

# 3D Face Reconstruction using Color Photometric Stereo with Uncalibrated Near Point Lights

Zhang Chen, Yu Ji, Mingyuan Zhou, Sing Bing Kang, *Fellow, IEEE*, and Jingyi Yu, *Member, IEEE*

**Abstract**—We present a new color photometric stereo (GPS) method that recovers high quality, detailed 3D face geometry in a single shot. Our system uses three uncalibrated near point lights of different colors and a single camera. For robust self-calibration of the light sources, we use 3D morphable model (3DMM) [1] and semantic segmentation of facial parts. For reconstruction, we address the inherent spectral ambiguity in color photometric stereo by incorporating albedo consensus, albedo similarity, and proxy prior into a unified framework. In this way, we jointly exploit multiple cues to resolve under-determinedness, without the need for spatial constancy of albedo. Experiments show that our new approach produces state-of-the-art results from single image with high-fidelity geometry that includes details such as wrinkles.

**Index Terms**—Color photometric stereo, 3D face reconstruction, uncalibrated near point lights, single shot capture, normal estimation.



## 1 INTRODUCTION

STATE-of-the-art photometric stereo solutions for 3D face reconstruction [2], [3], [4], [5] are capable of producing movie-quality, photo-realistic results. However, these systems tend to be bulky and expensive and generally require taking multiple shots. Even with elaborate time-multiplexing, it is difficult to capture fine facial geometry movements unless using an ultra-fast speed camera coupled with high precision synchronized light sources. The light sources and cameras also require accurate calibration to avoid distortions in the final reconstruction.

In this paper, we present a novel lightweight one-shot solution based on uncalibrated color photometric stereo method that simply uses a camera and three uncalibrated near point light sources of different color. Our approach eliminates the need of time multiplexing, and therefore can be used to recover dynamic facial motions. Compared with distant light sources which require relatively strong power, the use of near point light sources makes the system more portable by reducing the cost and space requirement. However, for near-field lighting, one needs to know the relative positions between light sources and face geometry. Even with light positions calibrated using special calibration targets (e.g., sphere and planar light probes), one still require extra depth information of the captured object. We instead propose a self-calibration method exploiting the shape prior of human faces encoded in 3D morphable model (3DMM) [1] and can directly self-calibrate the

relative positions between light sources and face geometry with a single image.

For objects with non-gray albedo, color photometric stereo is inherently under-determined due to spectral inconsistencies of surface reflectance: albedo is not identical under different spectra and therefore there are more unknown variables than there are constraints. We address the spectral ambiguity problem by proposing albedo similarity and proxy prior, and incorporating them with albedo consensus into a unified framework. As a result, our approach does not need to assume spatial constancy of albedo. We also present a new measure for albedo similarity based on the albedo norm profile. The proposed albedo similarity and proxy prior effectively correct distortions caused by incorrect albedo consensus in prior work. Experiments show that our new approach can produce state-of-the-art results from single image with high-fidelity geometry that includes details such as wrinkles.

Our technical contributions are as follow:

- A self-calibration method utilizing 3DMM proxy face for color photometric stereo with near point lights.
- A per-pixel formulation for solving normal and albedo from color photometric stereo.
- A framework that incorporates albedo similarity and proxy prior with albedo consensus to produce accurate 3D reconstruction.

## 2 RELATED WORK

Structured light [6], [7] and multi-view stereo [8] have been used to reconstruct faces. While they can accurately reconstruct coarse shapes, they are less successful in recovering high frequency details such as wrinkles. On the other hand, photometric stereo [9] is capable of recovering high frequency details. Techniques that combine stereo and photometric stereo exist [5], [10], [11], but the combination is at the expense of a complicated hardware setup. Recently, Gotardo *et al.* [12] achieves high-quality dynamic face reconstruction with multi-view stereo and constant white

- Z. Chen is with School of Information Science and Technology, ShanghaiTech University, Shanghai, China.  
E-mail: chenzhang@shanghaitech.edu.cn
- Z. Chen is also with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China.
- Z. Chen is also with University of Chinese Academy of Sciences, China.
- Y. Ji and M. Zhou are with DGene Inc., Baton Rouge, LA, USA.  
E-mail: yu.ji@dgene.com, mingyuan.zhou@dgene.com
- S.B. Kang is with Zillow Group, Seattle, WA, USA.  
E-mail: singbingk@zillowgroup.com
- J. Yu is with School of Information Science and Technology, ShanghaiTech University, Shanghai, China.  
E-mail: yujingyi@shanghaitech.edu.cn

lights through an inverse rendering framework. However, they still require careful geometric and photometric calibration as well as HDR light probe of the surrounding environment.

## 2.1 Photometric Stereo (PS)

Traditional PS [9] uses 3 or more distant lights (of the same color) and sequentially creates different directional illumination by turning on only one light at a time. A sequence of images is captured, each with a different light source. The surface orientation map can then be inferred from image intensities using an over-determined linear system. Normal integration is then applied to obtain a 2.5D reconstruction. We refer readers to [13], [14] for a comprehensive review of classical PS methods. The distant light requirement has since been relaxed; much work has been done using more practical near point light sources [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. Notably, Liu *et al.* [25] use an LED ring with a radius of only  $30mm$  centered at camera lens. Alternative self-calibrating methods [26], [27], [28], [29], [30], [31] provide simpler and more flexible solutions under various assumptions [32]. It is also possible to use uncalibrated near point light sources [33], [34], [35], [36], but they all require sequential capture.

## 2.2 Color Photometric Stereo (CPS)

CPS has the key benefit of acquiring only one image and hence can be directly used to reconstruct dynamic objects. Most existing approaches use red, green, and blue lights along with a color camera [37], [38], [39]. Hernández *et al.* [40] apply such a technique to dynamic cloth reconstruction; they use a planar board with cloth sample fixed in the center to calibrate the coupled matrix containing reflectance, camera response, lighting spectrum, and lighting directions. Vogiatzis and Hernández [41] first construct a coarse 3D face using structure from motion and then impose the constant chromaticity constraint for shape refinement. Kludiny *et al.* [42] use a specular sphere to estimate lighting directions. To ensure constant chromaticity, they apply uniform make-up to faces. Bringier *et al.* [43] explicitly calibrate the spectral response of camera and assume gray color or known uniform color.

To eliminate the need of constant chromaticity, there are methods [44], [45] that combine spectral and time-multiplexing; optical flow is then used to align adjacent frames. Jankó *et al.* [46] make use of temporal constancy of surface reflectance to eliminate the need for time-multiplexing, but an image sequence is still required as input. Gotardo *et al.* [11] simultaneously solve for color photometric stereo, optical flow, and stereo matching within each 3-frame time window, but require 9 color lights. Rahman *et al.* [47] arrange complementary color lights on a ring. Their approach requires using 2 images under complementary illuminations as input. Anderson *et al.* [48] assume piecewise constant chromaticity by segmenting a scene into different chromaticities. To calibrate chromaticities, they also require a stereo camera pair to obtain coarse geometry.

Fyffe *et al.* [49] extend the usual 3 color channels to 6 by using 2 RGB cameras and a pair of Dolby dichroic filters. An extension of their work [50] employ polarized color gradient illumination but require a complex setup with 2040 LED light sources. Chakrabarti and Sunkavalli [51] observe that the reflectance and normal within a uniform color region can be uniquely recovered from spectrally demultiplexed image by assuming piecewise constant albedo. Ozawa *et al.* [52] densely discretize albedo chromaticity

and enforce consensus on albedo norms to reconstruct objects with spatially-varying albedo. However, most of these approaches assume directional lighting and require pre-calibrating them. It is possible to use near light sources [53], but they still require pre-calibration. In contrast, our technique focuses on face reconstruction and exploits prior face information to enable self-calibration of near point lights. We assume unknown light positions and spatially-varying albedo. The former enables more feasible capture while the latter fulfils the physical property of real faces.

## 2.3 Single Image Techniques

There are methods for inferring face geometry from a single unconstrained image; see [54] for an overview of state-of-the-art methods. However, they tend to produce less accurate results compared with multi-view stereo and photometric stereo. Piontraschke and Blanz [55] demonstrate the usefulness of semantic segmentation to improve reconstruction quality. In our work, we use the 3D morphable model [1] to obtain an initial proxy face for light source calibration.

Shape-from-shading and deep learning based approaches have also been adopted to recover details [56], [57], [58], [59], [60], [61], [62], [63]. Jiang *et al.* [64] combined local corrective deformation fields with photometric consistency constraints. Yamaguchi *et al.* [65] use a large corpus of high-fidelity face captures from the USC Light Stage [10] to learn the mapping from texture to highly-detailed displacement map. These solutions can provide visually pleasing results but their accuracy is heavily dependent on illumination.

