Practical and Scalable Desktop-based High-Quality Facial Capture

Alexandros Lattas[2] ![ORCID], Yiming Lin[1] ![ORCID], Jayanth Kannan[1] ![ORCID], Ekin Ozturk[1,2] ![ORCID], Luca Filipi[1] ![ORCID], Giuseppe Claudio Guarnera[1,3] ![ORCID], Gaurav Chawla[1] ![ORCID], and Abhijeet Ghosh[1,2] ![ORCID]

[1] Lumirithmic Ltd, London, UK
{yiming.lin,jay,luca,gaurav.chawla}@lumirithmic.com
[2] Imperial College London, London, UK
{a.lattas,ekin.ozturk17,ghosh}@imperial.ac.uk
[3] University of York, York, UK claudio.guarnera@york.ac.uk

Abstract. We present a novel desktop-based system for high-quality facial capture including geometry and facial appearance. The proposed acquisition system is highly practical and scalable, consisting purely of commodity components. The setup consists of a set of displays for controlled illumination for reflectance capture, in conjunction with multi-view acquisition of facial geometry. We additionally present a novel set of modulated binary illumination patterns for efficient acquisition of reflectance and photometric normals using our setup, with diffuse-specular separation. We demonstrate high-quality results with two different variants of the capture setup – one entirely consisting of portable mobile devices targeting static facial capture, and the other consisting of desktop LCD displays targeting both static and dynamic facial capture.

Keywords: Facial capture, active illumination, reflectance, photometric normals, diffuse-specular separation, dynamic capture

(a) Tablet-based Setup   (b) Render   (c) Monitor-based Setup   (d) Render
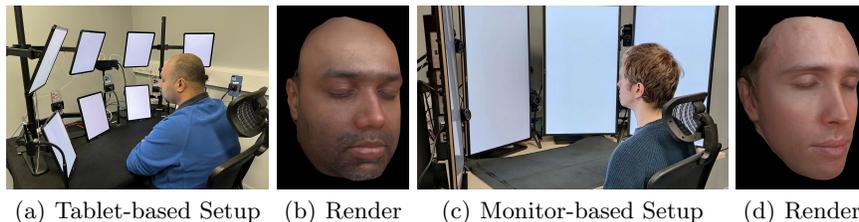
Fig. 1: Two proposed novel desktop-based setups (a, c) for high-quality facial capture (b, d). Left: setup consisting of a set of portable mobile devices – tablets and smartphones, for static facial capture. Right: setup consisting of a set desktop LCD displays for static and dynamic facial capture.

1   Introduction

Realistic reconstruction and rendering of human faces has been a long standing goal in computer vision and graphics. It has wide ranging applications in well known sectors such as entertainment, advertising, AR/VR, training systems, etc., as well novel applications of digital avatars in social media, and virtual presence in the envisioned *metaverse*. Great strides have been made towards achieving this goal over the last two decades, both in terms of development of photo-realistic rendering algorithms, as well as realistic modeling of facial shape and appearance. The latter has been revolutionized with the development of acquisition techniques for high-quality 3D facial capture. However, creation of realistic digital characters from acquired data requires very high-quality data capture typically acquired using custom designed and expensive capture systems such as the Lightstage [4][29][21][10]. This usually restricts the application of such acquisition techniques in practice to big-budget film and games VFX, and/or to organizations that can either afford to access such setups at specific locations or to construct such capture systems in-house. Also, statistical [27] and deep-learning methods [9][19][20] still lack in quality compared to the above. This limits the accessibility of such 3D acquisition technology for wider applications.

Instead, this work aims to make acquisition of high quality facial geometry and appearance of a subject much more practical and accessible, thereby democratizing the creation of realistic virtual humans for various applications. Specifically, we propose a practical and scalable desktop-based setup for high-quality facial capture. The setup consists of a set of displays for active control of illumination for facial appearance capture, in conjunction with multiview acquisition of facial geometry. We present two different variants of the capture setup constructed entirely using commodity components – a highly portable setup consisting of a set of mobile devices (tablets and smartphones) targeting static facial capture, and a setup consisting of a set of desktop LCD displays targeting both static and dynamic facial capture (see Fig. 1). We additionally present a novel set of modulated binary illumination patterns emitted by the display panels in our setup for efficient acquisition of diffuse-specular separated reflectance and photometric normals. We present high-quality results of facial capture using both the proposed setups, and demonstrate the results to be qualitatively competitive with those obtained using dedicated, expensive capture setups. To summarize, this work introduces:

1. Two novel, scalable desktop setups for reflectance and shape capture, comprising only of commodity components.
2. A novel photometric acquisition pipeline with diffuse-specular separation employing modulated binary illumination that achieves high-quality results with the proposed desktop setups.
3. Novel design of spectral multiplexing for binary illumination that enables high-quality two-shot capture and estimation of spatially-varying specular roughness.
4. Demonstration of applicability of proposed setup and spectral multiplexing for dynamic facial appearance capture.

2   Related Work

Here we limit the discussion to the most related works on facial capture. For a broader discussion, we refer the reader to an excellent surveys on the topic [18][28].

