states to reconstruct the geometry and to recover appearance properties. We do so by means of an inverse rendering framework, which solves *per texel diffuse albedo, specular intensity, and high-resolution normals, as well as global specular roughness* considering the subsurface scattering nature of skin. We show results for a variety of human subjects of different ages and skin typology, illustrating how the captured fine-detail skin surface and subsurface scattering effects lead to realistic renderings of their digital doubles, also in different illumination conditions.

CCS Concepts: • **Computing methodologies → Reflectance modeling**; **3D imaging**; *Appearance and texture representations.*

Additional Key Words and Phrases: Passive Photogrammetry, Dynamic Face Capture, Appearance Capture, High-Detail Surface, Inverse Rendering

Authors' addresses: Jérémy Riviere, jeremy.riviere@disneyresearch.com; Paulo Gotardo, paulo.gotardo@disneyresearch.com; Derek Bradley, derek.bradley@disneyresearch.com, DisneyResearch|Studios; Abhijeet Ghosh, ghosh@imperial.ac.uk, Imperial College London; Thabo Beeler, thabo.beeler@gmail.com, DisneyResearch|Studios.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2020/7-ART81 $15.00
https://doi.org/10.1145/3386569.3392464

ACM Reference Format:
Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-Shot High-Quality Facial Geometry and Skin Appearance Capture. *ACM Trans. Graph.* 39, 4, Article 81 (July 2020), 12 pages. https://doi.org/10.1145/3386569.3392464

1 INTRODUCTION

Digital humans have become omnipresent in today's entertainment landscape, making an appearance in nearly every blockbuster movie and triple-A video game. To create such digital characters it is common practice to 3D scan real humans, digitally capturing their likeness. To accomplish this, passive photogrammetry solutions have become the method of choice for two reasons. Firstly, passive photogrammetry systems can be constructed from off-the-shelf consumer hardware, such as digital cameras and flashes, and are hence much less complex and more cost effective than active technologies, such as structured light scanning or lightstage acquisition. Secondly, a number of software solutions exist, both commercial and open-source, that allow to reconstruct high-quality 3D geometry from the acquired images. This makes 3D shape acquisition readily and widely available.

Standard photogrammetry alone, however, is not sufficient to create photorealistic digital human assets. In addition to 3D shape, high-quality diffuse and specular reflectance properties are also required for realistic rendering. Furthermore, the level of geometric detail provided by photogrammetry has been typically inferior when compared to 3D shapes obtained with more complex setups based on active lighting, such as lightstages [Debevec et al. 2000] and other recent videogrammetry solutions [Gotardo et al. 2018]. Thus, to acquire these appearance properties and fine-detail geometry, studios and digital artists are currently forced to employ costly and complex setups that require expert knowledge to build and operate. As a result, high-quality appearance acquisition is currently only viable for hero assets in high-budget productions.

In this paper, we propose the first light-weight, inexpensive, single-shot acquisition system that can capture both high-quality

facial appearance and 3D geometry. The proposed method can be readily integrated with current widely employed photogrammetry setups, requiring only minimal hardware changes. Our key result is to effectively upgrade these widespread setups into *one-stop-shop* acquisition systems for high-quality face capture.

More specifically, the captured face data consists of a single exposure that is simultaneously captured by multiple, conventional cameras located around the actor's face. The main requirement is that a subset of such cameras be cross-polarized with respect to incoming light, which provides our inverse rendering algorithms with an effective means to separate surface and subsurface appearance parameters. The remainder of the cameras are parallel-polarized, allowing to sample direct reflectance information like specular highlights. This single-shot mixture of cross and parallel-cross polarization acts like a form of view multiplexing, where our method can gather different information from different viewpoints and combine it in a single inverse rendering optimization. The output texture maps provided by our technique include fine-detail 3D geometry (displacement map), diffuse RGB albedo, specular intensity, overall specular roughness.

The proposed method allows to further democratize the creation of digital human assets, making high-quality 3D shape and appearance acquisition affordable for lower budget productions, realistic digital avatar creation in mixed reality applications, and also more appealing to fields outside the entertainment industry, such as academia, psychology, or life sciences.

2 RELATED WORK

We restrict the discussion specifically to related work on facial geometry and appearance capture and refer to recent surveys on the topic [Klehm et al. 2015; Weyrich et al. 2009] for a more in-depth discussion. In the following, we discuss facial capture in computer graphics in the context of both active and passive capture setups, employing multi vs single-shot capture, targeting static and dynamic facial appearance acquisition.

Active capture of static appearance: Active illumination based techniques have long been the methods of choice for high-quality facial capture. Debevec et al. [2000] introduced a specialized light stage setup to acquire a dense reflectance field of a human face for photo-realistic image-based relighting. They also employed the acquired data to estimate a few view-dependent reflectance maps that could be interpolated for viewpoint animation in conjunction with structured light scanning of facial geometry. Weyrich et al. [2006] employed an LED sphere with 150 lights and 16 cameras to densely record facial reflectance and computed view-independent estimates of facial reflectance from the acquired data including per-pixel diffuse and specular albedos, and per-region specular roughness parameters. They also employed a specialized skin contact probe to estimate a skin translucency parameter based on dipole diffusion [Jensen et al. 2001]. The facial geometry was acquired using two commercial 3D structured light scanners in their setup. These initial works, while very influential, involved a significant amount of data capture to acquire a face including its appearance. Hence, more recent works have focused on reducing the amount of acquired data for high-quality face scans. Ma et al. [2007] introduced polarized spherical gradient illumination for acquisition of the separated diffuse and specular albedos and photometric normals using just eight photographs, and reconstructed high quality facial geometry including skin mesostructure as well as realistic rendering with hybrid normal mapping. Ghosh et al. [2008] further extended the acquisition method to practically acquire layered facial reflectance within a capture budget of 20 photographs (burst mode of DSLR), using a combination of polarization and structured lighting. The acquisition method of Ma et al. was however restricted to just a frontal stereo pair of cameras due to the view-dependent polarization of the LED sphere employed for diffuse-specular separation. This was later extended to multi-view capture with polarized gradient illumination by Ghosh et al. [2011]. They employed two orthogonal polarization patterns (lines of latitude and longitude) on the LED sphere, allowing fast capture (due to static polarizers on the cameras) and separation of diffuse and specular reflectance from multiple viewpoints around the equator of the LED sphere. Graham et al. [2013] further extended the technique to acquire facial microgeometry of 1 cm^2 skin patches. They employed constrained texture synthesis to then add microscale details to underlying skin meso-structure and also fit micro-scale skin BRDF for increased realism of skin rendering. Recently, Kampouris et al. [2018] have proposed employing binary spherical gradient illumination in conjunction with color-space analysis for efficient acquisition of separated diffuse and specular reflectance and photometric normals. While enabling faster acquisition than polarized spherical gradients (half the number of photographs), the method still requires active illumination using an LED sphere for acquisition. Closer to our approach, Fyffe et al. [2016] have proposed a solution for static facial capture that employs consumer hardware for near-instant capture of facial geometry and reflectance. Their setup uses a combination of 24 DSLR cameras and 6 flashes that are triggered in sequence within a few milliseconds. However, the approach does not extend to dynamic facial appearance capture due to active triggering mechanism of the cameras and flash units. In comparison, our method relies on a simpler more practical single-shot capture setup while estimating high quality facial reflectance including spatially varying specular albedo and an improved diffuse albedo accounting for subsurface scattering. Moreover, our method can be flexibly applied to dynamic facial capture.