## 3 COLOR PHOTOMETRIC STEREO WITH NEAR POINT LIGHTS

Traditional color photometric stereo uses 3 distant lights with different lighting directions and spectrum (usually red, green and blue) together with an RGB camera to spectrally multiplex different illumination in a single image. By assuming distant lights, each surface point is illuminated by three directional lights with direction  $\mathbf{l}_j \in \mathbb{R}^3$  and spectral distribution  $\mathcal{E}_j(\lambda)$ , where  $j = 1, 2, 3$  and  $\lambda$  is the wavelength. We denote the normal and reflectance function at any pixel  $(x, y)$  as  $\mathbf{n}(x, y)$  and  $\mathcal{R}(x, y, \lambda)$ , respectively. Let  $\mathcal{S}_i(\lambda)$  with  $i = 1, 2, 3$  be the spectral response of each camera color channel. For a Lambertian surface, the image pixel intensity  $c_i(x, y)$  can be expressed as

$$c_i(x, y) = \sum_j \mathbf{l}_j^\top \mathbf{n}(x, y) \int \mathcal{S}_i(\lambda) \mathcal{R}(x, y, \lambda) \mathcal{E}_j(\lambda) d\lambda. \quad (1)$$

We denote  $\mathbf{A}(x, y) \in \mathbb{R}^{3 \times 3}$  as the albedo matrix whose element at  $i$ th row and  $j$ th column is

$$\mathbf{A}_{i,j}(x, y) = \int \mathcal{S}_i(\lambda) \mathcal{R}(x, y, \lambda) \mathcal{E}_j(\lambda) d\lambda. \quad (2)$$

Each element of  $\mathbf{A}(x, y)$  thus represents the albedo under one light-channel pair. Letting  $\mathbf{c} = [c_1, c_2, c_3]^\top$  and  $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]^\top$ , we can rewrite Eq. 1 in matrix form as

$$\mathbf{c}(x, y) = \mathbf{A}(x, y) \mathbf{L}(x, y) \mathbf{n}(x, y). \quad (3)$$

Note that for distant lights,  $\mathbf{L}$  is identical for all pixels. As a result, with initial coarse normal  $\mathbf{n}'$ , one can self-calibrate the product of  $\mathbf{A}$  and  $\mathbf{L}$  by assuming constant albedo or constant chromaticity [41]. However, for near point lights, lighting direction is

spatially-varying. By further taking into account the inverse square illumination attenuation due to distance, we obtain

$$I_j(x, y) = \frac{\mathbf{p}_j - \mathbf{v}(x, y)}{\|\mathbf{p}_j - \mathbf{v}(x, y)\|_2^3}, \quad (4)$$

where  $\mathbf{p}_j$  is the 3D position of  $j$ th light source and  $\mathbf{v}(x, y)$  is the corresponding 3D position for pixel at  $(x, y)$ .

#### 4 NEAR POINT LIGHT SELF-CALIBRATION

The benefits of self-calibration are two-fold: first, it eliminates the need for a special calibration target (e.g., sphere and planar light probes) and the laborious procedure usually involved when calibrating near point lights; second, it can handle unexpected movements of hardware devices (e.g., light sources), making the capture process more robust. To the best of our knowledge, our work is the first to address self-calibration of near point lights under color photometric stereo. For traditional photometric stereo with near point lights of same color, numerous self-calibration methods exist [33], [34], [35], [36], but these methods are not directly applicable due to more unknowns in color photometric stereo.

In order to self-calibrate the light source positions, we first require a coarse proxy mesh, from which we obtain initial rough estimates for normal  $\mathbf{n}$  and position  $\mathbf{v}$  at every pixel  $(x, y)$ . Unlike other methods that use multi-view stereo [41] or stereo matching [48] to obtain the proxy mesh, our approach makes use of the 3D morphable model (3DMM) [1] and needs only one image as input. To compensate for the inaccuracies in the proxy mesh, we use RANSAC followed by hypothesis merging to robustly estimate light source positions. We provide details of our method in the following two sections.

Most relevant to our work, Cao *et al.* [35] also exploit 3DMM for self-calibration. However, a significant difference with our work is that they resolve ill-posedness by jointly solving for all lights and require the albedo of a pixel to be identical under each light. This assumption no longer holds for color photometric stereo due to spectral inconsistencies of surface reflectance, as shown in Eq. 2. In contrast to [35], we propose a RANSAC-based approach in this paper.

##### 4.1 Proxy Mesh Generation

3DMM is a deformable template for the mesh of a human face. It consists of Principal Component Analysis (PCA) linear basis along three dimensions: shape, expression, and albedo. Since we are concerned with only shape and expression associated with the proxy mesh, we omit the albedo dimension. 3DMM interprets the face mesh  $\mathbf{m} \in \mathbb{R}^{3n}$  as a linear combination of shape and expression bases:

$$\mathbf{m} = \mathbf{a}_s + \mathbf{a}_e + \sum_i \alpha_i \mathbf{b}_i^s + \sum_i \beta_i \mathbf{b}_i^e, \quad (5)$$

where  $\mathbf{a}_s, \mathbf{a}_e \in \mathbb{R}^{3n}$  are PCA means and  $\mathbf{b}_i^s, \mathbf{b}_i^e \in \mathbb{R}^{3n}$  are  $i$ th PCA bases of shape and expression, respectively.  $n$  is the number of mesh vertices, and  $\alpha_i, \beta_i$  are  $i$ th coefficients for linear combination of the bases. We adopt the *Basel Face Model 2017* [1] for 3DMM, and use the iterative linear method from [66] to jointly solve for PCA coefficients and camera parameters (intrinsics and extrinsics). We then rasterize the generated proxy mesh to recover initial normal and 3D position for each pixel. While the proxy mesh resembles a human face with a reasonable pose, its geometry is usually inaccurate.

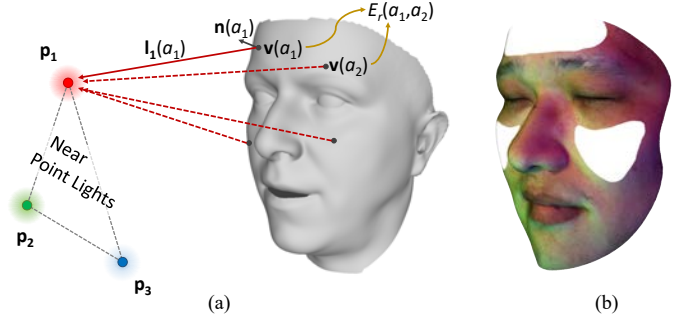


Fig. 1. Self-calibration of near point light positions using a proxy face. (a) Parameters involved in estimating  $\mathbf{p}_1$ . (b) Regions (white) on the face used for RANSAC pixel sampling.

##### 4.2 Estimation of Light Source Positions

As with [51], [52], we assume that there is no crosstalk between light sources and camera channels, i.e., the spectrum of each light source can only be observed in its corresponding camera channel. As a result, the albedo matrix  $\mathbf{A}(x, y)$  is diagonal. For simplicity, let  $\boldsymbol{\rho}(x, y) = [\mathbf{A}_{1,1}(x, y), \mathbf{A}_{2,2}(x, y), \mathbf{A}_{3,3}(x, y)]^\top$ , Eq. 3 then becomes

$$\mathbf{c}(x, y) = \boldsymbol{\rho}(x, y) \odot \mathbf{L}(x, y) \mathbf{n}(x, y), \quad (6)$$

where  $\odot$  is the Hadamard product operator. For two pixels  $(x_1, y_1), (x_2, y_2)$  with equal albedo in the  $i$ th channel, i.e.,  $\rho_i(x_1, y_1) = \rho_i(x_2, y_2)$ , we have

$$\frac{\mathbf{c}_i(x_1, y_1)}{\mathbf{L}_i(x_1, y_1) \mathbf{n}(x_1, y_1)} = \frac{\mathbf{c}_i(x_2, y_2)}{\mathbf{L}_i(x_2, y_2) \mathbf{n}(x_2, y_2)}, \quad (7)$$

where  $\mathbf{L}_i$  is the  $i$ th row of  $\mathbf{L}$ , representing the lighting direction of  $i$ th light source. Substituting Eq. 4 into Eq. 7 and moving all variables to the left hand side, we obtain

$$\frac{\mathbf{c}_i(x_1, y_1) \|\mathbf{p}_i - \mathbf{v}(x_1, y_1)\|_2^3}{(\mathbf{p}_i - \mathbf{v}(x_1, y_1)) \mathbf{n}(x_1, y_1)} - \frac{\mathbf{c}_i(x_2, y_2) \|\mathbf{p}_i - \mathbf{v}(x_2, y_2)\|_2^3}{(\mathbf{p}_i - \mathbf{v}(x_2, y_2)) \mathbf{n}(x_2, y_2)} = 0. \quad (8)$$

Once  $\mathbf{n}$  and  $\mathbf{v}$  are extracted from proxy mesh, we can now recover  $\mathbf{p}_i$  (in the same coordinate system as proxy mesh), which has 3 unknowns. We require at least 3 constraints, which means a minimum of 4 pixels with equal albedo in the  $i$ th channel. Since there is no correlation between different lights or channels in Eq. 8, we can estimate the position of each light independently. However, since the albedo is unknown, we cannot deterministically locate pixels with equal albedo. Our solution is to employ RANSAC to randomly sample quadruplets of pixels. Since we only require each sampled quadruplet to have equal albedo in one channel, there is still a high probability that at least one sampling provides a qualified quadruplet.

Notice that in Eq. 8, the numerators have a higher order of distance between light source and surface point than those in the denominators. This biases the solution towards closer light

positions. We instead use an unbiased form of Eq. 8 to measure the residual between two pixels  $a_1, a_2$ :

$$E_r(a_1, a_2) = \frac{\mathbf{c}_i(a_1)(\mathbf{p}_i - \mathbf{v}(a_2))\mathbf{n}(a_2) \|\mathbf{p}_i - \mathbf{v}(a_1)\|_2}{\|\mathbf{p}_i - \mathbf{v}(a_2)\|_2^2} - \frac{\mathbf{c}_i(a_2)(\mathbf{p}_i - \mathbf{v}(a_1))\mathbf{n}(a_1) \|\mathbf{p}_i - \mathbf{v}(a_2)\|_2}{\|\mathbf{p}_i - \mathbf{v}(a_1)\|_2^2}. \quad (9)$$

For each quadruplet  $\mathcal{Q}$  (an example is shown in Fig. 1(a)), a hypothesis of the light position is computed by solving

$$\min_{\mathbf{p}_i} \sum_{a_k \in \mathcal{Q}} \sum_{a_l \in \mathcal{Q} - a_k} (E_r(a_k, a_l))^2, \quad (10)$$

which is a squared sum of residuals between each pair of pixels in a quadruplet. We use the Levenberg-Marquardt algorithm to solve the nonlinear optimization.