2.1   Active Illumination

Photometric stereo systems have traditionally been popular in computer vision for estimating surface normals, in conjunction with color-multiplexed illumination for dynamic capture [14][17]. Debevec et al. [4] first proposed a specialized light stage setup to acquire a dense reflectance field of a human face for photo-realistic image-based relighting applications. They also fit microfacet-BRDF based reflectance using the acquired data. Weyrich et al. [29] instead employed an LED sphere and multiple cameras to densely record facial reflectance and computed view-independent estimates of spatially varying facial reflectance from the acquired data for driving a microfacet BRDF model-based rendering. Subsequently, Ma et al. [21] introduced polarized spherical gradient illumination (using an LED sphere) for efficient acquisition of the separated diffuse and specular albedos and photometric normals of a face, and demonstrated high quality facial geometry including skin mesostructure as well as realistic rendering with the acquired data.

Ghosh et al. [11] further extended the acquisition method to acquire layered facial reflectance using a combination of polarization and structured lighting, and further extended the view-dependent solution of Ma et al. for multi-view facial acquisition with polarized spherical gradient illumination [10]. These techniques have had significant impact in film VFX. Fyffe et al. [8] employed the method of [30] for temporal alignment of spherical gradient illumination for dynamic performance capture. Their method involves a heuristic based separation of reflectance which is not accurate and the capture process requires high speed cameras. Fyffe & Debevec [6] further proposed a single-shot method using a complex setup of a polarized RGB LED sphere for color-multiplexed spherical gradients that are further polarized using the method of [10] for diffuse-specular separation.

Closer to our work, Kampouris et al. [16] have employed binary spherical gradient illumination using an LED sphere for facial appearance capture including diffuse-specular separation of albedo and photometric normals. We employ a similar analysis of binary illumination in our work for diffuse-specular separation. However, we demonstrate our lighting patterns in conjunction with our illumination setup to result in fewer measurements for facial capture, even enabling high-quality results with spectral multiplexing which is not achieved by [16]. More recently, Guo et al. [13] have employed complementary pair of color-multiplexed spherical gradient illumination patterns using a spectral LED sphere for estimating diffuse and specular reflectance and surface normals for full-body volumetric capture. Unlike their approach, we employ a much simpler acquisition setup and estimate a broader set of spatially-varying reflectance parameters from each camera viewpoint using our spectral multiplexing approach.

An alternate approach that does not require an LED sphere is the work of Fyffe et al. [7] on static facial appearance capture employing off-the-shelf photography hardware. However, the approach does not extend to dynamic facial appearance capture and still requires quite a complex hardware setup consisting of 24 cameras and 6 flashes that are triggered in sequence within a few milliseconds. Our approach is inspired by these previous works. However, we aim to significantly simplify the capture process using commodity components and propose highly scalable desktop-based active illumination systems for high-quality facial capture. Very related to our approach is the recent work of Sengupta et al. [26] who employ a single desktop monitor to illuminate a face using a regular video sequence in order to learn a facial reflectance field with limited angular coverage. However, the goal of this work is to support facial relighting for video conferencing applications. Our desktop setup with brighter and wider illumination coverage is much more suitable for high-quality multiview capture, and we propose lighting patterns for direct acquisition of reflectance and photometric normals instead of learning based inference.

2.2   Passive Capture

Researchers have also investigated approaches for passive acquisition of faces and full-body. Such acquisition is particularly well suited for performance capture since active approaches usually require time-multiplexed illumination, imposing requirements of high frame rate acquisition, lighting control and synchronization. A popular approach has been to employ uniform constant illumination for multi-view facial capture [1][3]. Such an approach enables estimation of an albedo texture under flat lit illumination for rendering purposes besides facial geometry reconstruction based on multi-view stereo. High-frequency mesostructure is further embossed on the base-geometry using a high-pass filter on the diffuse texture [1]. The approach has been extended for reconstructing facial performances with drift-free tracking over long sequences using anchor frames [2]. While producing very good qualitative results for facial geometry, the estimated albedo is not completely diffuse and contains a small amount of specular reflectance baked into the texture and the embossed mesostructure is plausible rather than accurate.

More recent works have extended passive capture systems for achieving facial reflectance estimation using inverse rendering in conjunction with view-multiplexing [12][22]. However, while achieving impressive results for passive capture, the quality of acquired reflectance maps is not quite at par with active illumination systems. Gotardo et al. [12] are also able to acquire dynamic facial appearance like our monitor-based setup. However, their method requires initialization of neutral facial appearance through dense multiview capture in conjunction with measurement of person-specific change in skin color due to blood flow to constrain the optimization under passive illumination. Our method does not require any initialization or person-specific measurements and directly estimates high-quality appearance maps for each frame of dynamic capture.

3   Desktop-based Capture System

We propose two desktop-based setups for facial capture in this work. Both setups employ specific types of display panels for controlled illumination, while employing a set of cameras for multiview acquisition. Both setups have been designed to provide controlled illumination over a frontal hemispherical zone of directions spanning $\beta < \pm 90°$ along the longitudinal directions ϕ, while spanning $\alpha < \pm 45°$ along the latitudinal directions θ (see supplemental material). Such a zone has previously been shown to include majority of facial surface normals [7].

3.1   Tablet-based Setup

Our first desktop setup consists purely of a set of mobile devices – eight tablets and five smartphones, that are mounted on a desk as shown in Fig. 1 (a). The tablets are arranged in two rows (latitudes) of four devices and oriented longitudinally so as to cover a significant zone of the frontal hemisphere around a subject's face. The screens of the tablets are oriented towards the subject and provide controlled piece-wise continuous illumination for acquiring facial reflectance. The selfie cameras on the tablets are employed to capture facial photometric response due to illumination during a capture sequence. These cameras are typically lower in resolution so they mostly serve to provide data for multiview 3D reconstruction (e.g., using structure-from-motion (SFM) [23]). We also place five smartphones in the setup along the equatorial plane, one between each column of the tablets, and employ their high-resolution back cameras (zoom lens) for acquiring facial reflectance and photometric normals. The tablets and the smartphones are mounted on a desk using appropriate table mounts. In this work, we employed iPad Air devices (4th generation) for the tablets and iPhone 12 Pro devices as the smartphones in our setup. Other types of tablets and smartphones (e.g., Android based) could also be used instead for the setup.