Active capture of dynamic appearance: Hawkins et al. [2004] extended the approach of [Debevec et al. 2000] to acquire dynamic facial reflectance fields of a set of key facial poses, and interpolated between the reflectance fields of these key poses at run-time for synthesizing relightable facial animations. Wenger et al. [2005] employed an LED sphere and high speed photography to acquire the response to a dense set of illumination conditions in order to relight each frame of a target facial performance. They also employed the data to estimate photometric surface normals, and diffuse and specular albedos for relighting of the facial performance. These above techniques relied on dense capture of dynamic facial reflectance which can be impractical. To reduce the acquired data, Ma et al. [2008] employed spherical gradient illumination in conjunction with high speed acquisition to capture short sequences

of facial performances (formation of expressions). They employed the acquired facial displacement maps (extracted from photometric normal) in conjunction with marker-based correspondences to fit polynomial functions as a way of encoding facial mesostructure dynamics during a performance. Fyffe et al. [2011] instead applied complementary spherical gradient illumination based alignment of Wilson et al. [2010] in conjunction with high speed photography to acquire longer facial performance sequences, and employed a heuristics based diffuse-specular separation on the acquired data to obtain albedo and normal maps for rendering. Nagano et al. [2015] have extended [Graham et al. 2013] to acquire microgeometry of various skin patches under stretch and compression and employed the acquired data for efficient real-time rendering of dynamic facial microgeometry using texture space filtering. For true video-rate dynamic capture, Fyffe & Debevec [2015] have proposed employing spectral multiplexing with polarized spherical gradient illumination (using an RGB LED sphere) for facial performance capture. This has similarities with our setup in requiring multiple polarized cameras per acquisition viewpoint, but requires a much more complicated lighting and capture hardware while acquiring a spectrally saturated diffuse albedo due to RGB illumination. Gotardo et al. [2015] have proposed a simpler binocular setup with spectral and temporal multiplexing of nine light sources to compute dynamic diffuse albedo and normal maps. They however do not estimate any specular reflectance. Most recently, Meka et al. [2019] have proposed efficient dynamic performance relighting using capture with RGB-multiplexed unpolarized spherical gradient illumination and complement pairs which are then used as input to a convolutional deep network to predict relit facial performance under a novel lighting. This however requires a database of acquired facial reflectance fields in multiple expressions to train the deep network, and does not result in reflectance maps or geometry that can be used in a standard rendering pipeline.

Passive capture of facial geometry and texture: With advancement in photogrammetry techniques, passive facial acquisition has become a popular alternative to active capture techniques that require specialized acquisition setups. Besides simplifying static facial capture, such acquisition is particularly well suited for dynamic facial performance capture without requiring high frame rate acquisition and synchronization. A popular approach has been multi-view facial capture under uniform passive illumination [Beeler et al. 2010; Bradley et al. 2010], with such a capture providing estimate of an albedo texture under flat lit illumination for rendering purposes besides facial geometry reconstruction based on multi-view stereo. Beeler et al. [2010] further augmented the reconstructed facial base geometry with mesostructure detail extracted from the albedo texture using a high-pass filter. They later extended the approach for reconstructing facial performances with drift-free tracking over long sequences using anchor frames [Beeler et al. 2011]. The method produces very good qualitative results for facial geometry. However, the estimated albedo is not completely diffuse and contains a small amount of baked-in specular reflectance. The approach has been extended to static facial geometry and performance capture with simpler binocular [Valgaerts et al. 2012] and monocular setups [Cao et al. 2015; Garrido et al. 2013; Ichim et al. 2015; Shi et al. 2014] in uniform, uncontrolled illumination settings including indoor and outdoor environments. These methods assume that skin reflectance is Lambertian, and employ low-frequency lighting estimation with spherical harmonics for geometric refinement. In addition, they strongly rely on facial geometry priors (e.g. blendshape models) and although shading-based geometry refinement reveals facial wrinkles at larger scales, they cannot resolve fine scale detail. Importantly, these method do not provide a high quality estimate of facial reflectance.

Passive facial capture with reflectance: Fyffe et al. [2014] employed a database of high quality facial scans (acquired using the method of [Ghosh et al. 2011]) to augment a monocular video sequence of a facial performance acquired under passive illumination with high resolution facial geometry and reflectance maps for realistic rendering. The approach achieves impressive qualitative results but requires a dense set of facial scans (in different poses) with reflectance information of the same target subject. Saito et al. [2017] first proposed a deep learning approach for data-driven inference of high resolution facial texture map of an entire face for realistic rendering from an input of a single low resolution face image with partial facial coverage. They further extended this to inference of facial mesostructure given a diffuse albedo texture [Huynh et al. 2018], and complete facial reflectance and displacement maps besides albedo texture given partial facial image as input [Yamaguchi et al. 2018]. These approaches focus on simple creation of a believable digital avatar rather than accurate reconstruction of facial appearance, and rely on a facial database acquired with polarized spherical gradients for training the deep network for super-resolution and augmentation tasks. Closest to our work is that of Gotardo et al. [2018] who employ a 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. Their method however requires an initialization of the reflectance estimate using a video-sequence capture for a neutral expression where the subject has to rotate their face in various directions. In contrast, our capture solution is truly single-shot for geometry and reflectance estimation and can be naturally extended to dynamic capture without requiring cumbersome subject motion for any initialization. Furthermore, thanks to our approach of view-multiplexing with cross and non-cross polarization, our method is capable of better separating diffuse and specular components, leading naturally to better surface normals. Additionally our approach improves sharpness of the diffuse albedo for rendering by accounting for subsurface scattering in skin, and we improve reflectance estimation by optimizing for the per-subject specular lobe. In this respect, our proposed single-shot capture method is the first to target the quality of reflectance usually acquired by state-of-the-art active illumination systems.

3 BASE GEOMETRY CAPTURE AND PRE-PROCESSING

This section describes the simple capture setup we used to acquire facial images, followed by the initial geometry reconstruction pipeline, and the preprocessing steps that are performed prior to the main process of simultaneous appearance estimation and geometry refinement, described in Section 4.

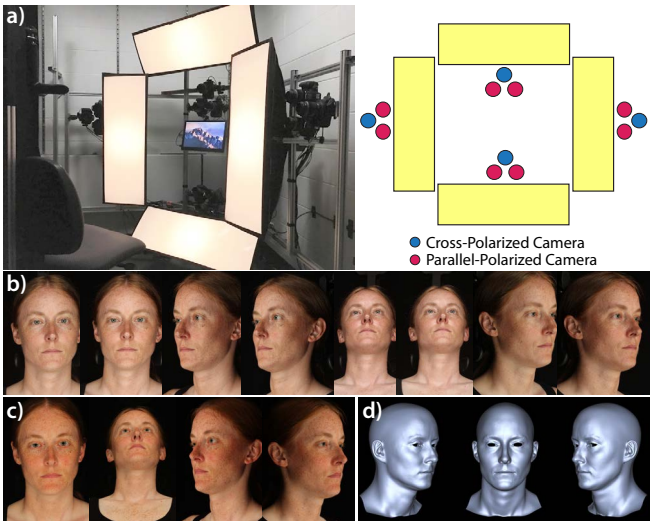


Fig. 2. Our simple capture setup consists of 12 DSLR cameras and 4 flashes, all with linear polarization filters (a). Of these, 8 cameras (4 stereo pairs) have parallel polarization (b), and 4 have cross-polarization (c). We reconstruct the base geometry from the 4 stereo pairs (d) and use all 12 cameras for appearance estimation and geometry refinement.

3.1 Capture Setup

Our design goal is to operate with a simple configuration of inexpensive consumer-level hardware, making our approach widely applicable in both high-end and low-budget scenarios alike. Our single-shot capture setup consists of 4 studio flashes and 12 DSLR cameras arranged as 4 triplets (see Fig. 2). We apply linear polarizing filters to both flashes (horizontal) and cameras (horizontal or vertical). Each camera triplet consist of (i) a narrow baseline stereo pair with polarization filters parallel to the flash filters, and (ii) a central camera that is cross-polarized with respect to the lights, a common approach for canceling specular reflection [Ghosh et al. 2011]. We therefore dedicate a larger number of (parallel-polarized) cameras to capture view-dependent specular signal and full surface detail, which we have found to provide for better stereo matching and triangulation under soft illumination. Figure 2 shows this camera configuration and the full set of 12 images recorded for one shot of a target subject. This particular setup consists of Canon Rebel T5 cameras with EF-S 60mm f/2.8 lenses and Elinchrom D-Lite RX 4 flashes with soft-boxes, leading to a total hardware cost of below $10K for the entire setup. However, we note that the system is not limited to operate with such specific choice of hardware and can directly accommodate cheaper or more expensive models. In Section 5 we also show results from an alternative setup, for dynamic facial performance capture, highlighting the flexibility of our method.