In voting for a hypothesis, a pixel  $a_w$  is considered an inlier if the squared sum of residuals between it and the pixels in  $\mathcal{Q}$  satisfies

$$\sum_{a_k \in \mathcal{Q}} (E_r(a_k, a_w))^2 < \tau^2, \quad (11)$$

where  $\tau$  is a threshold and set as 0.01 in our experiments.

Instead of using all pixels for sampling and voting, we only use the pixels on left cheek, right cheek, and forehead, as shown in Fig. 1(b). This is to avoid potential highly non-Lambertian regions such as facial hair and shadows. The segmentation of these regions only needs to be done once on a 3DMM mean face, which can then be projected to different face images [35].

Unlike standard RANSAC which chooses the hypothesis with the most number of inliers as the final estimate, we perform an additional filtering and merging process on all the hypotheses. The reason is that the 3DMM-based proxy mesh is inaccurate even as low-frequency geometry. As a result, the initial normals deviate from true normals at most pixels, making consensus less concentrated and potentially drifting away from the correct hypothesis. Instead, we take a set of hypotheses into account to produce a more robust estimate.

In the filtering step, we determine a plausible region for hypotheses and ignore all hypotheses outside this region. We first use the four-point algorithm in [41] to produce the calibration matrix, which is the product of dominant albedo and directional lighting directions. We then factor out the dominant albedo and extract lighting direction  $\mathbf{l}'_i$  for each light by normalizing each row of the calibration matrix. Hypothesis  $\mathbf{p}'_i$  (for the  $i$ th light source position) is dropped if it does not satisfy

$$\arccos \frac{(\mathbf{p}'_i - \mathbf{v}_c)^\top \mathbf{l}'_i}{\|\mathbf{p}'_i - \mathbf{v}_c\|_2} < \eta, \quad (12)$$

where  $\mathbf{v}_c$  is the mean 3D position of all pixels.

Eq. 12 forms a cone region with half-angle  $\eta$  around  $\mathbf{l}'_i$ ; all hypotheses outside this region are ignored. We use  $\eta = 15^\circ$  in our experiments. Subsequently, we merge the remaining hypotheses  $\mathcal{P}_i$  with weighted linear combination to obtain final estimate for a light source position:

$$\mathbf{p}_i = \frac{\sum_{\mathbf{p}'_i \in \mathcal{P}_i} w(\mathbf{p}'_i) \mathbf{p}'_i}{\sum_{\mathbf{p}'_i \in \mathcal{P}_i} w(\mathbf{p}'_i)}, \quad (13)$$

where  $w(\mathbf{p}'_i)$  is the number of inliers for hypothesis  $\mathbf{p}'_i$ .

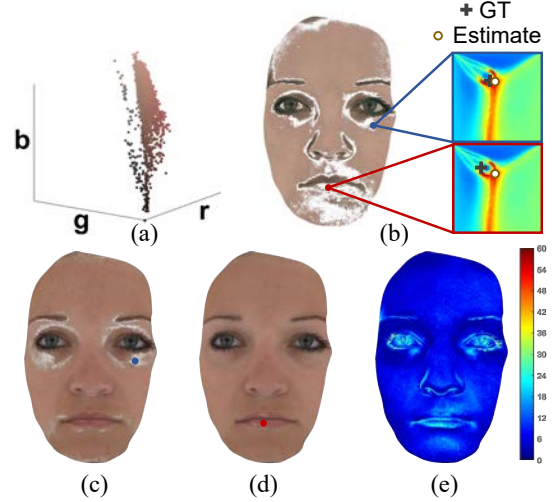


Fig. 2. Effect of consensus term, illustrated on a face with ground truth. (a) Albedo distribution. (b) Two pixels that contribute to a consensus. The close-ups show the magnitude of negative consensus term at the two pixels in chromaticity space. The skin pixel is accurately estimated while the lip pixel is not. (c, d) Distribution of ground truth pixels that form consensus with a given pixel (indicated as blue and red dots, respectively). (e) Normal error map when using only the consensus term.

## 5 FACE RECONSTRUCTION

Once the light source positions have been determined, we can obtain per-pixel lighting directions  $\mathbf{L}(x, y)$  using light positions and proxy face. We then set out to estimate per-pixel photometric normal. With the albedo unknown, this problem is pixel-wise underdetermined (from Eq. 6). This is because there are 5 degrees of freedom (3 for albedo and 2 for normal) but only 3 constraints. It has been shown [51], [52] that 3 pixels with equal albedo and linearly independent normals can uniquely determine the albedo and normals at these pixels. To exploit this property, Chakrabarti and Sunkavalli [51] modeled the albedo as being piece-wise constant and used a polynomial model for surface depth. However, their method tends to produce overly-smoothed results. On the other hand, Ozawa *et al.* [52] developed an iterative voting scheme based on consensus of albedo norms under albedo chromaticity candidates (see Sec. 5.1) to simultaneously classify pixels into different albedos and compute their normals. Since their method assumes no spatial constancy on albedo, high-frequency details can be recovered. In the extreme case where all pixels share the same albedo, the correct albedo chromaticity can be estimated by finding the one that produces the strongest consensus on the albedo norm. However, for a multi-colored surface, their method may produce albedo consensus that leads to incorrect estimation for some pixels. This is because a pixel can be interpreted by any albedo chromaticity and corresponding albedo norm. There may exist situations where, under consensus albedo chromaticity, a pixel with a different albedo has a similar albedo norm with consensus. For human faces, the albedo distribution tends to spread out instead of being of a single albedo, as shown in Fig. 2a. Consensus usually arrives at a reasonable estimation for major clusters because the number of inliers tends to be large, which improves robustness. The skin pixel at the blue dot in Fig. 2(b, c) shows an example. On the other hand, for minor clusters, consensus tends to provide unreliable estimation as shown by the

red dot in Fig. 2(b, d), where the lip pixel can be better interpreted by an incorrect albedo chromaticity.

By comparison, we propose a pixel-wise formulation which incorporates albedo consensus, albedo similarity between pixels as well as proxy mesh for high-quality reconstruction. From Eq. 6, we can decompose albedo  $\boldsymbol{\rho}$  into albedo chromaticity  $\hat{\boldsymbol{\rho}}$  and albedo norm  $\tilde{\rho}$ :

$$\begin{aligned} \mathbf{c}(x, y) &= \hat{\boldsymbol{\rho}}(x, y) \odot \mathbf{L}(x, y)(\tilde{\rho}(x, y)\mathbf{n}(x, y)), \\ \tilde{\rho}(x, y)\mathbf{n}(x, y) &= \mathbf{L}(x, y)^{-1}(\mathbf{c}(x, y) \oslash \hat{\boldsymbol{\rho}}(x, y)), \end{aligned} \quad (14)$$

where  $\oslash$  is the Hadamard division operator. We only need to solve for albedo chromaticity because albedo norm and normal can then be trivially computed.

To make the problem more tractable, as with [51], [52], we discretize albedo chromaticity in the space of positive unit sphere  $\mathcal{S}_+^2$  into candidates  $\mathcal{C} = \{\hat{\boldsymbol{\rho}}^{(1)}, \hat{\boldsymbol{\rho}}^{(2)}, \dots\}$ . Then, for each pixel  $a_i$ , we solve for its albedo chromaticity using

$$\begin{aligned} \hat{\boldsymbol{\rho}}(a_i) &= \arg \min_{\hat{\boldsymbol{\rho}} \in \mathcal{C}} E_c(a_i, \hat{\boldsymbol{\rho}}) + \lambda_s w_s(a_i) E_s(a_i, \hat{\boldsymbol{\rho}}) \\ &\quad + \lambda_p w_p(a_i) E_p(a_i, \hat{\boldsymbol{\rho}}), \end{aligned} \quad (15)$$

where  $E_c$  is the albedo consensus term (Sec. 5.1),  $E_s$  the albedo similarity term (Sec. 5.2), and  $E_p$  the proxy prior term (Sec. 5.3).  $w_s(a_i)$  and  $w_p(a_i)$  modulate the influence of similarity term and proxy term at different pixels. After solving for albedo chromaticity at each pixel, we can then compute the normal and use Poisson integration to obtain geometry. Compared with proxy mesh, our final reconstruction is more accurate for both macro- (shape, expression) and micro- (wrinkles, etc.) geometries. We detail each term in the following three sections.