The devices are all controlled in synchronization during a capture process where one device acts as the master and the other devices act as slaves and they wirelessly communicate with each other over wifi. Once a capture command is initiated by an operator on the master device, the master broadcasts the start time for the capture process to all slave devices. All devices in the setup pool the global GPS clock-time from the GPS signal, and having received the capture instruction and start-time from the master device, then execute their own individual capture processes. Each device has a pre-built capture process that for the tablets includes illuminating a sequence of patterns from the device screen and simultaneously recording images using the selfie-camera, while for the smartphones involves capturing high-resolution images using the back zoom camera. All devices start and end their capture processes together in synchronized fashion in a few seconds due to pre-set capture/illumination timings. Afterwards, the acquired data from all devices is transferred to a single machine for off-line processing. The data transfer is also done wirelessly once initiated by an operator. The entire setup is highly portable since it consists of only the mobile devices and their table mounts. Hence, it can be easily moved from one

location to another and quickly re-assembled for use without having to deal with any custom electronics and wiring.

3.2 Monitor-based Setup

Our second setup consists of four desktop LCD monitors that we mount together on a desk in portrait mode as shown in Fig. 1 (c). We employ four 27" 4K monitors with 16:9 aspect ratio (Asus ProArt PA279CV) in our setup. The monitors are arranged longitudinally facing the subject to cover a similar frontal hemispherical zone of directions with screen illumination. The four monitors are all controlled by single workstation (running Windows 10) via HDMI and display ports. The workstation includes a high-end graphics card (Nvidia RTX 3070 Ti) which is employed to drive the four monitors. We additionally mount stereo pairs of digital cameras within the vertical gaps between the monitors for multiview capture. Eight cameras are mounted using small desk-tripods and we chose mirrorless cameras with small form factor (Canon EOS M200) for ease of mounting in the setup. During acquisition, the workstation controls the monitors and the set of cameras in synchronization to rapidly capture a sequence of images under a set of controlled illumination patterns for 3D shape and reflectance capture. The monitor-based setup has the advantage over the tablet-based setup of much more fine-grained synchronization between multiple displays supporting much faster capture, even supporting video rate capture required for acquiring dynamic facial performance.

3.3 Modulated Binary Illumination

We employ horizontally and vertically aligned binary illumination patterns in this work, over the hemispherical zone of illumination of the proposed capture setups, for acquiring albedo and photometric normals with diffuse-specular separation. While this is similar to the approach of Kampouris et al. [16], our patterns require additional form-factor modulation due to being near-field with limited angular extent. Furthermore, we make the crucial observation that over the zone of hemispherical illumination covered by the displays in our setup, we do *not* need all three axis aligned (X, Y, Z) binary illumination conditions and their complements employed by [16]. Specifically, when we illuminate a subject with the horizontally aligned binary pattern and its complement (H and H' respectively), the lit portion of the zone does not have its centroid aligned with X-axis but is actually centered around $+45°$ and $-45°$ directions respectively in the XZ plane (with $0°$ corresponding to the $+Z$ axis). This results in the complementary pair of horizontal binary patterns H and H' to have their centroids orthogonal to each other in the XZ plane (albeit with a $45°$ in-plane rotation), and hence sufficient for determining the x and z components of a photometric normal. We additionally only need measurement of the vertically aligned binary pattern V and its complement V' to determine the y component of the photometric normal. This reduces the number of measurements to only four photographs under the horizontal and vertical binary illumination patterns (see Fig. 2, a).

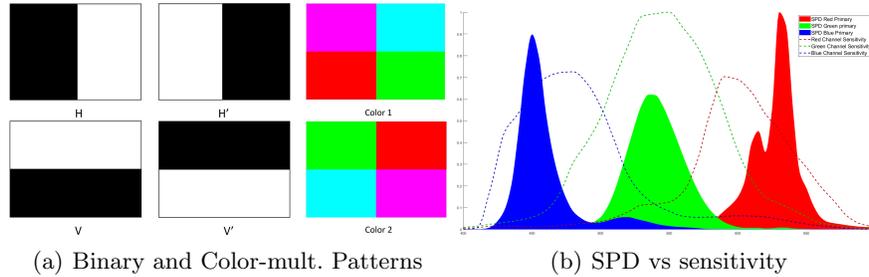(a) Binary and Color-mult. Patterns   (b) SPD vs sensitivity

Fig. 2: (a) Four binary illumination patterns (H, H', V, V'), and color-multiplexed patterns used for 2-shot acquisition. (b) Display illumination SPD (solid fill) plotted against the spectral sensitivity of iPhone 12 Pro camera (dashed lines). The color patterns are designed to avoid cross-talk between green and red channels.