3.2 Initial Geometry Reconstruction

Facial geometry is reconstructed using the 8 parallel-polarized cameras arranged as 4 stereo pairs. We use the stereo method of Beeler et al. [2010] but without mesoscopic augmentation, as we perform our own geometry refinement to capture fine-scale surface details, as a byproduct of appearance estimation. This geometry refinement step outputs a displacement map defined in UV texture space; the UV parameterization also makes it easier to pool together data from the different cameras for appearance estimation. We therefore require the initial 3D face mesh to be texture mapped. This step can be achieved via automatic parameterization methods, but we opted to manually fit the triangulated raw geometry using a template face mesh with a well-formed topology, which is a typical step when creating digital human assets. An example 3D face mesh, after fitting to the raw geometry, is shown in Fig. 2 (d).

3.3 Input Texture Maps

As both the input and output of our main appearance estimation step consist of data in the form of UV texture-space maps, we also compute the following input maps in a preprocessing step, similar to [Gotardo et al. 2018]: (i) per-camera texture maps containing the input image data, also encoding the per-camera visibility maps; (ii) per-camera weights that downweight less reliable data due to high foreshortening of camera view and due to defocus from shallow depth of field; and (iii) initial geometry maps comprising of an initial normal map, an initial vertex map (low-frequency geometry of the fitted template mesh), and an initial "macro" displacement map from multiview stereo (with mid-frequency geometry to be refined during appearance estimation).

3.4 Additional Calibration

We also precompute an environment map encoding the spatial distribution of incoming light from our 4 flash panels. Following standard practice, we capture an HDR image of a mirror sphere with known radius, using a frontal camera without polarizer. Given calibrated cameras, we triangulate the 3D position of the sphere, shoot rays through the pixels of this HDRI, which are then reflected on the sphere to yield the environment map [Gotardo et al. 2018]. We also use the initial face geometry to ray-trace shadow maps to be used during inverse rendering. Finally, we use a standard Macbeth color chart to color calibrate the frontal, cross-polarized camera to serve as reference color space at normalized exposure; all other cameras are automatically self-calibrated to this color space, as described in Section 4.2.

4 HIGH-QUALITY APPEARANCE AND GEOMETRY

Given the base geometry of the precomputed template face mesh, the next step is to simultaneously compute high-quality appearance and fine-detail geometry maps from the single-shot input. This section describes our forward rendering models, photometric calibration of the input data, and estimation method via inverse rendering.

4.1 Appearance and Geometry Models

We model the reflectance of facial skin using a spatially varying, bidirectional scattering-surface reflectance distribution function (BSSRDF) [Pharr et al. 2016]. Let $\mathbf{x}_i$ denote the position of a surface patch with normal $\mathbf{n}_i$, and $L_i(\mathbf{x}_i, \omega_i)$ be the incident light from direction ω_i. The proportion of light that radiates out of a nearby position $\mathbf{x}_o$ along the (view) direction ω_o is given by our discretized

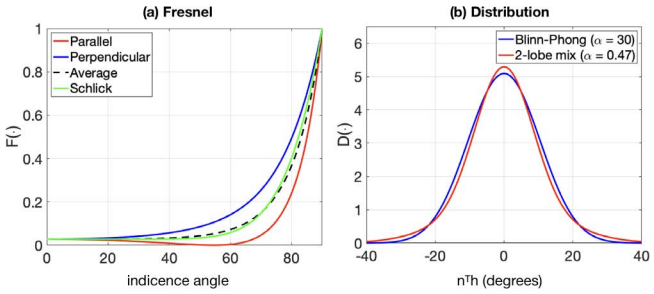


Fig. 3. **Surface reflection** – (a) for parallel-polarized views, specular reflection vanishes near Brewster's angle of incidence, which must be accounted for in inverse rendering for accurate results; (b) our 2-lobe microfacet distribution has stronger tails than the equivalent Blinn-Phong lobe (or Beckmann with roughness $\alpha = 0.27$) to better render skin [Walter et al. 2007].

rendering equation,

$$L_o(\mathbf{x}_o, \omega_o) = \sum_A \sum_\Omega S(\mathbf{x}_o, \omega_o, \mathbf{x}_i, \omega_i) L_i(\mathbf{x}_i, \omega_i)(\mathbf{n}_i^\top \omega_i) \Delta\omega_i \Delta A_i \quad (1)$$

for unoccluded directions $\omega_i \in \Omega$ with solid angle $\Delta\omega_i$ and nearby patches $\mathbf{x}_i \in A$ with area ΔA_i. Our BSSRDF model comprises two components that define surface (specular) and subsurface (diffuse) contributions to reflected light,

$$S(\mathbf{x}_o, \omega_o, \mathbf{x}_i, \omega_i) = \delta_{ij} S_r(\mathbf{x}_o, \omega_o, \omega_i) + S_d(\mathbf{x}_o, \omega_o, \mathbf{x}_i, \omega_i), \quad (2)$$

where the Kronecker delta $\delta_{ij} = 1$ if $\mathbf{x}_i = \mathbf{x}_o$. When rendering cross-polarized views, we model surface reflectance as $S_r(\mathbf{x}_o, \omega_o, \omega_i) = 0$. For parallel-polarized views, we use the Cook-Torrance BRDF

$$S_r(\mathbf{x}_o, \omega_o, \omega_i) = \rho_s(\mathbf{x}_o) \frac{D(\omega_o, \omega_i, \mathbf{n_o}, \alpha) G(\omega_o, \omega_i) F(\eta, \omega_o, \omega_i)}{4(\mathbf{n}_o^\top \omega_i)(\mathbf{n}_o^\top \omega_o)}, \quad (3)$$

which is modulated by the spatially varying specular intensity parameter ρ_s that captures variability in skin reflectance due to, for example, surface oiliness. The standard geometry attenuation term is given by G, and F denotes the Fresnel curve. The index of refraction for skin is fixed at $\eta = 1.4$. Instead of the typically used Fresnel curve for unpolarized light, we use the Fresnel curve for *parallel-polarized* light, Fig. 3(a). This is somewhat important in our setup given the use of horizontal polarizers on the lightboxes, which results in predominately parallel polarized reflection on the face along the equatorial directions. The distribution term $D(\cdot) = \alpha D_{12}(\cdot) + (1 - \alpha) D_{48}(\cdot)$ is a linear combination of two Blinn-Phong basis lobes with exponents 12 and 48, Fig. 3(b); it provides slightly stronger tails for more realistic face rendering and enables the estimation of the lobe size (roughness) α via a linear fit (Section 4.2).

The diffuse reflection term accounts for subsurface scattering and absorption of light for the given color channel wavelength λ,

$$S_d(\cdot) = \frac{1}{\pi} F_t(\mathbf{x_o}, \omega_o) \rho_\lambda(\mathbf{x}_o) R_\lambda(\|\mathbf{x}_o - \mathbf{x}_i\|_2) \rho_\lambda(\mathbf{x}_i) F_t(\mathbf{x}_i, \omega_i), \quad (4)$$

where F_t is the Fresnel transmittance, ρ_λ is the (red, green, or blue) spatially-varying albedo, and $R_\lambda(r)$ is the sum-of-Gaussians diffusion profile proposed by [d'Eon et al. 2007]. In our experiments, we fix the per-channel Gaussian weights as computed to approximate the three-layer skin model of [Donner and Jensen 2005].