## 5.1 Albedo Consensus

Albedo consensus measures the number of pixels that have similar albedo norm under an albedo chromaticity candidate [52]. To compute the consensus term, for each albedo chromaticity candidate  $\hat{\boldsymbol{\rho}}^{(j)}$ , we find the corresponding albedo norms of all pixels  $\mathcal{N}^{(j)} = \{\tilde{\rho}^{(j)}(a_1), \tilde{\rho}^{(j)}(a_2), \dots\}$  and build a histogram with bin width  $\delta_b \cdot \text{median}(\mathcal{N}^{(j)})$  [52]. Let  $\mathcal{B}^{(j,k)}$  be the  $k$ th bin under  $\hat{\boldsymbol{\rho}}^{(j)}$ ,  $|\mathcal{B}^{(j,k)}|$  its cardinality, and  $b_{i,j}$  the index for the bin that contains the albedo norm of pixel  $a_i$  under  $\hat{\boldsymbol{\rho}}^{(j)}$ . We define

$$E_c(a_i, \hat{\boldsymbol{\rho}}^{(j)}) = \frac{m - |\mathcal{B}^{(j,b_{i,j})}|}{m}, \quad (16)$$

where  $m$  is the total number of pixels. However, it should be noted that pixels of different albedo may also have similar albedo norm under incorrect albedo chromaticities. We propose using *albedo similarity* and *proxy prior* to handle this problem.

## 5.2 Albedo Similarity

Directly inferring albedo similarity from image intensity is error-prone, since the difference in image intensity can be caused by either albedo or shading or both. Instead, the albedo norms of a pixel under all albedo chromaticities form an albedo norm profile. We reason that if two pixels have similar albedo norm profile, then they are likely to have similar albedos. From Eq. 14, letting  $\mathbf{H} = [\mathbf{c}_1 \mathbf{L}_{:1}^{-1}, \mathbf{c}_2 \mathbf{L}_{:2}^{-1}, \mathbf{c}_3 \mathbf{L}_{:3}^{-1}]$  (where  $\mathbf{L}_{:i}^{-1}$  is the  $i$ th column of  $\mathbf{L}^{-1}$ ) and  $\hat{\boldsymbol{\rho}}^j = [\frac{1}{\hat{\rho}_1}, \frac{1}{\hat{\rho}_2}, \frac{1}{\hat{\rho}_3}]^T$ , we have

$$\tilde{\rho}(x, y)\mathbf{n}(x, y) = \mathbf{H}(x, y)\hat{\boldsymbol{\rho}}^j(x, y). \quad (17)$$

The albedo norm profile of a pixel is controlled by  $\mathbf{H}$ . Hence, we measure the similarity between two pixels as

$$M(a_1, a_2) = -\|\mathbf{H}(a_1) - \mathbf{H}(a_2)\|_F, \quad (18)$$

where  $\|\cdot\|_F$  is the Frobenius norm. The albedo similarity term is then computed as

$$E_s(a_i, \hat{\boldsymbol{\rho}}^{(j)}) = \frac{1}{|\mathcal{B}^{(j,b_{i,j})}|} \sum_{a \in \mathcal{B}^{(j,b_{i,j})}} -M(a_i, a), \quad (19)$$

which is the mean similarity between a pixel and its same-bin pixels under the  $j$ th albedo chromaticity candidate.

We further multiply a per-pixel weight to the similarity term to suppress its effect at pixels where the similarity term is large for all albedo chromaticity candidates. More specifically, we compute the weight as

$$w_s(a_i) = e^{-(\min(E_s(a_i, :)) - \min(E_s(:, :)))^2 / \sigma_s^2}. \quad (20)$$

## 5.3 Proxy Prior

The proxy albedo chromaticity map can be computed from the proxy mesh using Eq. 6 and is used to penalize implausible estimations produced by the consensus term. The proxy term is expressed as

$$E_p(a_i, \hat{\boldsymbol{\rho}}^j) = 1 - \hat{\boldsymbol{\rho}}_p(a_i)^T \hat{\boldsymbol{\rho}}^j, \quad (21)$$

where  $\hat{\boldsymbol{\rho}}_p(a_i)$  is the proxy albedo chromaticity at pixel  $a_i$ . We apply this term only to pixels where the consensus term gives estimations largely deviated from proxy albedo chromaticity. Otherwise, it will bias reconstruction towards the proxy mesh. We multiply the proxy term with the following per-pixel weight:

$$w_p(a_i) = e^{-(\min(E_p(a_i, :)) / E_p(a_i, \hat{\boldsymbol{\rho}}_c(a_i)))^2 / \sigma_p^2}, \quad (22)$$

where  $\hat{\boldsymbol{\rho}}_c(a_i)$  is the estimated albedo chromaticity at pixel  $a_i$  using the consensus term alone.

## 6 EXPERIMENTAL RESULTS

In this section, we first report results on synthetic face images generated using a high-quality face dataset and synthetic lighting. We then show results for real data captured using our setup. To self-calibrate each light, we use 2,000 iterations for RANSAC. The reconstruction parameters are set as follows:  $\delta_b = 0.025$ ,  $\lambda_s = 1.5$ ,  $\lambda_p = 0.5$ ,  $\sigma_s = 0.003$ ,  $\sigma_p = 0.01$ . We discretize albedo chromaticity in spherical coordinates as  $\{0^\circ, 1^\circ, \dots, 90^\circ\} \times \{0^\circ, 1^\circ, \dots, 90^\circ\}$ .

We compare our performance against those of representative state-of-the-art techniques [41], [51], [52]. VH12 [41] assumes directional lighting with single albedo chromaticity, and uses the same proxy face as our method for self-calibration. Since CK16 [51] requires directional lighting directions as input, we compute approximated lighting directions as the rays from face center to ground truth light positions. OS18 [52] originally assumes directional lighting, but we adapted it to work for near point lighting by simply using per-pixel lighting directions during computation. The per-pixel lighting directions are obtained using our estimated light positions and proxy face, which are the same as with our method. Notice that only VH12 [41] and our method work under uncalibrated light sources while the other two methods require additional calibration information. After obtaining normal map, we use Poisson integration [67] to get geometry for both our method and comparison methods.

## 6.1 Experiments Using Synthetic Data

To evaluate our method objectively, we apply it to synthetic input images with known ground truth. The synthetic images are generated by rendering high-quality face data from the USC Light Stage [5], [68] under near point lighting and orthographic projection, with resolution of  $2048 \times 1536$ . (Note that while real cameras are not based on orthographic projection, we use it in our simulations to exclude the influence of perspective and focus.) The synthetic lights are distributed with equal azimuth angles between neighboring lights, and at the same elevation angle. The distance between each light and the face center is identical. During rendering, we retain self-shadows on the face while ignoring other shadowing effects on the background. We also avoid saturation by scaling each image so that the maximum pixel intensity is 255.

We first report our system’s performance under different light distances, elevation angles, anisotropy, and crosstalk using a single face data from [5]. Then, we use the face dataset ICT-3DRFE [68] to evaluate our method for different gender, skin appearance, and expression. We also compare with competing techniques in each analysis.

### 6.1.1 Light Distance

In this experiment, we vary the distance between the light sources and face mesh while fixing the elevation angle at  $65^\circ$ . The distance is specified in terms of vertical span of the face; it ranges from 0.5 to 10 with increments of 0.5. The rendered images for the first 8 distances are shown in Fig. 3a.

Fig. 3b compares the calibration errors for vanilla RANSAC and our method. We first transform the calibration results to the same coordinate system as the ground truth light positions before computing errors. We compute the relative position error as Euclidean position error normalized by light source distance. The angular error is computed with regard to the face center. We can see that vanilla RANSAC is less accurate with large fluctuations in error over distance. By comparison, our calibration results are more accurate and robust to changing light source distance, with the relative position error around 0.1 and the angular error around  $5^\circ$  for most distances.

We also compare the reconstruction accuracy of our method using our estimated light positions with VH12 [41] in Fig. 3c. We can see that our method consistently performs better, even at distance 10 (where lighting is almost directional). There is considerable shape deformation for [41] across the different distances as shown in Fig. 4, while our method produces reasonable shapes starting from distance 1.5. At very close distances such as 0.5 and 1, both methods do not perform well due to significant self-shadowing.

Fig. 3c also shows comparisons with using ground truth light positions and mean albedo chromaticity (which are the conditions that should result in the best accuracy under the single chromaticity assumption). In this case, our method using ground truth light positions out-performs the others by a significant margin under almost all distances; this shows the importance of spatially-varying albedo chromaticity. The degraded accuracy at distance 0.5 is due to significant self-shadowing.

### 6.1.2 Light Elevation Angle

The elevation angle of light sources have a direct impact on light source baseline. A large elevation angle results in a small light source baseline, which enables the equipment to be more

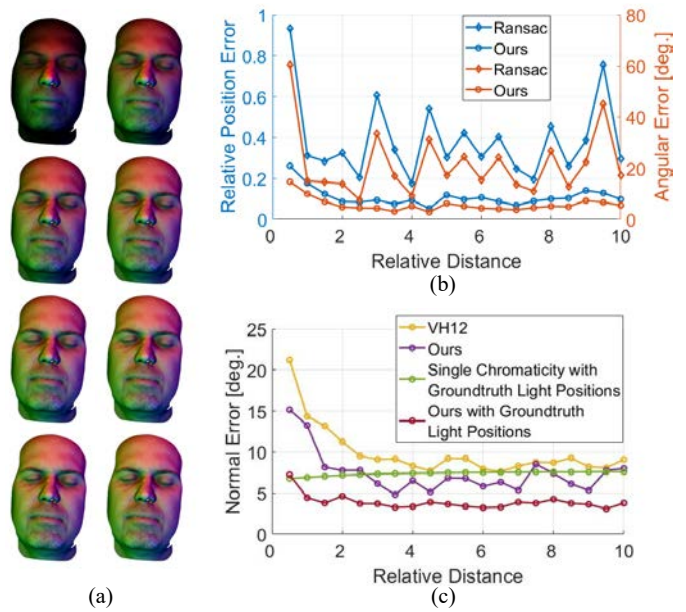


Fig. 3. Effect of changing light source distances. (a) Rendered images under the first 8 distances (distance increases from left to right and from top to bottom). Comparisons on (b) self-calibration and (c) reconstructed normal errors at different light source distances, including against VH12 [41].