Spectral Multiplexing For faster acquisition, we further reduce the number of measurements to only two photographs by employing spectral multiplexing of the binary patterns using the R, G and B color channels of the display panels. While this is similar to the spectral multiplexing approach of [16], we demonstrate much higher quality results for faces with our spectral multiplexing scheme which are almost at par with results obtained using four measurements under binary white illumination. Unlike previous work employing spectral multiplexing for photometric measurements, we observe that typical RGB illumination observed by standard RGB cameras can suffer from a bit of spectral cross-talk between neighboring wavelengths. Specifically, depending on the spectral power distribution (SPD) of the RGB source and the spectral sensitivity of the camera, this can result in some cross-talk between the blue and green channels, or the green and red channels sensed by the camera. Both our setups employ LCD displays with DCI P3 color gamut. We measured the SPD of our displays using a spectrometer (Sekonic SpectoMaster C700) and camera spectral sensitivity (using the procedure of [15]), and found the green and red illumination emitted by the displays to be spectrally closer to each other due to an amber peak in the red illumination spectrum. This causes the green channel of the cameras we employ to be sensitive to red illumination emitted by the displays, and hence causes some spectral cross-talk between these two color channels (Fig. 2(b), see supplemental material for details).

In order to deal with this spectral cross-talk, we exploit our horizontal and vertical binary patterns to design two novel complementary spectral multiplexing patterns for high-quality photometric measurements. Specifically, we ensure that the patterns do not have any overlap of green and red illumination and we spectrally encode the horizontal and vertical binary patterns and their complements in the two patterns shown in Fig. 2(a). Each pattern consists of both the horizontal pattern H and its complement H' encoded in the red and green channels respectively (or vice versa), while the blue channel encodes either the vertical pattern V or its complement V'. The two patterns are complementary and sum to uniform white illumination.

Form-factor Modulation As previously mentioned, we additionally need to modulate the intensity of the binary illumination patterns to compensate for the form-factor of the measurement setup. This is required since the illumination is restricted to a hemispherical zone and hence is not uniformly incident from all directions (unlike an LED sphere). Specifically, due to the limited extent of the zone in the vertical (latitudinal) direction compared to the horizontal (longitudinal) direction, we need to modulate the intensity along the vertical direction to create an even distribution of illumination. We do this by reducing the intensity nearer to the equator compared to latitudes near the top and bottom of the zone. This form-factor modulation also needs to account for any local effects of illumination due to the setup geometry. We pre-compute this vertical modulation function using inverse rendering a diffuse sphere lit by display panel arrangements of our two setups. We then apply this vertical modulation function to the binary illumination patterns during acquisition.

Multi-view Capture The acquired patterns are axis-aligned with the global coordinates of the capture setup. We observe that the patterns form a steerable basis and rotate the corresponding pairs into the local coordinate frame of each camera before employing the photometric estimation process described in Sec. 4.

4   Reflectance and Shape Estimation

4.1   Acquisition using White Illumination

Given the acquired set of photographs of a subject under the four binary patterns $\mathbf{H}, \mathbf{H}', \mathbf{V}, \mathbf{V}'$ (Fig. 2), we first rotate the $\mathbf{H}$ and $\mathbf{H}'$ patterns around the Y axis by $\pm 45°$ to additionally obtain the axis aligned patterns $\mathbf{X}, \mathbf{X}'$, and $\mathbf{Z}$. Each binary-complement pair can be added to acquire a full-on observation $\mathbf{F}$, e.g., $\mathbf{F} = \mathbf{X} + \mathbf{X}'$. These observations can then be used to estimate photometric normals and separated diffuse and specular albedo.

Photometric Normals The set of acquired binary patterns can be used to compute mixed photometric normals following [16]. However, due to incomplete hemispherical illumination (limited to $< \pm 45°$ latitudinal span) employed in our setups, this would result in skewed normals. Instead, we employ Singular Value Decomposition (SVD) to extract photometric normals with more correct global orientation. We directly stack the flattened $\mathbf{H}, \mathbf{H}', \mathbf{V}, \mathbf{V}'$ measurements into a matrix $\mathbf{A} \in \mathbb{R}^{WH \times 4}$, where W, H are the width and height of each measurement. As shown by [31], the eigenvectors of the three largest eigenvalues of $\mathbf{A}\mathbf{A}^{\top}$ correspond to the normals components $\mathbf{N^X}, \mathbf{N^Y}, \mathbf{N^Z}$.

The orientation and order of the components of normals is arbitrary with SVD. Therefore we order them in the following way: We create a facial mask $\mathbf{M}$ by using a face detector [5] on F, and design a function $\mathcal{A}(\mathbf{N}, \mathbf{M})$, that aligns the components with the +X,+Y,+Z axes. Specifically, we fit for each component a linear regression on the X and Y axes, and assign each component to the axis

with the highest absolute trend, while using the sign of the trend to orient the component. For $\mathbf{N}^Z$, the regression is fit on half of the X axis. For an $\mathbf{USV}^\top$ SVD decomposition of $\mathbf{A}$ and $\mathbf{U}'$ including the first three vectors of $\mathbf{U}$:

$$\mathbf{N} = \mathcal{A}(\mathbf{U}', \mathbf{M}), \quad \mathbf{A} = \mathbf{USV}^\top \tag{1}$$

Diffuse Normals Shorter-wavelength channels exhibit sharper normal details, because of the wavelength-dependent diffuse scattering of light [21]. Under white illumination, we observe that the different color channels share the same amount of specular reflectance, which is white, but different amounts of diffuse reflectance. We can therefore acquire pure diffuse signal using the spectral differencing between the brightest (red) and darkest channels (blue) of each pattern, i.e., $\mathbf{X}_D = \mathbf{X}^R - \mathbf{X}^B$. The pure diffuse patterns can be used with Eq. 1 to compute the diffuse normals $\mathbf{N}_D$.