Typically, inverse rendering approaches that do not take into account subsurface scattering yield blurry normal and albedo estimates with attenuated high-frequency detail. To improve the level of recovered surface detail, our approach focuses on data from surface (specular) reflectance. Since specular reflection maintains light polarization, our parallel-polarized cameras filter out half of the diffuse reflection and effectively increase the specular-to-diffuse reflection ratio. However, our single-shot approach observes the skin only under a single illumination condition, and thus the specular signal alone may not be enough to fully disambiguate normal estimation. For this reason, we leverage the fact that subsurface scattering is significantly lower in the blue image channel and estimate fine-scale detail using predominantly specular and blue-diffuse constraints (see Section 4.3). For this reason, we use diffusion profiles that are relative to the typical diffusion observed for a blue wavelength [d'Eon et al. 2007].

To further constrain the estimation of the normals, we directly enforce integrability (zero curl) as a hard constraint in our geometry model. A similar idea was explored by [Gotardo et al. 2018], but their method required a final post-processing step [Nehab et al. 2005] to compute an actual displacement map from their estimated normal field, which typically causes some loss of geometric detail.

We thus parameterize our refined normal field directly in terms of a displacement map $d(u, v)$, with one surface patch per texel in UV texture space. This displacement map is optimized for from the outset and can be trivially applied to emboss fine-detail geometry onto our initially fitted template face mesh. Given the input vertex and normal maps of the template mesh, let $\hat{\mathbf{n}}$, $\hat{\mathbf{t}}_u$, and $\hat{\mathbf{t}}_v$ denote a texel's unit normal and tangent vectors (computed by simple finite differences). Also, let $\hat{s}_u$ and $\hat{s}_v$ be the original lengths of the tangent vectors, encoding texel size. Then, after applying the desired high-detail displacement map $d(u, v)$, the non-unit normal of the new, refined mesh can be expressed from the new, non-unit tangents as

$$\mathbf{n} = (\hat{s}_u \hat{\mathbf{t}}_u + d_u \hat{\mathbf{n}}) \times (\hat{s}_v \hat{\mathbf{t}}_v + d_v \hat{\mathbf{n}}) \quad (5)$$

$$= \begin{bmatrix} \hat{\mathbf{t}}_u & \hat{\mathbf{t}}_v & \hat{\mathbf{n}} \end{bmatrix} \begin{bmatrix} \hat{s}_u & 0 & 0 \\ 0 & \hat{s}_v & 0 \\ 0 & 0 & \hat{s}_u \hat{s}_v \end{bmatrix} \begin{bmatrix} -d_u \\ -d_v \\ 1 \end{bmatrix}, \quad (6)$$

where d_u and d_v are the partial derivatives of $d(u, v)$ computed via finite differencing. The simple form in Eq. 5 is achieved by leveraging the fact that a triangle in the initial template mesh spans multiple texels in its normal map, resulting in locally constant $\hat{\mathbf{n}}$. Another key difference to [Gotardo et al. 2018] is in properly accounting for texel size, which improves scaling of constraints and allows for optimization in a coarse-to-fine, multi-resolution manner for better convergence.

4.2 Photometric Self Calibration

Before computing appearance and geometry refinement, we must account for the differences in color space (exposure, black level) and polarization filter attenuation amongst all cross- and parallel-polarized cameras. Often, color calibration using a standard color chart can be mislead by specular reflection when both view and light directions are at an oblique angle. It can also be difficult and laborious to properly measure per-camera attenuation of image

intensity due to the use of polarization filters. Therefore, to facilitate the use of our system, we introduce an automatic self calibration procedure that uses the captured face itself along with renderings of our model as the calibration target.

Our technique requires that only one camera be color calibrated towards a color chart, to provide a reference color space that will be matched by all cameras. We take as reference the cross-polarized camera in the frontal stereo pair. Then, each of the other three cross-polarized cameras is automatically calibrated to match the colors of the frontal one, by estimating a 3×4 affine color matrix in a least-squares sense.

The other 8 cameras are parallel polarized and exhibit strongly view-dependent specular reflection. To color calibrate each of these cameras, we compute an initial rendering of our appearance model and choose it as the calibration target, to ensure that each camera agrees with the model as closely as possible. More specifically, given the initial geometry of our template face mesh, for each camera c we render two specular reflection images, $S_{c_1}(\mathbf{x})$ and $S_{c_2}(\mathbf{x})$, one for each of the specular basis lobes in our BRDF model. The diffuse term $I_{xp}(\mathbf{x})$ is the image of the closest cross-polarized camera.

The self-calibration procedure of each parallel-polarized camera image $I_c(\mathbf{x})$ estimates the camera color matrix $\mathbf{M}_c$ satisfying

$$\mathbf{M}_c \begin{bmatrix} I_c(\mathbf{x}) \\ 1 \end{bmatrix} \approx \begin{bmatrix} S_{c_1}(\mathbf{x}) & S_{c_2}(\mathbf{x}) & I_{xp}(\mathbf{x}) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ 1 \end{bmatrix}, \quad \forall c, \forall \mathbf{x}\,. \tag{7}$$

The specular weights $w_1 > 0$ and $w_2 > 0$ are related to our BRDF parameters in *Eq.* 3. That is, specular intensity $\rho_s = w_1 + w_2$ and specular lobe size $\alpha = w_1/(w_1 + w_2)$. Since these weights are initially unknown for the given face, self-calibration estimates them (globally) in addition to the 8 matrices $\mathbf{M}_c$ via alternated least squares.

We initialize w_1 and w_2 using known measurements of facial skin reflectance [Weyrich et al. 2006] and observe convergence within 10 iterations. As a by-product of self calibration, we compute the initial global estimates ρ_{s0} and α_0 that can be used to regularize per-texel estimates in the subsequent, main inverse rendering step.

4.3 Inverse Rendering

This main step for computing appearance and refined geometry maps operates in UV texture-space using the precomputed geometry maps and the self-calibrated textures alongside the captured image data. The output is a multi-channel map $\Theta(\mathbf{x}_u, \mathbf{x}_v)$ encoding per-texel RGB albedo, specular intensity and lobe size, and a fine-detail displacement map. For each texel $\mathbf{x}$, these parameters are encoded in the vector of unknowns $\Theta_{\mathbf{x}} = \{\rho_r, \rho_g, \rho_b, \rho_s, \alpha, d\} \in \mathbb{R}^6$. Due to the soft nature of the lighting in our setup (as well as in most photogrammetry stages), estimating per-texel specular lobe sizes is an ill-posed problem [Ghosh et al. 2008]. We therefore fix $\alpha = \alpha_0$ as estimated during self calibration. Still, spatial variation in skin roughness (e.g. due to skin stretch) are partially captured in the computed specular intensity and displacement maps.