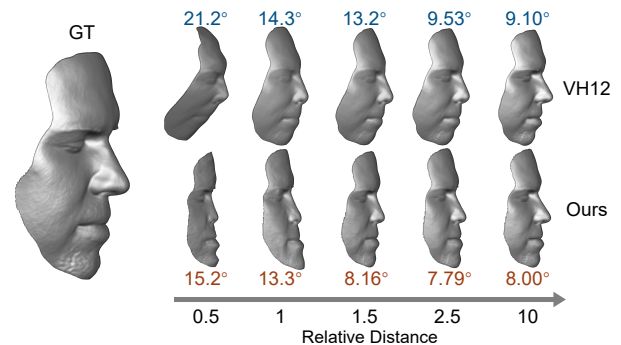


Fig. 4. Comparison of reconstructed geometry with VH12 [41] at different light source distances. The colored numbers are the mean normal errors.

portable. However, the angular difference between light sources decreases as the elevation angle increases, which in turn makes reconstruction less robust. In the extreme case where the elevation angle is  $90^\circ$ , the three light sources degenerate into a single light source, with their spectra combined. On the other hand, small elevation angles results in more self-shadowing, which also negatively impacts reconstruction.

We fix the distance at 2.0 and vary the elevation angle from  $85^\circ$  to  $30^\circ$  with a decrement of  $5^\circ$ . At the elevation angle of  $30^\circ$ , about 30% of facial pixels are in shadow for green and blue lights. Fig. 6 shows the mean normal error of our method against [41], [51], [52]. It can be seen that our method consistently performs the best under all elevation angles. In addition, all methods display a trend to perform worse at two ends of elevation angles, although [51] is less affected by severe self-shadowing at small elevation angle. While [52] produces smaller errors than [41], [51] under medium elevation angles, its accuracy drastically degrades for

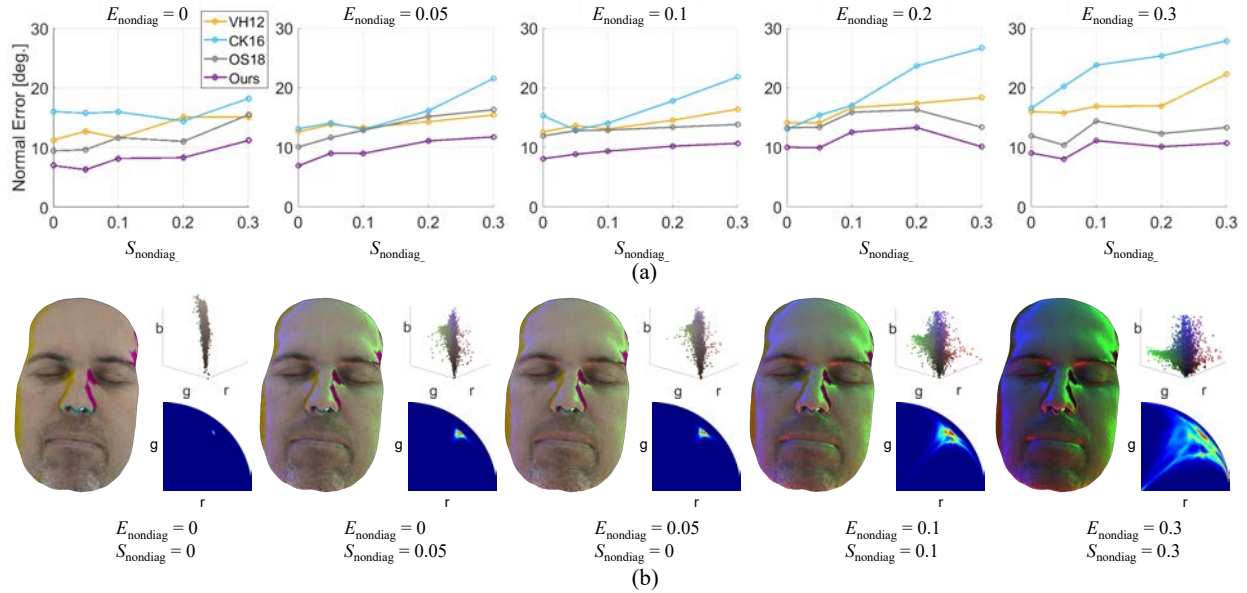


Fig. 5. Experimental results for different crosstalk parameters  $E_{\text{nondiag}}$ ,  $S_{\text{nondiag}}$ . (a) Mean normal errors (in degrees), compared with VH12 [41], CK16 [51], and OS18 [52]. (b) Apparent albedo maps and albedo/albedo chromaticity distributions for five representative examples. The albedo values are rescaled so that most pixels are within  $[0, 1]$ . (Albedo values at pixels near shadows can be very large; they are not used for rescaling.)

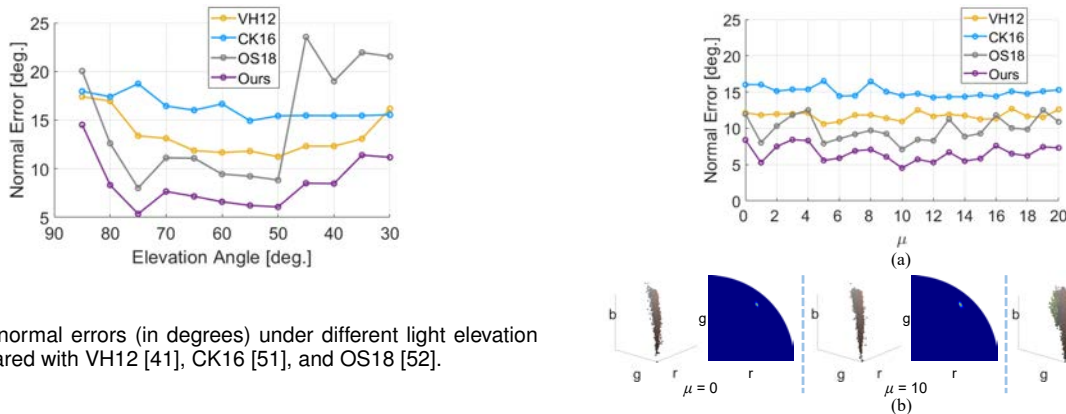


Fig. 6. Mean normal errors (in degrees) under different light elevation angles, compared with VH12 [41], CK16 [51], and OS18 [52].

small elevation angles due to self-shadowing. This highlights the importance of our proposed albedo similarity and proxy prior in correcting errors led by albedo consensus.

### 6.1.3 Light Anisotropy

Unlike the ideal point light model, real LEDs exhibit anisotropic intensity patterns. To analyze its effect on our method, we further render images using an anisotropic point light model [24]:

$$\mathbf{c}_i(x, y) = \rho_i(x, y) \left( \frac{\mathbf{n}_s^i \cdot \mathbf{L}_i(x, y)}{\|\mathbf{L}_i(x, y)\|_2} \right)^{\mu^i} \mathbf{L}_i(x, y) \mathbf{n}(x, y), \quad (23)$$

where  $\mathbf{n}_s^i$ ,  $\mu^i$  are the (unit-length) principal direction and anisotropy parameter of the  $i$ th light source. The anisotropy parameter equals 0 for ideal point light source while a larger value indicates stronger radial attenuation around the principal direction. We render the images at distance 2.0 and elevation angle  $65^\circ$ , with anisotropy parameter ranging from 0 to 20. With  $\mu = 20$ , the half-intensity angle is only about  $15^\circ$ , revealing very strong radial attenuation.

As shown in Fig. 7(a), light anisotropy has no noticeable adverse effect on both our method and comparison methods. We further compute apparent albedo under different anisotropy

Fig. 7. Experimental results for different anisotropy parameter  $\mu$ . (a) Mean normal errors (in degrees), compared with VH12 [41], CK16 [51], and OS18 [52]. (b) Albedo distribution (left) and albedo chromaticity distribution (right) for  $\mu = 0, 10, 20$ .

parameters by using Eq. 6 along with ground truth light position, normal, and per-pixel 3D position. (Note that the apparent albedo is not the true albedo, since it incorporates any outlier effect such that the rendering equation adheres to Eq. 6.) Here, light anisotropy is entirely incorporated as part of albedo. Fig. 7(b) shows the albedo/albedo chromaticity distributions for three cases, where there is only minor difference. This illustrates why light anisotropy has little influence on the accuracy of all methods and validates our use of ideal point light model.