Specular Normals We follow the observation of [16] that the mixed normals are a mixture of diffuse and specular normals. However, we employ a simpler empirical signal processing step for estimating specular normals from the mixture. The calculation, although heuristic in nature, produces sharp specular normals, which are able to generate highly photorealistic shading. Formulating mixed photometric normals $\mathbf{N}$ as a combination of the diffuse normals $\mathbf{N}_D$ and a portion of the specular signal $\alpha\mathbf{S}$, we separate the specular signal using a Gaussian high-pass filter $g()$. We empirically set $\alpha = 0.5$ for our results. To obtain the final specular normals $\mathbf{N}_S$, we add the separated specular signal to the diffuse normals and re-normalize the normals.:

$$\mathbf{N}_S = \frac{\mathbf{N}_D + \frac{1}{\alpha}(\mathbf{N} - g(\mathbf{N}))}{\left|\left| \mathbf{N}_D + \frac{1}{\alpha}(\mathbf{N} - g(\mathbf{N})) \right|\right|} \tag{2}$$

Diffuse-Specular Albedo Separation For separating the reflectance albedo into diffuse $\mathbf{A}_D$ and specular $\mathbf{A}_S$ components respectively, we employ the linear system in Eq. 3 originally proposed by [16] for color-multiplexed data. We find the linear system based separation to also be well suited for binary patterns with white illumination. Assuming the binary observation along X is brighter than its complement X', the linear system is expressed as:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{X}' \end{bmatrix} = \begin{bmatrix} \mathbf{N_D}.\mathbf{x} & 1 \\ (1 - \mathbf{N_D}.\mathbf{x}) & 0 \end{bmatrix} \begin{bmatrix} \mathbf{A}_D \\ \mathbf{A}_S \end{bmatrix}, \tag{3}$$

where $\mathbf{N_D}.\mathbf{x}$ is the x component of the diffuse normal scaled to $[0, 1]$. Eq. 3 can be solved separately for each binary-complement pair to acquire an estimate of $\mathbf{A}_S$. Instead of the median of three axis-aligned solutions proposed by Kampouris et al., we find the best estimate as the average $\mathbf{A}_{S_{avg}}$ of solutions for only the $\mathbf{X}$ and $\mathbf{Y}$ binary pairs. Finally, for computing the diffuse albedo $\mathbf{A}_D$, we solve the linear system independently for each color channel and then subtract the channel wise $\mathbf{A}_{S_{avg}}$ from full-on $\mathbf{F}$ to obtain $\mathbf{A}_D$.

4.2   Color-Multiplexed Illumination

We reduce measurements to a two-shot capture process using color multiplexed binary patterns shown in Fig. 2. These patterns yield two spectral pairs of horizontally aligned patterns $\mathbf{H}_1, \mathbf{H}_1', \mathbf{H}_2, \mathbf{H}_2'$ and one pair of vertically aligned patterns $\mathbf{V}, \mathbf{V}'$. Similarly to the white pattern captures, the orthogonal horizontal patterns can be rotated to create two sets of X and Y axis-aligned patterns. We then follow a similar approach to Sec. 4.1, but with the following key differences:

Diffuse and Specular Normals We acquire two horizontal pattern sets, using the green channel $\mathbf{H}_1, \mathbf{H}_1'$ and the red channel $\mathbf{H}_2, \mathbf{H}_2'$. Photometric normals from shorter-wavelength channels exhibit sharper normal details [21]. Hence, we use the SVD method (Eq. 1) to estimate the mixed photometric normals $\mathbf{N}$ using the green-channel set $\mathbf{H}_1, \mathbf{H}_1'$ which has substantial specular signal. Additionally, we use SVD to estimate the normals using the red-channel set $\mathbf{H}_2, \mathbf{H}_2'$, which exhibits weaker specular signal, and treat these as the diffuse normals $\mathbf{N}_D$. Finally, we use Eq. 2 to calculate the specular normals given $\mathbf{N}$ and refined $\mathbf{N}_D$.

Diffuse-Specular Separation Similar to white illumination, we employ the linear system of Eq. 3 to estimate the specular albedo $\mathbf{A}_{S_{avg}}$ by averaging the solutions for the X and Y axes. However, unlike the case of white illumination, to estimate the diffuse albedo $\mathbf{A}_D$ we first scale the average $\mathbf{A}_{S_{avg}}$ by the ratio r_c of the respective channel wise solutions of $\mathbf{A}_S$ relative to red channel solution, before subtracting it from full-on image $\mathbf{F}$: $\mathbf{A}_D = \mathbf{F} - r_c * \mathbf{A}_{S_{avg}}$. Here, $\mathbf{F}$ is obtained by adding the two color-multiplexed patterns.

4.3   Specular Roughness Estimation

We further exploit the observations under the color-multiplexed patterns to estimate spatially varying specular roughness. We make the observation that the two color patterns illuminate a subject with four saturated colors and increasing specular roughness blurs these colors to make them less saturated. Hence, we employ normalized color as a novel metric to estimate spatially varying specular roughness using an inverse rendering process. We employ the estimated diffuse albedo $\mathbf{A}_D$ and diffuse normal $\mathbf{N}_D$ to relight the subject under the two color multiplexed lighting conditions and subtract these rendered diffuse components from the photographs to isolate their specular components. We then employ the estimated specular albedo $\mathbf{A}_S$ and specular normal $\mathbf{N}_S$ to render the specular response to these lighting patterns and compare the normalized color of the renderings with various roughness parameters (Cook-Torrance BRDF) to the isolated specular component in the photographs. The estimated specular roughness corresponds to the best matching normalized color (L1 norm). This way, we exploit color as an additional cue to separate specular albedo and roughness.