To compute the optimal parameter map Θ, we implemented an auto-differentiable renderer that seeks to match the input image data $I_c(\mathbf{x})$ of all 12 cameras c as closely as possible. This is done by minimizing the energy (loss) term,

$$E_{img}(\Theta) = \sum_{\mathbf{x}} \sum_{c} W_c(\mathbf{x}) \left\| I_c(\mathbf{x}) - L_o(\mathbf{x}, \omega_c) \right\|_2^2, \tag{8}$$

where the rendered texel colors $L_o(\cdot)$ are given by our BSSRDF model in Eq. 1. The precomputed per-camera weight maps $W_c(\mathbf{x})$ provide a measure of confidence in the data due to defocus and view foreshortening. By themselves, the data terms in Eq. 8 may not be sufficient to completely constrain all parameters of all texels. We thus introduce additional regularization constraints that are needed to disambiguate parameter estimation in small regions of the face. The overall energy term minimized during inverse rendering is

$$\begin{aligned} \min_{\Theta} E_{img}(\Theta) + \lambda_1 \left\| d - d_0 \right\|_F^2 + \lambda_2 \left\| \nabla d \right\|_F^2 \\ + \lambda_3 \left\| \rho_s - \rho_{s0} \right\|_F^2 + \lambda_4 \left\| \nabla \rho_s \right\|_F^2 . \end{aligned} \tag{9}$$

The refined displacement map is weakly constrained to be close to the initial one, $d_0(u, v)$, as it only updates high-frequency geometry components of the template face mesh ($\lambda_1 = 0.03$). A small 3×3 Laplacian operator (∇) is also applied to ensure smoothness in underconstrained regions ($\lambda_2 = 0.02$). Similarly, we regularize specular intensity towards the global, self-calibrated value ρ_{s0} in underconstrained areas where specular reflection is very weak ($\lambda_3 = 0.03$). These areas often include the extreme sides of the face (when there is no illumination from behind), underneath the jaw and in concave regions (multiple indirect bounces of light are not accounted for). We initially apply a strong Laplacian operator to smooth the specular intensity map ($\lambda_4 = 0.3$), which forces fine-detail surface geometry to be explained mostly by the displacement map. To improve convergence and avoid loss of geometric detail, we account for the wavelength dependence of subsurface scattering and apply different weights for each color channel of the energy in Eq. 8, $w_R = 0.1, w_G = 0.3, w_B = 1.0$. Upon convergence, we fix the displacement map and continue optimization with uniform channel weights and disabled Laplacians. This final step fits sharp albedo and allows specular intensity to also model sharp specular reflection occlusion effects that were not explained by the optimized geometry.

To compute displacement maps with stronger mid-frequencies (e.g. deeper skin wrinkles and creases, larger moles), appearance and geometry optimization are computed in a coarse-to-fine strategy, with results first computed at lower resolutions and then used to initialize optimization at higher resolutions. Typically, optimization begins with $2K \times 2K$ and ends with 4K or 8K maps, upsampling with a factor of 2x. We employ the non-linear Ceres solver [Agarwal et al. 2016] to optimize for Θ. Convergence to poor local minima has not been observed, but surface detail can be less sharp with insufficient iterations. Solving at 4K resolution takes approximately 45 minutes on an 8-core 2019 MacBook Pro laptop. The coarse-to-fine solver also reduces runtime by better initializing high-resolution levels.

5 EXPERIMENTAL RESULTS

In this section, we assess the quality of the results achieved by our method by showing the high quality of reconstruction of both geometry and appearance across subjects of different ethnicities, ages and genders, while they performed different facial expressions. We also illustrate how the captured fine-detail skin surface and

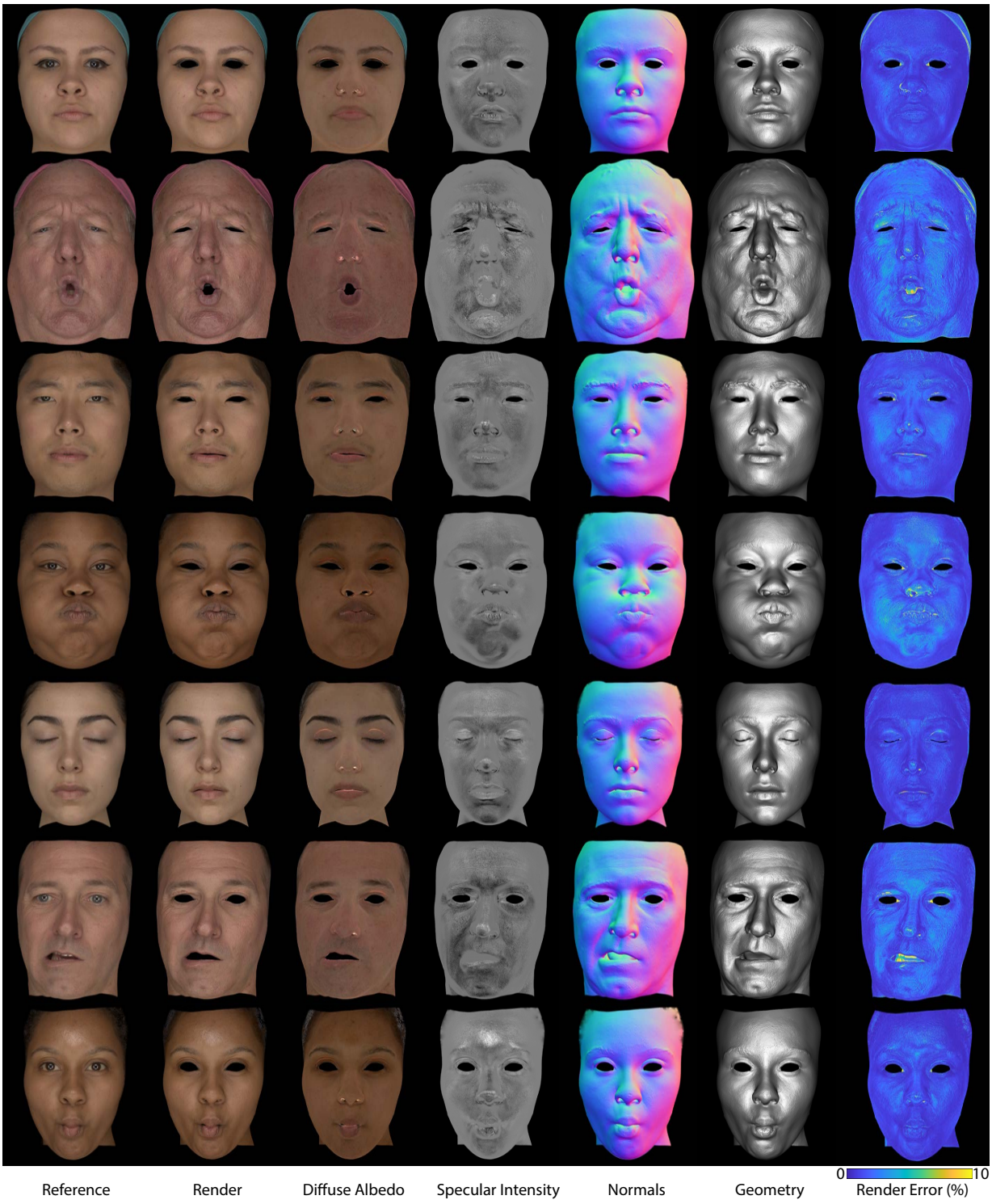


Fig. 4. From a single shot, the proposed system reconstructs geometry and high-quality appearance maps: RGB albedo, specular intensity, and displacement map with fine surface detail. These maps lead to realistic renderings that closely match the real face images, as indicated by the small rendering errors (averaged over the RGB channels, relative to range of image intensities).



Fig. 5. We validate our result by re-rendering a face under novel illumination conditions, comparing to ground truth captured data. The optimized result captured under all 4 flashes (a) is relit with the right side flash removed (b), only the top flash on (c), as well as top and bottom flashes on (d). Our re-render closely matches the ground truth images.

subsurface scattering effects lead to realistic renderings of their digital doubles also in different illumination conditions.

Our method outputs a set of appearance maps, namely diffuse albedo and specular intensity, and a tangent-space displacement map, as well as global parameters for our two-lobe specular BRDF model. Figure 4 shows these maps projected onto our base mesh for seven different subjects in a variety of facial expressions. Fine-scale geometry is shown both as a normal map and as displaced high-resolution mesh. As demonstrated in the first two columns, these maps can be used with our rendering model to generate images that faithfully reproduce the appearance of the captured real faces.

To assess the generalization quality of our results, we also compare our renderings to actual face images captured under new illumination conditions. Figure 5 shows three additional, real face images that were captured under different illumination conditions. The renderings generated for these novel illumination conditions closely matches the real images, even though these conditions were not used for the inverse rendering step.