### 6.1.4 Crosstalk

Similar to light anisotropy, the influence of crosstalk can also be interpreted as modification on albedo/albedo chromaticity distribution. To simulate crosstalk, due to a lack of hyperspectral reflectance data, we only consider the wavelengths corresponding

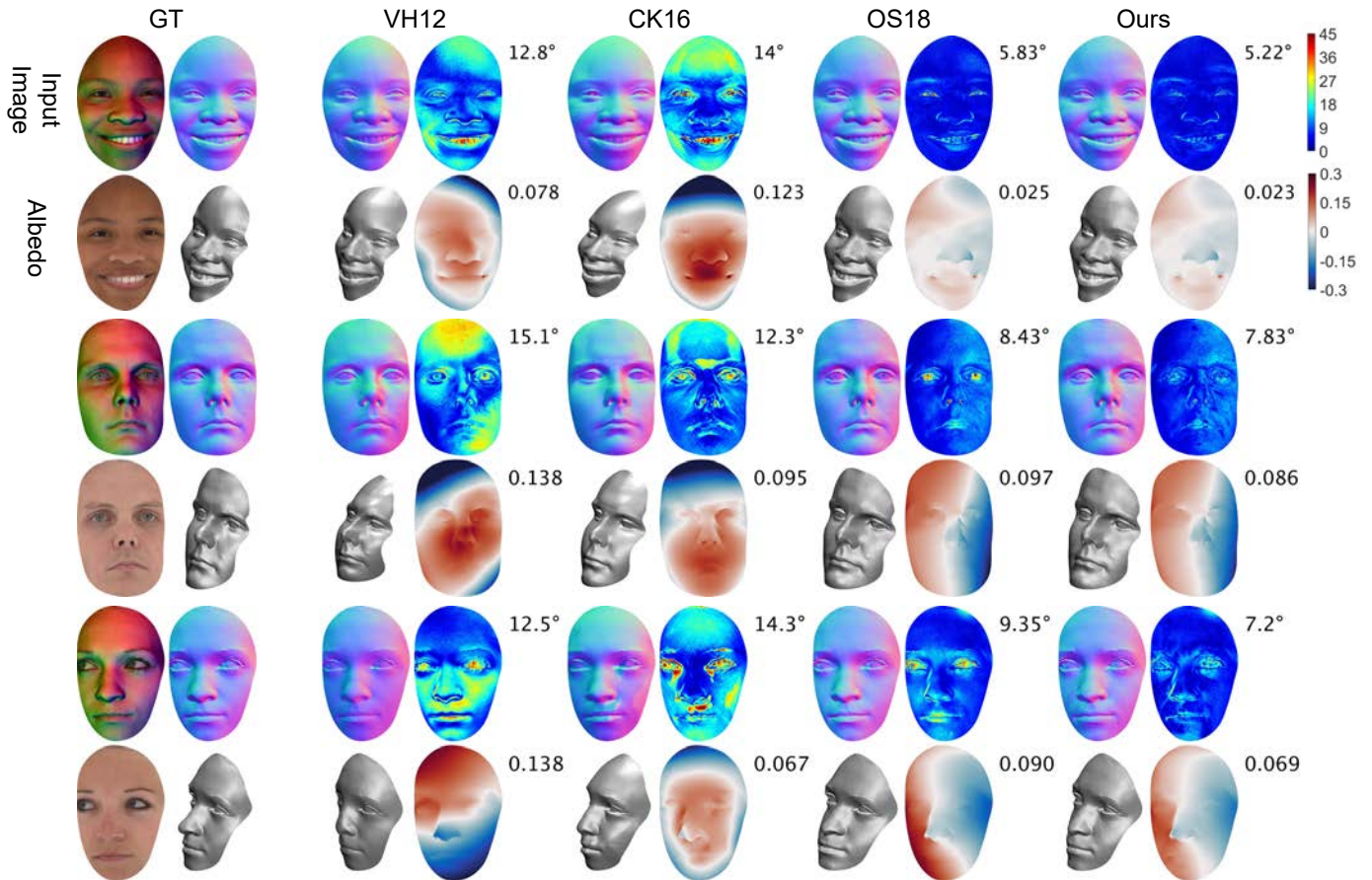


Fig. 8. Comparison with competing techniques (VH12 [41], CK16 [51], OS18 [52]) using data from the ICT-3DRFE dataset. GT is ground truth. The mean normal and geometry errors are listed in the odd and even rows, respectively. More results can be found in the supplementary file.

to red, green, and blue when evaluating Eq. 2, which can be rewritten as:

$$\mathbf{A}(x, y) = \mathbf{S} \text{diag}(\mathbf{r}(x, y)) \mathbf{E}, \quad (24)$$

where  $\mathbf{S}, \mathbf{E} \in \mathbb{R}^{3 \times 3}$ ,  $\mathbf{r}(x, y) \in \mathbb{R}^{3 \times 1}$  and  $\mathbf{S}_{i,k} = \mathcal{S}_i(\lambda_k)$ ,  $\mathbf{E}_{k,j} = \mathcal{E}_j(\lambda_k)$ ,  $\mathbf{r}_k(x, y) = \mathcal{R}(x, y, \lambda_k)$ . Crosstalk exists when any non-diagonal element of  $\mathbf{S}$ ,  $\mathbf{E}$  is non-zero. We set the diagonal elements to 1 and gradually increase their non-diagonal elements  $\mathbf{S}_{\text{nondiag}}, \mathbf{E}_{\text{nondiag}}$  (non-diagonal elements are set as the same) to simulate increasing crosstalk.

Fig. 5(a) shows the normal errors under different combinations of  $\mathbf{S}$  and  $\mathbf{E}$ , where generally more crosstalk leads to worse accuracy for all methods. Fig. 5(b) shows the apparent albedo maps along with albedo/albedo chromaticity distributions for 5 cases. We can see that with more crosstalk, there is stronger spatial albedo variation, which violates the piecewise constancy assumption of [51]. Although [41] does not have a no-crosstalk requirement, it also significantly suffers from the spreading-out of albedo chromaticity distribution due to crosstalk. Our method is more robust to this phenomenon because of the incorporation of albedo similarity and proxy prior.

### 6.1.5 Evaluation on ICT-3DRFE

We further evaluate our method using the ICT-3DRFE dataset [68], which contains highly-detailed albedo and geometry for 23 subjects (22 with 15 expressions each, and one with 11 expressions, with a total of 341 face inputs). The dataset has vastly

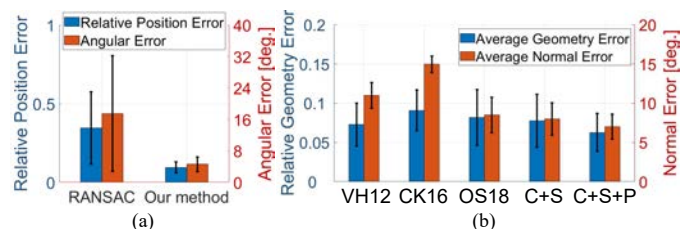


Fig. 9. Error statistics on ICT-3DRFE dataset. (a) Self-calibration errors. (b) Reconstruction errors of VH12 [41], CK16 [51], OS18 [52], our “Consensus + Similarity” and our “Consensus + Similarity + Proxy”.

different skin reflectance as well as face geometry. We rendered images at light source distance 2.0 and elevation angle  $65^\circ$ , with no anisotropy or crosstalk. We also added Gaussian noise ( $\sigma_{\text{noise}} = 2/255$ ) to simulate real images.

As shown in Fig. 9a, our self-calibration method significantly improves over vanilla RANSAC. In Fig. 9b, we compare the accuracy of our face reconstruction method with those of [41], [51], [52]. For our method, we show results of two variants (“Consensus + Similarity” and “Consensus + Similarity + Proxy”) to analyze the influence of each term. We compute relative geometry error as depth error of integrated geometry normalized by depth range of ground truth geometry. Methods using near point light model outperform those using directional light model in terms of normal



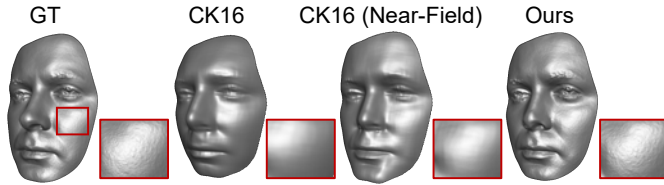


Fig. 10. Example reconstruction results of CK16 [51], near-field version of CK16 and our method on data from the ICT-3DRFE dataset.

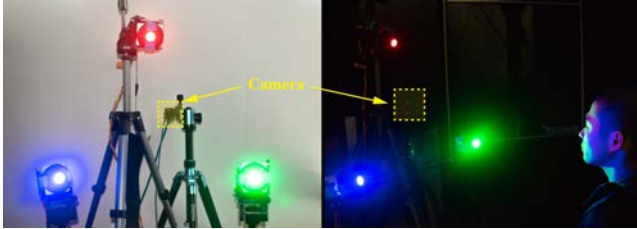


Fig. 11. Hardware setup of our capture system.

error. Each proposed term improves over using consensus only.

Although [51] handles multi-chromaticity, it performs worse than [41]. It is likely that its polynomial model for depth is not suitable for complex geometry. While [41] has lower geometry error than using consensus only [52], our full method improves over this metric and yields the best accuracy. Fig. 8 shows 3 comparisons. Our method works reasonably well in the lip and eyebrow regions, even though they contain non-dominant albedos (which tend to cause incorrect consensus). Shadows, as with light anisotropy and crosstalk, can also be explained by apparent albedo; they result in additional albedo variation (see the leftmost albedo map in Fig. 5(b)). Since our formulation does not enforce spatial constancy, it can better handle such variation compared with [41], [51]. Still, our reconstructions contain errors at shadowed regions near the nose due to inaccurate proxy mesh around the nose. Please see the supplementary material for detailed error statistics and more results.

We have also adapted [51] to work for near point lighting, and use the per-pixel lighting computed from our estimated light positions and proxy face as input. In this case, [51] and our approach use the same illumination model and differ only in the priors used for reconstruction. The near-field version of [51] indeed shows a noticeable improvement over the original one using directional lighting: 10.36 vs. 14.94 degrees for average normal error and 0.080 vs. 0.091 for average geometry error. However, The full model of our approach still outperforms by a noticeable margin, with the average normal error of 6.99 degrees and average geometry error of 0.063. Fig. 10 shows example reconstruction results, where we can see that although the near-field version of [51] contains less low-frequency distortions than the original one, it still lacks recovery of high-frequency details.

