Sparse Photometric 3D Face Reconstruction Guided by Morphable Models

Xuan Cao Zhang Chen Anpei Chen Xin Chen Shiying Li Jingyi Yu
ShanghaiTech University
393 Middle Huaxia Road, Pudong, Shanghai, China
{caoxuan,chenzhang,chenap,chenxin2,lishy1,yujingyi}@shanghaitech.edu.cn

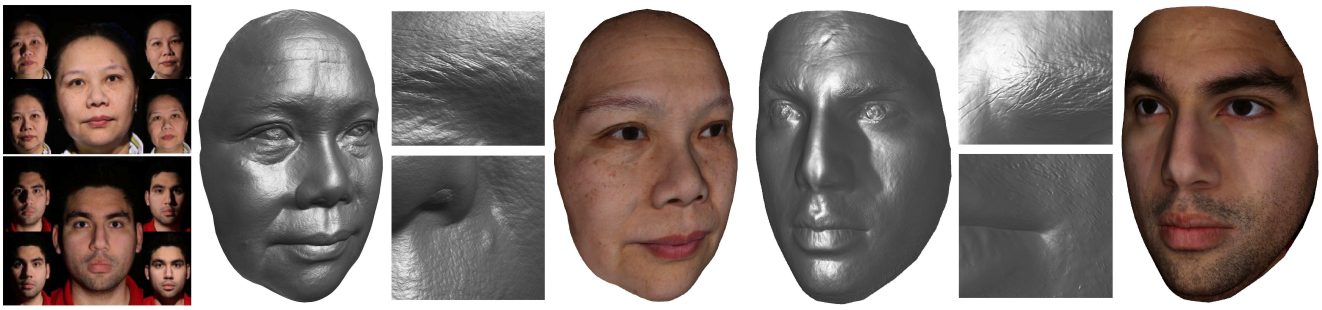

Figure 1: Sample results using our sparse PS reconstruction. By using just 5 input images (left), our method can recover very high quality 3D face geometry with fine geometric details.

Abstract

We present a novel 3D face reconstruction technique that leverages sparse photometric stereo (PS) and latest advances on face registration / modeling from a single image. We observe that 3D morphable faces approach [21] provides a reasonable geometry proxy for light position calibration. Specifically, we develop a robust optimization technique that can calibrate per-pixel lighting direction and illumination at a very high precision without assuming uniform surface albedos. Next, we apply semantic segmentation on input images and the geometry proxy to refine hairy vs. bare skin regions using tailored filter. Experiments on synthetic and real data show that by using a very small set of images, our technique is able to reconstruct fine geometric details such as wrinkles, eyebrows, whelks, pores, etc, comparable to and sometimes surpassing movie quality productions.

1. Introduction

The digitization of photorealistic 3D face is a long-standing problem and can benefit numerous applications, ranging from movie special effects [2] to face detection and recognition [17]. Human faces contain both low-frequency geometry (e.g., nose, cheek, lip, forehead) and high-frequency details (e.g., wrinkles, eyebrows, beards, and pores). Passive reconstruction techniques such as stereo matching [19], multiview geometry [37, 5], structure-from-motion [3], and most recently light field imaging [1] can now reliably recover low frequency geometry. Recovering high-frequency details is way more challenging. Successful solutions still rely on professional capture systems such as 3D laser scans or ultra-high precision photometric stereo such as the USC Light Stage systems [15, 26]. Developing commodity solutions to simultaneously capture low-frequency and high-frequency face geometry is particularly important and urgent.

To quickly reiterate the challenges, PS requires knowing the lighting direction at a very high precision. It is common practice to position a point light at a far distance to emulate a directional light source for easy calibration. In reality, such setups are huge and require strong lighting power. Alternatively, one can use near-field point light sources [40, 9, 28] to set up a more portable system. However, calibrating the lighting direction for each face vertex becomes particularly difficult: one needs to know the relative position between the light source(s) and the face geometry. The light position can be estimated by using sphere [42, 48, 36] or planar light probes [29, 46]. However, the human face would have to be positioned at approximately the same location as the probe. Then the relative position between point lights and facial vertices can be measured. One may resolve this

problem by measuring relative position between probe and human face in the camera coordinates. This would need extra depth camera to locate human face. Our method uses a single camera and perfectly solves this problem by directly calibrating the light position relative to the individual face model. Latest techniques such as [30] uniformly attempt to calibrate lights from subjects. In our paper, we have shown that using morphable face model as proxy, we can already calibrate the light source positions at a high accuracy and our approach achieves higher quality face model than [30].

We leverage recent advances on 3D morphable faces for lighting calibration and geometric reconstruction [10, 8, 21, 35, 33, 39, 12, 13, 22, 43, 34, 51]. Such solutions only use very few or even a single image as input. The 3D face model can then be inferred by morphing the canonical model. Their results are impressive for neutral facial expressions [8, 12]. But high frequency details are still largely missing [10, 7, 35, 31]. The seminal work of [22, 34] manages to recover high frequency geometry to some extent but the results are still not comparable to high-end solutions (e.g., from the USC Light Stage [15, 26]).