4.4   Dynamic Capture

The reflectance estimation approach using color-multiplexed patterns (Sec.4.2) is well-suited for dynamic capture. As proof-of-concept, we employ our monitor-based setup for dynamic capture and synchronize a machine vision camera (FLIR Grasshopper 3) with the refresh rate of the monitors. Using this setup, we capture alternating color-multiplexed patterns at 60 fps. The monitors take 16ms to update the pattern on the screens. In order to handle this, we double up the alternating patterns for two frames at a time and use every other frame in the sequence while ignoring the in-between frames corresponding to the screen updates. This provides an effective capture rate of 30 fps for the alternating patterns. Each pattern can be used in conjunction with the next pattern in the sequence for reflectance estimation at the effective capture rate.

4.5   Base Geometry Acquisition

We reconstruct base geometry using multiview capture for our static capture examples. We employ widely used structure-from-motion software COLMAP [24,25] for this purpose and provide it a full-on image $\mathbf{F}$ (computed as sum of a binary pair) from each camera viewpoint. Each camera's processed reflectance and normals are projected on to the reconstructed mesh.

5   Results

5.1   Evaluation

Fig. 3 presents comparison of reflectance and photometric normal maps estimated using our 4-shot method employing white binary illumination (top-left), and maps estimated using our 2-shot method using spectral multiplexing (top-right), vs those estimated using the 6-shot method (bottom-left) and the 2-shot method (bottom-right) proposed by Kampouris et al. [16]. Here, all methods have been implemented on our tablet-based capture setup. Both our 4-shot and 2-shot method achieve high quality results that are quite comparable, while achieving results that are superior to those obtained with the 6-shot and the 2-shot method respectively of [16] for both albedo and photometric normals. We note that the method of [16] was designed for an LED sphere and hence it does not perform as optimally in our setup. Finally, Fig. 11 shows a comparison of our reconstruction with a single-image learning-based method [20].

5.2   Static Capture

Fig. 4 presents qualitative comparisons of acquired reflectance and photometric normal maps for a subject using both our proposed desktop-based capture setups. As can be seen, both setups acquire high-quality data. Fig. 5 presents

(i) $\mathbf{A}_D$ (j) $\mathbf{A}_S$ (k) $\mathbf{N}_D$ (l) $\mathbf{N}_S$ (m) $\mathbf{A}_D$ (n) $\mathbf{A}_S$ (o) $\mathbf{N}_D$ (p) $\mathbf{N}_S$
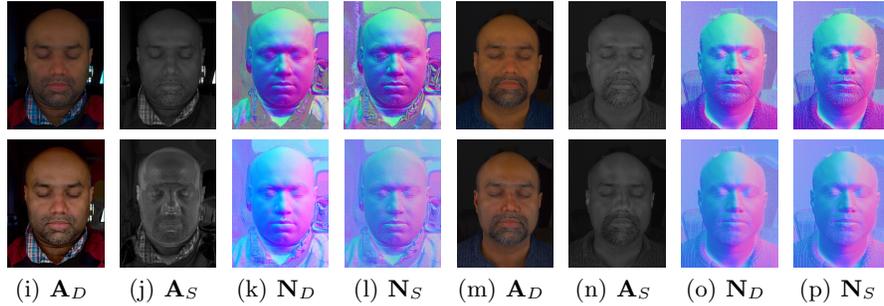
Fig. 3: Comparison of our results (top) with 4-shot white illumination (left), and 2-shot spectral multiplexing (right), against that of [16] (bottom) using 6-shot white illumination (left) and 2-shot spectral multiplexing (right).

maps of spatially varying specular roughness of a few subjects that were estimated using novel color cues given observations of the subjects under the two color multiplexed illumination conditions.

Fig. 9 presents results of several subjects acquired using both of our proposed desktop-based capture setups. Here, we present results of our 2-shot capture process using spectral multiplexing. As can be seen, our estimated diffuse-specular separated reflectance and normals maps (c, d) enable camera-space renderings (b) that are a good qualitative match to the comparison photographs (a). The acquired data can be combined from multiple viewpoints for high quality renderings of 3D geometry (e), and appearance (f). We show multiview data projected to UV maps in Fig. 7. We compare the quantitative error between the validation photos (a) and our camera-space renderings (b) and report an avrage MSE of 0.0019 and PSNR of 27.205 across 7 subjects.

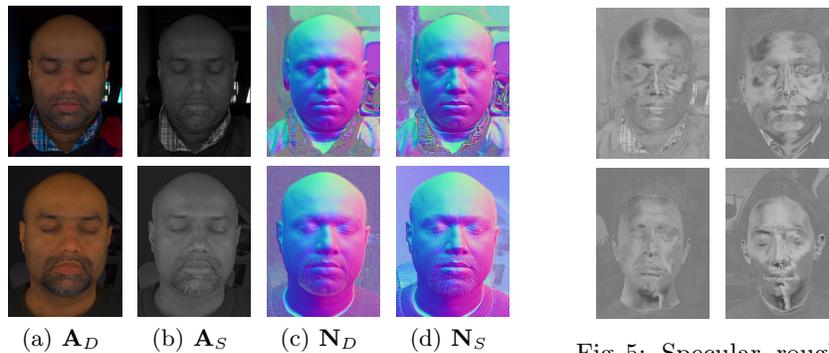(a) $\mathbf{A}_D$ (b) $\mathbf{A}_S$ (c) $\mathbf{N}_D$ (d) $\mathbf{N}_S$

Fig. 4: Comparison of reflectance separation between the Tablet-based setup (top), and the monitor-based setup (down).