## 6.2 Experiments Using Real Data

To collect real data, we built a color photometric capture system as shown in Fig. 11. It consists of 3 LED (red, green, blue) near point lights and a PointGrey Flea3 FL3-U3-88S2C color camera (4096 × 2160). The distance between the light sources and subject is roughly 70cm. We mounted orthogonal linear polarizers in front of the light sources and camera to reduce specular reflection.

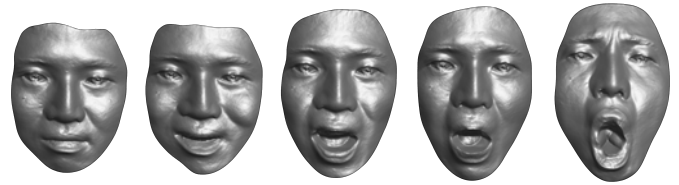


Fig. 12. Reconstruction results for a video clip of a face with changing expressions. Each frame is processed independently.

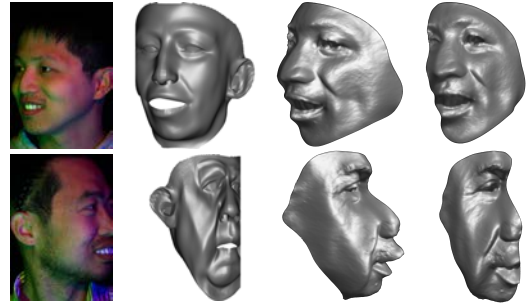


Fig. 13. Failure cases for our method. The first two columns are input image and proxy face, while the last two columns are two views of our reconstruction.

To reduce crosstalk, we compute a de-crosstalk matrix from three images of a white paper, namely, one image for each of the three lights. (This step is done only once.) Specifically, the de-crosstalk matrix is computed as

$$\mathbf{M} = \begin{bmatrix} 1 & \text{med}(I_r^g \oslash I_g^g) & \text{med}(I_r^b \oslash I_b^b) \\ \text{med}(I_r^g \oslash I_r^r) & 1 & \text{med}(I_g^b \oslash I_b^b) \\ \text{med}(I_b^g \oslash I_r^r) & \text{med}(I_b^g \oslash I_g^g) & 1 \end{bmatrix}^{-1}, \quad (25)$$

where  $I_r^g$  is the red channel of the image under green light,  $\oslash$  is Hadamard division operator and  $\text{med}(\cdot)$  yields the median value. This matrix is left multiplied with the RGB value of each pixel.

The final mesh consists of about 3,000,000 vertices. The whole process takes about 12 minutes on a 6-core 3.7GHz CPU with 64GB memory, whereas self-calibration takes 8 minutes and face reconstruction takes 4 minutes. Same with synthetic experiments, proxy faces are made available to [41] for self-calibration while [51], [52] are provided with calibration information.

We captured faces of different people and expressions; Fig. 14 shows results for 3 examples (including different gender and expressions). Results from competing techniques (VH12 [41], CK16 [51], and OS18 [52]) feature local and global geometric distortion as well as over-smoothing. These results also have issues at the lips, and this is because the albedo at the lips differ from those at the rest of the face. Notice that our method works well for exaggerated expressions (such as the second example) even though the proxy face does not accurately depict the expression. Please refer to the supplementary material for more results. We have also compared our self-calibration results using captured real faces with light source positions calibrated using sphere light probes. The average angular error and relative position error are 7.26 degrees and 0.137, respectively. The errors are slightly higher than the performance reported in the synthetic experiment on ICT-



Fig. 14. Reconstruction results of VH12 [41], CK16 [51], OS18 [52] and our method on real data. More examples are in the supplementary file.

3DRFE dataset, possibly due to the experimental ground truth being not perfect.

We have also captured a video clip of a face with changing expressions and reconstructed each frame independently. Fig. 12 shows results for 5 representative frames. The mouth interior was not reconstructed well due to significant self-shadowing.

Fig. 13 shows two failure cases for our method, which contain extreme poses. The reason is that the proxy face generated by 3DMM fitting is significantly less accurate under such poses. This affects our algorithm in two ways: (1) self-calibration of light sources is less robust due to significant pose misalignment of proxy face, and (2) highly incorrect proxy normals adversely affect face reconstruction due to incorrect proxy term.

## 7 CONCLUSION

We have presented a novel color photometric stereo (CPS) method with only 3 uncalibrated near point lights. Our method is capable of reconstructing high-quality face geometry from a *single image*. Self-calibration of the near point lights relies on the geometric

prior from the 3DMM proxy face. We apply RANSAC, followed by hypothesis merging to robustly estimate light positions. We also propose a per-pixel formulation for reconstruction that incorporates albedo consensus, albedo similarity, and proxy prior to handle the ill-posedness of CPS. Synthetic and real experiments show that our method outperforms previous CPS methods that similarly use a single image as input.

In our work, we did not exploit the albedo prior of human faces; this prior may further improve the accuracy of self-calibration and face reconstruction. While not trivial, it would also be interesting to explicitly handle self-shadows. Another possible future work would be extending our method to general objects by learning from depth sensor observations.

## ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program (2018YFB2100500), the programs of NSFC (61976138 and 61977047), STCSM (2015F0203-000-06), and SHMEC (2019-01-07-00-01-E00003).