In the next experiment, we assess the importance of modeling subsurface scattering, an effect that is often neglected in inverse rendering methods for face capture. When subsurface scattering is ignored, its natural blurring effect is baked into normals and the albedo map. This leads to poor recovery of detail in these maps, despite reasonably good fits (re-renderings) for mostly diffuse objects. For shinier surfaces, highlights are rendered incorrectly by these blurry normals. Conversely, even with accurate sharp recovery of the normals, the highlights will appear correct but the diffuse layer will appear unnatural for skin without subsurface scattering, leading to poor re-renderings as illustrated in the ablation result in Fig. 6. Subsurface scattering is indeed important to give skin its soft and organic appearance. Furthermore, modeling subsurface scattering is important because accounting for its spatial low-pass filtering effect allows other computed appearance maps to become sharper. In our experiments, we observed that the computed RGB albedo maps



Fig. 6. Subsurface scattering is often neglected in facial appearance capture and can lead to unnatural rendering when not accounted for (a); our model considers this effect explicitly and can more realistically reproduce the soft, organic appearance of skin (b).

become sharper when this phenomenon is accounted for, as shown in Fig. 7. Thus, our method not only estimates sharper normal and albedo maps, but also renders with high fidelity both diffuse and specular layers since we account for subsurface scattering.



Fig. 7. Subsurface scattering, if not accounted for, can lead to a loss of detail in the recovered RGB albedo (left); by modeling this effect we can recover sharper albedo maps (right).

We also compare the quality of our appearance capture results with those computed with the algorithm of Gotardo et al. [2018]. Although their capture method also operates under constant illumination, their approach requires a sequence of about 20 images of a neutral face, rotating relative to the lights, to compute the neutral albedo map since they factorize diffuse and specular signals algorithmically. Separating these signals algorithmically is extremely challenging, especially in a single-shot, since many more observations are required. When applied for single-shot capture, fitting only to our parallel polarized data, their results show significant artifacts and traces of specular reflection baked into the albedo. In contrast, our maps show much cleaner results since we employ cross-polarization to cancel the specular signal physically (Fig. 8). Additionally, detail loss is also observed as they suffer from motion blur and do not model subsurface scattering.

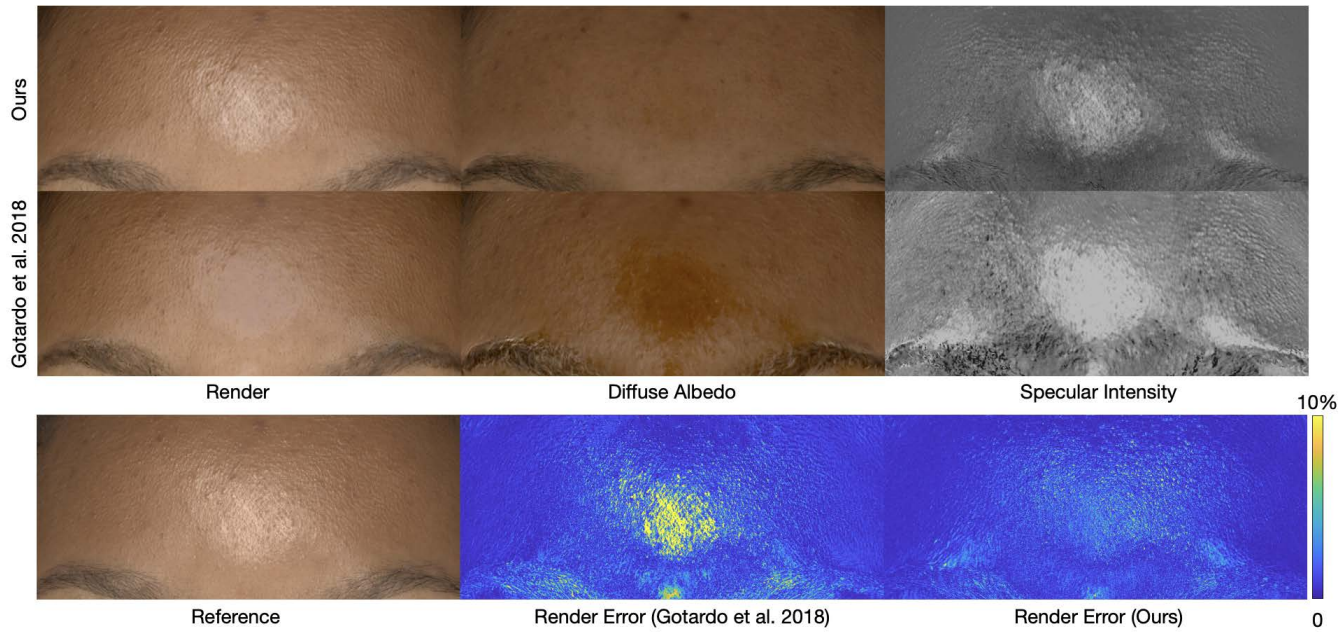


Fig. 8. Single-shot appearance capture of subject with darker skin tone and shiny forehead, showing strong specular highlights. On single-shot data, the (multi-frame) method of Gotardo et al. [2018] shows significant artifacts and specular residual in the albedo, as correct diffuse-specular layer separation is more difficult to achieve algorithmically. As the re-render errors illustrate, our new method can more faithfully reproduce the reference image appearance.

To assess the value of our geometry model and the proposed coarse-to-fine inverse rendering approach, the ablation study in Fig. 9 shows a visual comparison of the recovered level of fine surface detail when computing the displacement map directly at final 4K resolution (Fig. 9 (left)) versus when estimating in a coarse-to-fine manner, starting at 2K (Fig. 9 (right)). As illustrated in the figure, coarse-to-fine estimation improves the level of recovered geometric detail, providing more pronounced features such as deeper wrinkles and expression lines. The direct approach in Fig. 9 (left) is closer to that of Gotardo et al. [2018], which in addition require post-processing to convert the estimated normals to actual geometry, leading to further loss of detail. Coarse-to-fine allows to solve for even higher resolutions, where each additional resolution adds finer scale detail as shown in Fig. 10 at the example of an 8K zoom in. An additional comparison is given in Fig. 11, which compares the level of recovered detail against the popular dark-is-deep approach of [Beeler et al. 2010]. Here we see that the proposed method produces crisper geometric detail, whereas the dark-is-deep heuristic is not physically accurate and can be mislead by skin pigmentation.

The effect of increasing or decreasing the two geometric regularization weights $\{\lambda_1,\lambda_2\}$ in Eq. 9 is illustrated in Fig. 12. Increased weights provide smoother geometry that is closer to the initially fitted, low-polygon template face mesh in Fig. 2. Less regularization provides sharper geometric detail but may allow localized artifacts to be introduced in under-constrained skin patches. Thus, an artist using our system to build a digital human could opt for lower weights and increased level of detail, at the cost of requiring manual touch up in localized areas. We note that no such post-processing was applied to any of the results shown in this paper.

An important aspect of our single-shot capture system is that it can be readily applied to *dynamic* sequences, containing for example facial performance, by recovering the appearance of each frame independently. In contrast to common polarization-based capture systems, we do not require temporal changes like active lighting or fast switching of polarization filters, as we can capture with constant, uniform illumination. This leads to better comfort for the actor and also maintains accuracy of markerless mesh tracking and motion capture using existing solutions [Beeler et al. 2011]. As a result, our proposed method enables dynamic face capture at camera frame rates, in full temporal correspondence, and with good temporal stability. This is illustrated in Fig. 13 and in the supplementary video. Here, instead of DSLR cameras and studio flashes, our capture setup included 12 machine vision cameras (20MP Ximea, also arranged as in Fig. 2) and constant lighting from horizontally polarized LED banks. Results were computed with the method as described above, showing the flexibility of the approach.

The ultimate goal of these digital assets is the ability to re-render them under novel illumination conditions to allow to integrate these digital character into arbitrary virtual scenes as shown in Fig. 14.