Fig. 5: Specular roughness maps, using novel color cues from the color multiplexed patterns.

Fig. 6: Example of dynamic capture using our monitor-based setup, acquired at 60 FPS for results at 30 FPS. From top-to-bottom: diffuse albedo, specular albedo, specular normals and rendering in the Grace Cathedral environment.
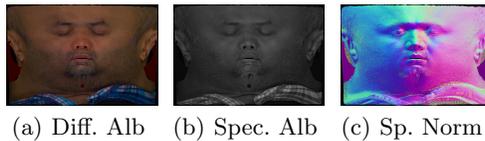
(a) Diff. Alb   (b) Spec. Alb   (c) Sp. Norm

Fig. 7: Complete facial UV maps of our captured reflectance, which shows the multi-view capabilities of our approach.

(a) Un.   (b) Pol.   (c) Ours

Fig. 8: Comparison between unpolarised, cross-polarised [10] and our diffuse albedo result.

5.3   Dynamic Capture

Fig. 6 presents a few frames from a dynamic capture sequence of a female subject acquired at effective 30FPS (every other frame of a 60FPS capture process) using a machine vision camera. As can be seen, the monitor-based setup, in conjunction with our proposed color multiplexed patterns, is well suited for obtaining video-rate diffuse-specular separated albedo maps and photometric normals of a dynamic capture sequence. We provide another example in the supplemental.

5.4   Limitations

Our proposed capture technique, while highly practical, has some limitations. The illumination is restricted to a frontal zone which may not be suitable for all object shapes, and limited illumination coverage can affect estimated photometric maps at the extreme sides. We do not model any single scattering in skin [11]. Finally, dynamic capture rate is limited by monitor refresh rate.

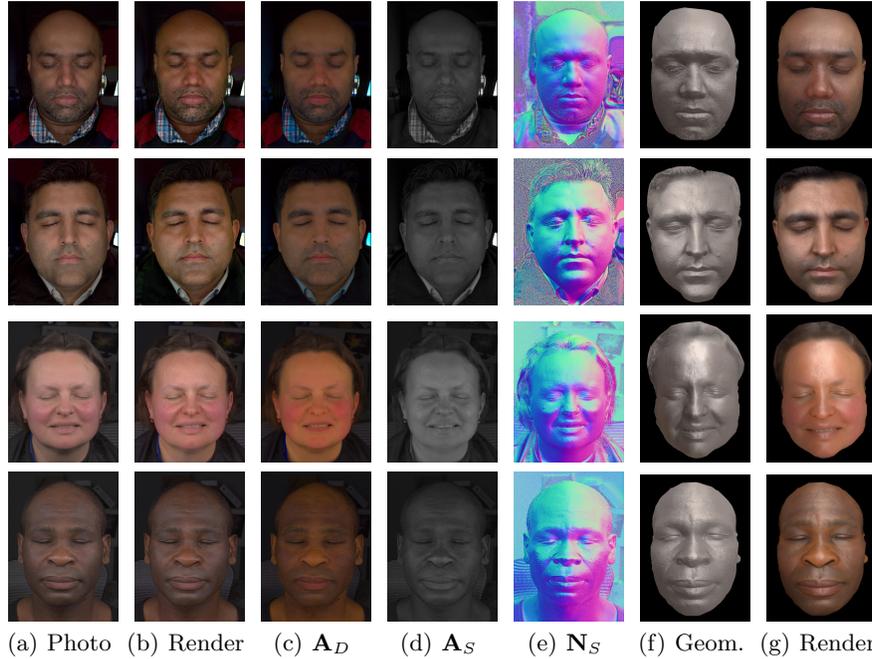(a) Photo (b) Render (c) $\mathbf{A}_D$ (d) $\mathbf{A}_S$ (e) $\mathbf{N}_S$ (f) Geom. (g) Render

Fig. 9: Various subjects acquired with our tablet-based setup (top two), and monitor-based setup (bottom two), captured using our 2-shot method.

6   Conclusions

We present two novel desktop-based setups for high quality facial capture, including reflectance and photometric normals, which are practical and scalable, consisting entirely of commodity components. We also present a novel analysis of binary illumination together with our setup, for efficient acquisition with reduced measurements. We achieve high-quality diffuse-specular separation with spectral multiplexing, exploit novel color cues for estimating specular roughness, and extend the two-shot capture for video-rate dynamic capture. Our proposed systems can make high-quality facial capture widely accessible for many applications.
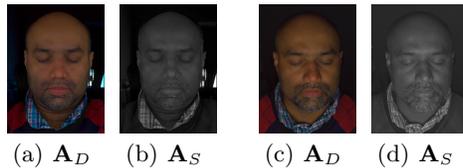

(a) $\mathbf{A}_D$ (b) $\mathbf{A}_S$ (c) $\mathbf{A}_D$ (d) $\mathbf{A}_S$

Fig. 10: Comparison of albedo measurement and separation with our method using our setup (left), vs using a single screen (right).
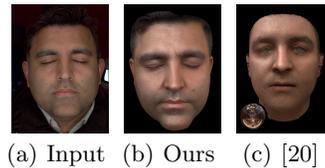

(a) Input (b) Ours (c) [20]

Fig. 11: Our results compared with AvatarMe[++], a deep-learning method [20].