## REFERENCES

- [1] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, "Morphable face models-an open framework," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 75–82.
- [2] "http://www.di4d.com/."
- [3] "http://www.3dmd.com/."
- [4] "https://www.eisko.com/."
- [5] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," in *Proceedings of the 18th Eurographics conference on Rendering Techniques*. Eurographics Association, 2007, pp. 183–194.
- [6] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: High-resolution capture for modeling and animation," in *Data-Driven 3D Facial Animation*. Springer, 2008, pp. 248–276.
- [7] S. Zhang and P. S. Huang, "High-resolution, real-time three-dimensional shape measurement," *Optical Engineering*, vol. 45, no. 12, p. 123601, 2006.
- [8] Y. Furukawa and J. Ponce, "Dense 3d motion capture for human faces," in *CVPR*. IEEE, 2009, pp. 1674–1681.
- [9] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, p. 191139, 1980.
- [10] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec, "Multiview face capture using polarized spherical gradient illumination," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 6. ACM, 2011, p. 129.
- [11] P. F. Gotardo, T. Simon, Y. Sheikh, and I. Matthews, "Photogeometric scene flow for high-detail dynamic 3d reconstruction," in *ICCV*. IEEE, 2015, pp. 846–854.
- [12] P. Gotardo, J. Riviere, D. Bradley, A. Ghosh, and T. Beeler, "Practical dynamic facial appearance modeling and acquisition," *ACM Trans. Graph.*, vol. 37, pp. 232:1–232:13, 2018.
- [13] S. Herbot and C. Wöhler, "An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods," *3D Research*, vol. 2, no. 3, p. 4, 2011.
- [14] J. Ackermann, M. Goesele *et al.*, "A survey of photometric stereo techniques," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 3-4, pp. 149–254, 2015.
- [15] J. J. Clark, "Active photometric stereo," in *CVPR*. IEEE, 1992, pp. 29–34.
- [16] T. Higo, Y. Matsushita, N. Joshi, and K. Ikeuchi, "A hand-held photometric stereo camera for 3-d modeling," in *ICCV*. IEEE, 2009, pp. 1234–1241.
- [17] F. Sakaue and J. Sato, "A new approach of photometric stereo from linear image representation under close lighting," in *ICCV Workshops*. IEEE, 2011, pp. 759–766.
- [18] R. Mecca, A. Wetzler, A. M. Bruckstein, and R. Kimmel, "Near field photometric stereo with point light sources," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 2732–2770, 2014.
- [19] A. Wetzler, R. Kimmel, A. M. Bruckstein, and R. Mecca, "Close-range photometric stereo with point light sources," in *3DV*, vol. 1. IEEE, 2014, pp. 115–122.
- [20] J. Ahmad, J. Sun, L. Smith, and M. Smith, "An improved photometric stereo through distance estimation and light vector optimization from diffused maxima region," *Pattern Recognition Letters*, vol. 50, pp. 15–22, 2014.
- [21] W. Xie, C. Dai, and C. C. Wang, "Photometric stereo with near point lighting: A solution by mesh deformation," in *CVPR*. IEEE, 2015, pp. 4585–4593.
- [22] F. Logothetis, R. Mecca, and R. Cipolla, "Semi-calibrated near field photometric stereo," in *CVPR*, vol. 3, no. 5. IEEE, 2017, p. 8.
- [23] J. Liao, B. Buchholz, J.-M. Thiery, P. Bauszat, and E. Eisemann, "Indoor scene reconstruction using near-light photometric stereo," *IEEE TIP*, vol. 26, no. 3, pp. 1089–1101, 2017.
- [24] Y. Quéau, B. Durix, T. Wu, D. Cremers, F. Lauze, and J.-D. Durou, "Led-based photometric stereo: modeling, calibration and numerical solution," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 3, pp. 313–340, 2018.
- [25] C. Liu, S. G. Narasimhan, and A. W. Dubrawski, "Near-light photometric stereo using circularly placed point light sources," in *ICCP*. IEEE, 2018, pp. 1–10.
- [26] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman, "Resolving the generalized bas-relief ambiguity by entropy minimization," in *CVPR*. IEEE, 2007, pp. 1–7.
- [27] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan, "Self-calibrating photometric stereo," in *CVPR*. IEEE, 2010, pp. 1118–1125.
- [28] Z. Wu and P. Tan, "Calibrating photometric stereo by holistic reflectance symmetry analysis," in *CVPR*. IEEE, 2013, pp. 1498–1505.
- [29] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato, "Uncalibrated photometric stereo for unknown isotropic reflectances," in *CVPR*. IEEE, 2013, pp. 1490–1497.
- [30] T. Papadimitri and P. Favaro, "A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima," *IJCV*, vol. 107, no. 2, pp. 139–154, 2014.
- [31] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Robust multiview photometric stereo using planar mesh parameterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1591–1604, 2017.
- [32] B. Shi, Z. Mo, Z. Wu, D. Duan, S. K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.
- [33] S. J. Koppal and S. G. Narasimhan, "Novel depth cues from uncalibrated near-field lighting," in *ICCV*. IEEE, 2007, pp. 1–8.
- [34] T. Papadimitri and P. Favaro, "Uncalibrated near-light photometric stereo," *BMVC*, 2014.
- [35] X. Cao, Z. Chen, A. Chen, X. Chen, S. Li, and J. Yu, "Sparse photometric 3d face reconstruction guided by morphable models," in *CVPR*. IEEE, 2018.
- [36] L. Xie, Y. Xu, X. Zhang, W. Bao, C. Tong, and B. Shi, "A self-calibrated photo-geometric depth camera," *The Visual Computer*, vol. 35, no. 1, pp. 99–108, 2019.
- [37] M. S. Drew and L. L. Kontsevich, "Closed-form attitude determination under spectrally varying illumination," in *CVPR*. IEEE, 1994.
- [38] L. Kontsevich, A. Petrov, and I. Vergelskaya, "Reconstruction of shape from shading in color images," *JOSA A*, vol. 11, no. 3, pp. 1047–1052, 1994.
- [39] R. J. Woodham, "Gradient and curvature from the photometric-stereo method, including local confidence estimation," *JOSA A*, vol. 11, no. 11, pp. 3050–3068, 1994.
- [40] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla, "Non-rigid photometric stereo with colored lights," in *ICCV*. IEEE, 2007, pp. 1–8.
- [41] G. Vogiatzis and C. Hernández, "Self-calibrated, multi-spectral photometric stereo for 3d face capture," *IJCV*, vol. 97, no. 1, pp. 91–103, 2012.
- [42] M. Kludiny, A. Hilton, and J. Edge, "High-detail 3d capture of facial performance," in *3DPVT Conference*, 2010.
- [43] B. Bringier, D. Helbert, and M. Khoudier, "Photometric reconstruction of a dynamic textured surface from just one color image acquisition," *JOSA A*, vol. 25, no. 3, pp. 566–574, 2008.
- [44] B. De Decker, J. Kautz, T. Mertens, and P. Bekaert, "Capturing multiple illumination conditions using time and color multiplexing," in *CVPR*. IEEE, 2009.
- [45] H. Kim, B. Wilburn, and M. Ben-Ezra, "Photometric stereo for dynamic surface orientations," in *ECCV*. Springer, 2010, pp. 59–72.
- [46] Z. Jankó, A. Delaunoy, and E. Prados, "Colour dynamic photometric stereo for textured surfaces," in *ACCV*. Springer, 2010, pp. 55–66.
- [47] S. Rahman, A. Lam, I. Sato, and A. Robles-Kelly, "Color photometric stereo using a rainbow light for non-lambertian multicolored surfaces," in *ACCV*. Springer, 2014, pp. 335–350.
- [48] R. Anderson, B. Stenger, and R. Cipolla, "Color photometric stereo for multicolored surfaces," in *ICCV*. IEEE, 2011, pp. 2182–2189.
- [49] G. Fyffe, X. Yu, and P. Debevec, "Single-shot photometric stereo by spectral multiplexing," in *ICCP*. IEEE, 2011, pp. 1–6.
- [50] G. Fyffe and P. Debevec, "Single-shot reflectance measurement from polarized color gradient illumination," in *ICCP*. IEEE, 2015, pp. 1–10.
- [51] A. Chakrabarti and K. Sunkavalli, "Single-image rgb photometric stereo with spatially-varying albedo," in *3DV*. IEEE, 2016, pp. 258–266.
- [52] K. Ozawa, I. Sato, and M. Yamaguchi, "Single color image photometric stereo for multi-colored surfaces," *Computer Vision and Image Understanding*, 2018.
- [53] T. Collins and A. Bartoli, "3d reconstruction in laparoscopy with close-range photometric stereo," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 634–642.
- [54] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt, "State of the art on monocular 3d face reconstruction, tracking, and applications," in *Computer Graphics Forum*, vol. 37, no. 2. Wiley Online Library, 2018, pp. 523–550.

- [55] M. Pietraschke and V. Blanz, "Automated 3d face reconstruction from multiple images using quality measures," in *CVPR*. IEEE, 2016, pp. 3418–3427.
- [56] C. Cao, D. Bradley, K. Zhou, and T. Beeler, "Real-time high-fidelity facial performance capture," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, p. 46, 2015.
- [57] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," in *3DV*. IEEE, 2016, pp. 460–469.
- [58] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *ICCV*. IEEE, 2017, pp. 1576–1585.
- [59] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *CVPR*. IEEE, 2017, pp. 1259–1268.
- [60] A. Tun Trn, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni, "Extreme 3d face reconstruction: Seeing through occlusions," in *CVPR*. IEEE, 2018, pp. 3935–3944.
- [61] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [62] Y. Li, L. Ma, H. Fan, and K. Mitchell, "Feature-preserving detailed 3d face reconstruction from a single image," in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*. ACM, 2018, p. 1.
- [63] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu, "Photo-realistic facial details synthesis from single image," in *ICCV*, October 2019.
- [64] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu, "3d face reconstruction with geometry details from a single image," *IEEE TIP*, vol. 27, no. 10, pp. 4756–4770, 2018.
- [65] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li, "High-fidelity facial reflectance and geometry inference from an unconstrained image," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 162, 2018.
- [66] P. Huber, G. Hu, J. R. Tena, P. Mortazavian, W. P. Koppen, W. J. Christmas, M. Rätzsch, and J. Kittler, "A multiresolution 3d morphable face model and fitting framework," in *VISIGRAPP*, 2016.
- [67] Y. Quéau and J.-D. Durou, "Edge-preserving integration of a normal field: Weighted least-squares, tv and l1 approaches," in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2015, pp. 576–588.
- [68] G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency, "Effect of illumination on automatic expression recognition: a novel 3d relightable facial database," in *Face and Gesture 2011*. IEEE, 2011, pp. 611–618.



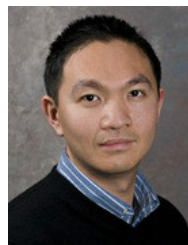
**Mingyuan Zhou** received his Ph.D. in computer science from University of Delaware in 2019. He received his M.E. degree in Computer Engineering from Stevens Institute of Technology in 2014, and B.E. degree in Intelligence Science and Technology from Beijing Information Science and Technology University, China, in 2011. He is now a research scientist at DGene. His research interests include computer vision and computational photography.



**Sing Bing Kang** received his Ph.D. degree in robotics from Carnegie Mellon University in 1994. He is Distinguished Scientist at Zillow Group, and was previously Principal Researcher at Microsoft Corporation. His research interests are computational photography and image-based modeling. Sing Bing has coedited two books ("Panoramic Vision" and "Emerging Topics in Computer Vision") and coauthored two books ("Image-Based Rendering" and "Image-Based Modeling of Plants and Trees"). On the community service front, he has served as Area Chair for the major computer vision conferences and as papers committee member for SIGGRAPH and SIGGRAPH Asia. He was Program Chair for ACCV 2007 and CVPR 2009, and was Associate Editor-In-Chief for IEEE Transactions on Pattern Analysis and Machine Intelligence from 2010-2014. He is a Fellow of the IEEE.



**Zhang Chen** received his Bachelor degree in Electrical and Computer Engineering from Shanghai Jiao Tong University, China in 2014. He is currently a Ph.D. candidate in computer science at ShanghaiTech University, China. He is also with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and University of Chinese Academy of Sciences. His research interests lie in computer vision, computer graphics and computational photography.



**Jingyi Yu** is currently Professor and Executive Dean of the School of Information Science and Technology at the ShanghaiTech University. He is also affiliated with the Department of Computer and Information Sciences at University of Delaware. He received B.S. from Caltech in 2000 and Ph.D. from MIT in 2005. He has published over 120 papers at highly refereed conferences and journals, and holds over 10 international patents on computational imaging. His research interests span a range of topics in computer vision and computer graphics, especially on computational photography and non-conventional optics and camera designs. He is a recipient of the NSF CAREER Award and the AFOSR YIP Award, and has served as an area chair of many international conferences including CVPR, ICCV, ECCV, ICCP and NIPS. He is currently an Associate Editor of IEEE TPAMI, IEEE TIP, and Elsevier CVIU, and will be program chair of ICPR 2020 and IEEE CVPR 2021.



**Ji Yu** received his PhD degree in computer science from the University of Delaware in 2014. He received his M.Sc. in digital media from Nanyang Technological University in 2011 and Bachelor degree in Electrical Engineering from Huazhong University of Science and Technology in 2009. He is now a principal scientist at DGene. His research interests include computational photography, computer vision, and computer graphics.