6 DISCUSSION

In this paper, we have presented a new light-weight face capture system for high-quality acquisition of both facial geometry and appearance. Typically, high-quality appearance acquisition has been performed only in expensive and complex capture stages that time-multiplex light with different polarization, while the face is observed by several cameras with equally oriented polarization filters. In contrast, we propose a system with constant lighting under a single

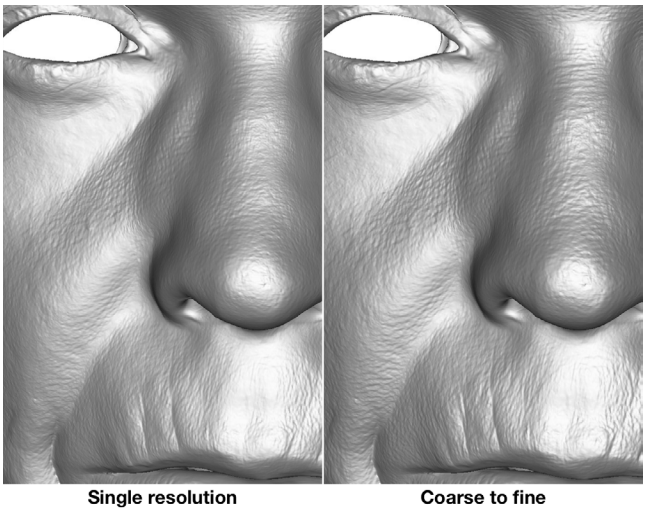


Fig. 9. Direct displacement map estimation at 4K resolution (left). Our coarse-to-fine optimization (right) provides better initialization at the final resolution, leading to more pronounced detail in shorter runtimes, as expected from a multi-resolution solver of differential equations. The RMSE of the re-rendered blue channel (sharpest channel), averaged over all views, is also smaller (0.0137) than that of the direct solver (0.0143).

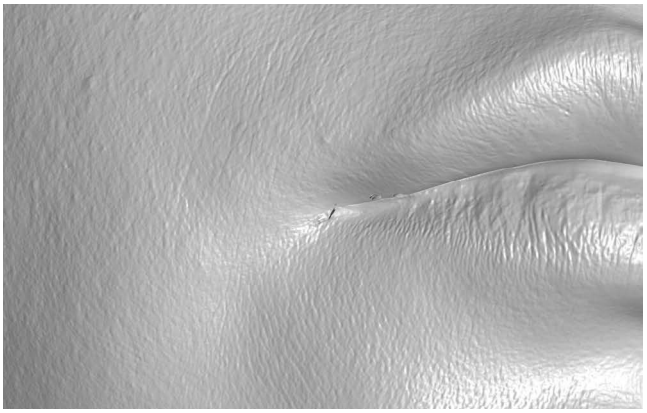


Fig. 10. Most results of the paper are using 4K appearance maps, but the system can also solve for higher resolutions, such as 8K in this zoom in.

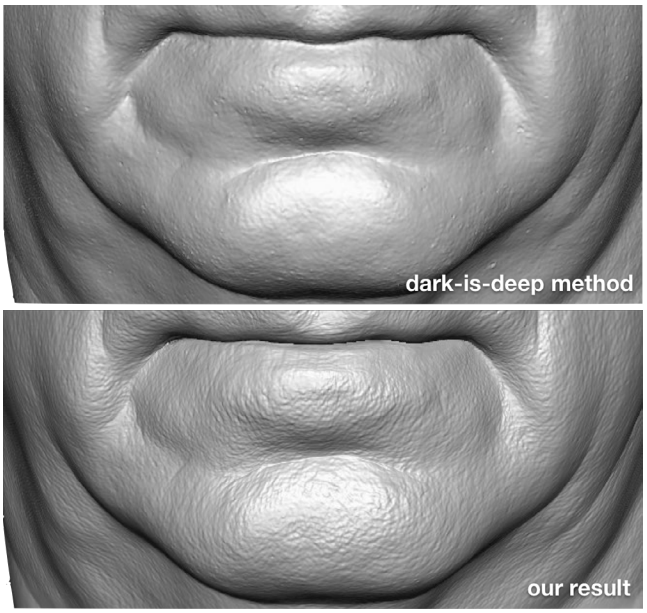


Fig. 11. Comparison of geometric detail against the popular dark-is-deep method proposed by [Beeler et al. 2010] (top), which is a heuristic approach that is not optimized for re-rendering and has been found to produce inaccurate results [Ghosh et al. 2011]. Our method is physically motivated and recovers crisper and more realistic geometric detail (bottom), leading to high-fidelity rerenderings as shown in Fig. 4.

polarization state (horizontal) and cameras that have different filter orientations. As a result, rather than operating in a time-multiplexed way, our system performs "view-multiplexed" capture of polarized light from a single exposure. It can be assembled from standard, off-the-shelf hardware components (from high- to low-end models) and, therefore, can be readily integrated into current low-cost, passive photogrammetry solutions that are already in widespread use. We believe the proposed system has the potential to democratize the creation of digital human assets, making it affordable also to lower-budget projects both inside and outside the entertainment industry.

We demonstrated the high-quality output of our new system on a number of human subjects with different gender, ages, and skin typology, leading to realistic rendering of their captured digital doubles. However, the system is not without its limitations. First, skin roughness (size of specular lobe) cannot be estimated on a per-texel basis due to the typically soft nature of lighting in photogrammetry setups such as ours [Ghosh et al. 2008]. As a result, some face regions may be rendered slightly more/less rough in comparison to the real image and specular intensity compensates for the fixed-sized specular lobe. Figure 4 also shows other localized rendering artifacts: the nose ring (and other accessories) presents less accurate base geometry and its appearance diverges more from the assumed skin model; also, correctly fitting the base geometry of eye lids is still challenging and mesh self-intersection can lead to inaccurate shadow maps and rendering results. Although our appearance model explicitly accounts for subsurface scattering effects, we use a fixed diffusion profile taken from the literature [d'Eon et al. 2007]. In practice, the diffusion profile varies across subjects, especially with age, and a person-specific profile estimate could lead to more realistic results. Finally, while we focus on modeling the appearance of human skin, an exciting direction in recent work is to realistically capture the likeness of the entire human head (including hair, teeth, eyes) with a single data-driven approach [Lombardi et al. 2019; Thies et al. 2019]. Nevertheless, our approach already provides accessible, high-fidelity skin appearance data that offers direct practical value for VFX and video game productions, and can also be used to train and advance future deep learning methods, which are still limited in terms of spatial resolution and level of detail.

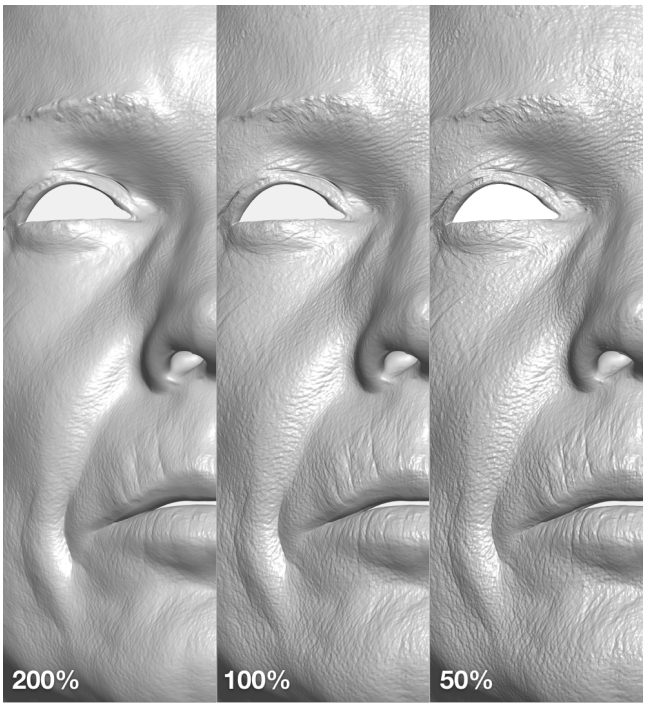


Fig. 12. Effect of regularization weights (λ_1,λ_2) on the recovered geometric detail: high-frequency detail is attenuated with increasing regularization (left); wrinkles and creases become deeper and sharper with less regularization, which can also introduce small reconstruction artifacts (right).