References

1. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Transactions on Graphics (TOG) **29**(3), 40:1–40:9 (2010)
2. 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. ACM Transactions on Graphics (ACM) **30**, 75:1–75:10 (August 2011)
3. Bradley, D., Heidrich, W., Popa, T., Sheffer, A.: High resolution passive facial performance capture. ACM Transactions on Graphics (TOG) **29**(4), 41 (2010)
4. 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. pp. 145–156 (2000)
5. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
6. Fyffe, G., Debevec, P.: Single-shot reflectance measurement from polarized color gradient illumination. In: International Conference on Computational Photography (ICCP). IEEE (2015)
7. Fyffe, G., Graham, P., Tunwattanapong, B., Ghosh, A., Debevec, P.: Near-Instant Capture of High-Resolution Facial Geometry and Reflectance. Computer Graphics Forum (2016)
8. Fyffe, G., Hawkins, T., Watts, C., Ma, W.C., Debevec, P.: Comprehensive facial performance capture. Computer Graphics Forum (CGF) **30**(2) (2011)
9. Gecer, B., Lattas, A., Ploumpis, S., Deng, J., Papaioannou, A., Moschoglou, S., Zafeiriou, S.: Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In: European conference on computer vision. pp. 415–433. Springer (2020)
10. Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., Debevec, P.: Multi-view face capture using polarized spherical gradient illumination. ACM TOG **30**(6) (2011)
11. Ghosh, A., Hawkins, T., Peers, P., Frederiksen, S., Debevec, P.: Practical modeling and acquisition of layered facial reflectance. ACM TOG **27**(5) (Dec 2008)
12. Gotardo, P., Riviere, J., Bradley, D., Ghosh, A., Beeler, T.: Practical dynamic facial appearance modeling and acquisition. ACM Trans. Graph. **37**(6) (Dec 2018)
13. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM Transactions on Graphics (ToG) **38**(6), 1–19 (2019)
14. Hernandez, C., Vogiatzis, G., Brostow, G.J., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2007)
15. Jiang, J., Liu, D., Gu, J., Süsstrunk, S.: What is the space of spectral sensitivity functions for digital color cameras? In: 2013 IEEE Workshop on Applications of Computer Vision (WACV). pp. 168–179. IEEE (2013)
16. Kampouris, C., Zafeiriou, S., Ghosh, A.: Diffuse-specular separation using binary spherical gradient illumination. In: EGSR (EI&I). pp. 1–10 (2018)
17. Klaudiny, M., Hilton, A.: High-detail 3d capture and non-sequential alignment of facial performance. In: Proceedings of 3DIMPVT (2012)
18. Klehm, O., Rousselle, F., Papas, M., Bradley, D., Hery, C., Bickel, B., Jarosz, W., Beeler, T.: Recent advances in facial appearance capture. Computer Graphics Forum (CGF) **34**(2), 709–733 (May 2015)

19. Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S.: Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 760–769 (2020)
20. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Ghosh, A., Zafeiriou, S.P.: Avatarme++: Facial shape and BRDF inference with rendering-aware GANs. TPAMI (2021)
21. Ma, W.C., Hawkins, T., Peers, P., Chabert, C.F., Weiss, M., Debevec, P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In: Proc. EGSR (2007)
22. Riviere, J., Gotardo, P., Bradley, D., Ghosh, A., Beeler, T.: Single-shot high-quality facial geometry and skin appearance capture. ACM Trans. Graph. **39**(4) (Jul 2020)
23. Schonberger, J.L., Frahm, J.: Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113. IEEE Computer Society, Los Alamitos, CA, USA (jun 2016)
24. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
25. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
26. Sengupta, S., Curless, B., Kemelmacher-Shlizerman, I., Seitz, S.M.: A light stage on every desk. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2420–2429 (2021)
27. Smith, W.A., Seck, A., Dee, H., Tiddeman, B., Tenenbaum, J.B., Egger, B.: A morphable face albedo model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5011–5020 (2020)
28. Weyrich, T., Lawrence, J., Lensch, H., Rusinkiewicz, S., Zickler, T.: Principles of appearance acquisition and representation. Foundations and Trends in Computer Graphics and Vision **4**(2), 75–191 (2008)
29. Weyrich, T., Matusik, W., Pfister, H., Bickel, B., Donner, C., Tu, C., McAndless, J., Lee, J., Ngan, A., Jensen, H.W., Gross, M.: Analysis of human faces using a measurement-based skin reflectance model. ACM Trans. Graphics (TOG) **25**(3), 1013–1024 (Jul 2006)
30. Wilson, C.A., Ghosh, A., Peers, P., Chiang, J.Y., Busch, J., Debevec, P.: Temporal upsampling of performance geometry using photometric alignment. ACM Transactions on Graphics (TOG) **29**(2), 17 (2010)
31. Yuille, A.L., Snow, D., Epstein, R., Belhumeur, P.N.: Determining generative models of objects under varying illumination: Shape and albedo from multiple images using svd and integrability. International Journal of Computer Vision **35**(3), 203–222 (1999)