Fig. 13. Our method is readily extendable to dynamic capture as we treat each video frame independently. Here we show renders as well as a breakdown of the different maps obtained using our method on video sequences of two different actors going in and out of dynamic facial expressions (please refer to the supplementary video).

ACKNOWLEDGMENTS

The authors would like to thank all the capture subjects, Mark Shriver and Sarita Greer for the help with data capture, and Prashanth Chandran for assisting with data processing.

REFERENCES

Sameer Agarwal, Keir Mierle, and Others. 2016. Ceres Solver. http://ceres-solver.org.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-Quality Single-Shot Capture of Facial Geometry. *ACM Transactions on Graphics (TOG)* 29, 3 (2010), 40:1–40:9.

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. *ACM Transactions on Graphics (ACM)* 30, Article 75 (August 2011), 10 pages. Issue 4.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High resolution passive facial performance capture. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 41.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time High-fidelity Facial Performance Capture. *ACM Transactions on Graphics (TOG)* 34, 4, Article 46 (July 2015), 9 pages.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., ACM, 145–156.

Eugene d'Eon, David Luebke, and Eric Enderton. 2007. Efficient Rendering of Human Skin. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques (EGSR'07)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 147–157.

Craig Donner and Henrik Wann Jensen. 2005. Light Diffusion in Multi-Layered Translucent Materials. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 1032–1039.

Graham Fyffe and Paul Debevec. 2015. Single-Shot Reflectance Measurement from Polarized Color Gradient Illumination. In *International Conference on Computational Photography (ICCP)*. IEEE.

Graham Fyffe, Paull Graham, Borom Tunwattanapong, Abhijeet Ghosh, and Paul Debevec. 2016. Near-Instant Capture of High-Resolution Facial Geometry and Reflectance. *Computer Graphics Forum (CGF)* 35, 2 (2016), 353–363. https://doi.org/10.1111/cgf.12837

Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul Debevec. 2011. Comprehensive Facial Performance Capture. *Computer Graphics Forum (CGF)* 30, 2 (2011).

Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. 2014. Driving High-Resolution Facial Scans with Video Performance Capture. *ACM Transactions on Graphics (TOG)* 34, 1, Article 8 (Dec. 2014), 14 pages. https://doi.org/10.1145/2638549

Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. *ACM Transactions on Graphics (TOG)* 32, 6 (November 2013), 158:1–158:10. https://doi.org/10.1145/2508363.2508380

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)* 30, 6 (2011), 129.

Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. 2008. Practical Modeling and Acquisition of Layered Facial Reflectance. *ACM Trans. Graph.* 27, 5 (Dec. 2008), 139:1–139:10.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical Dynamic Facial Appearance Modeling and Acquisition. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 232:1–232:13.

Paulo Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. 2015. Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction. In *IEEE International Journal of Computer Vision*. IEEE, 846–854.

Paul Graham, Borom Tunwattanapong, Jay Busch, Xueming Yu, Andrew Jones, Paul Debevec, and Abhijeet Ghosh. 2013. Measurement-Based Synthesis of Facial Microgeometry. *Computer Graphics Forum (CGF)* 32, 2 (2013), 335–344.

Tim Hawkins, Andreas Wenger, Chris Tchou, Andrew Gardner, Fredrik Göransson, and Paul Debevec. 2004. Animatable Facial Reflectance Fields. In *Proceedings of the Fifteenth Eurographics Conference on Rendering Techniques* (Norrköping, Sweden) *(EGSR'04)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 309–319. https://doi.org/10.2312/EGWR/EGSR04/309-319

Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. 2018. Mesoscopic Facial Geometry Inference Using Deep Neural Networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 45:1–45:14.



Fig. 14. The high-quality facial geometry and appearance data produced by the proposed system allows to re-render the faces under novel illumination conditions (forest on the left and city square on the right), fulfilling the ultimate goal when creating of digital human assets.

Henrik Wann Jensen, Steve Marschner, Marc Levoy, and Pat Hanrahan. 2001. A practical model for subsurface light transport. In *In Proceedings of ACM SIGGRAPH*. ACM, 511–518.

Christos Kampouris, Stefanos Zafeiriou, and Abhijeet Ghosh. 2018. Diffuse-specular Separation Using Binary Spherical Gradient Illumination. In *Proceedings of the Eurographics Symposium on Rendering: Experimental Ideas & Implementations* (Karlsruhe, Germany) *(SR '18)*. 1–10. https://doi.org/10.2312/sre.20181167

Oliver Klehm, Fabrice Rousselle, Marios Papas, Derek Bradley, Christophe Hery, Bernd Bickel, Wojciech Jarosz, and Thabo Beeler. 2015. Recent Advances in Facial Appearance Capture. *Computer Graphics Forum (CGF)* 34, 2 (May 2015), 709–733.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article Article 65 (July 2019), 14 pages. https://doi.org/10.1145/3306346.3323020

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques (EGSR'07)*. Eurographics Association, 183–194.

Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. 2008. Facial Performance Synthesis Using Deformation-driven Polynomial Displacement Maps. *ACM Transactions on Graphics (TOG)* 27, 5, Article 121 (Dec. 2008), 10 pages. https://doi.org/10.1145/1409060.1409074

Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, and et al. 2019. Deep Reflectance Fields: High-Quality Facial Reflectance Field Inference from Color Gradient Illumination. *ACM Trans. Graph.* 38, 4, Article Article 77 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323027

Koki Nagano, Graham Fyffe, Oleg Alexander, Jernej Barbic, Hao Li, Abhijeet Ghosh, and Paul E Debevec. 2015. Skin microstructure deformation with displacement map convolution. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 109.

Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. 2005. Efficiently combining positions and normals for precise 3D geometry. In *ACM transactions on Graphics (TOG)*, Vol. 24. ACM, 536–543.

Matt Pharr, Wenzel Jakob, and Greg Humphreys. 2016. *Physically Based Rendering: From Theory to Implementation (3rd ed.)* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 1266 pages.

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2017. Photorealistic Facial Texture Inference Using Deep Neural Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 222.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* 38, 4, Article Article 66 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323035

Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. 2012. Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting. *ACM Transactions on Graphics (TOG)* 31, 6 (November 2012), 187:1–187:11. https://doi.org/10.1145/2366145.2366206

Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (Grenoble, France) *(EGSR 2007)*. Eurographics Association, Goslar, DEU, 195–206.

Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 756–764.

Tim Weyrich, Jason Lawrence, Hendrik P. A. Lensch, Szymon Rusinkiewicz, and Todd Zickler. 2009. Principles of Appearance Acquisition and Representation. *Found. Trends. Comput. Graph. Vis.* 4, 2 (Feb. 2009), 75–191.

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. 2006. Analysis of Human Faces Using a Measurement-based Skin Reflectance Model. *ACM Transactions on Graphics (TOG)* 25, 3 (July 2006), 1013–1024.

Cyrus A Wilson, Abhijeet Ghosh, Pieter Peers, Jen-Yuan Chiang, Jay Busch, and Paul Debevec. 2010. Temporal upsampling of performance geometry using photometric alignment. *ACM Transactions on Graphics (TOG)* 29, 2 (2010), 17.

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-Fidelity Facial Reflectance and Geometry Inference From an Unconstrained Image. *ACM Transactions on Graphics (TOG)* 37, 4 (2018).