In this paper, we combine morphable face approach with sparse PS for ultra high quality 3D face reconstruction. We observe that morphable face approach [21] provides a reasonable geometry proxy for light position calibration. Specifically, we develop a robust optimization technique that can calibrate per-pixel incident lighting direction as well as brightness. Our technique overcomes the artifacts of geometric deformations caused by inaccurate lighting estimation and produces a high-precision normal map. Next, we apply semantic segmentation on input images and the approximated geometry to separately refine hairy vs. bare skin regions. For hairy regions, we adopt a bidirectional extremum filter for detail-preserving denoising. Comprehensive experiments on synthetic and publicly available datasets demonstrate our approach is reliable and accurate. For real data, we construct a capture dome composed of 5 near point light sources with an entry-level DSLR camera. Our technique is able to deliver high quality reconstruction with ultra-fine geometric details such as wrinkles, eyebrows, whelks, pores, etc. The reconstruction quality is comparable to and sometimes surpasses movie quality productions based on dense inputs and expensive setups.

2. Related Works

Photometric Stereo. In computer graphics and vision, photometric stereo (PS) [49] is a widely adopted technique for inferring the normal map of human faces. The normal map can then be integrated (e.g., using Poisson completion [38]) to reconstruct point cloud and then mesh. We refer readers to the comprehensive survey [18] for the benefits and problems of the state-of-the-art methods. In general, recovering high quality 3D geometry requires using complex setups. The most notable work is the USC Light Stage [26, 15] that utilizes 156 dedicatedly controlled light sources to simulate first-order spherical harmonics function. Their solution can produce very high-quality normal map using near point light sources and the superb results have been adopted in movie productions. The setup, however, is rather expensive in cost and labor. Developing cheaper solutions capable of producing similar quality reconstruction is highly desirable , but by far few solutions can match the Light Stage.

2D-to-3D Conversion. There is an emerging interest on directly converting a 2D face image to a 3D face model. Most prior works can be categorized into 3D morphable faces and learning-based techniques. Kemelmacher et al. [25, 23] and Suwajanakorn [41] reconstructed 3D face models from large unstructured photo collections (or video frames). Booth et al. [8] automatically synthesized a 3D morphable model from over 10,000 3D faces. Bolkar [7] utilized a multilinear model based learning framework that uses much smaller training datasets. [21] proposed a Surrey Face Model which provides high resolution 3D morphable model and landmarks alignment. Hu [20] further added hair for the face model. Face models obtained from these approaches are sensitive to pose, expression, illumination, etc, and the problem can be mitigated by using more images [35, 31] or special facial feature decoders [12].

In the past few years, a large volume of deep learning based approaches have shown great success on face pose and geometry estimations [13, 22, 34]. Trigeorgis et al. [43] tailored a deep CNN to estimate face normal map "in the wild" and then inferred the face shape. Tran et al. [44] applied regression to recover discriminative 3D morphable face models. The main goal of these approaches is face recognition and the recovered geometry is generally highly smooth. Most recent techniques [34] can recover certain medium-scale details such as deep wrinkles but the quality is still not comparable to professional solutions.

In a similar vein as ours, Park et al. [30] utilized a coarse initial 3D geometry to estimate lighting parameters where the initial geometry was obtained via multiview stereo and refined the geometry using a large set of photometric images. Compared with the large number of images, our method requires much fewer inputs. The 3D morphable face is smoother than that from multiview stereo but less accurate. We further conduct optimization and semantic segmentations for refinement.

3. Methods

Fig. 2 shows our processing pipeline. The input is a small set of images taken under different illumination. We first obtain a proxy face model through 3D morphable model [21], with pose and expression aligned with the input images. At the same time, initial normal map and seg-

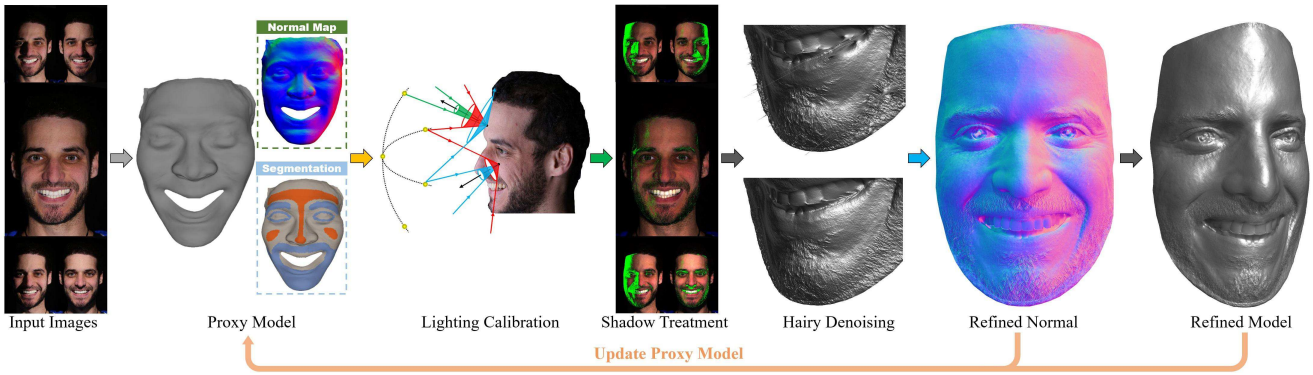

Figure 2: The processing pipeline of our proposed sparse PS face reconstruction framework.

mentation are inferred from proxy model. We then develop an optimization scheme that, without assuming uniform albedo, jointly estimates positions and illumination of all lights from the proxy model. It should be noted that the estimated light source positions are relative to the subject face, so the incident lighting direction for each face vertex can be easily calculated. This is the key idea in our method. Next, we detect shadows and choose at least three reliable incident lights to calculate high-resolution normal map. Subsequently, depth gradient maps are calculated from the normal map. We further develop a bidirectional extremum filter to denoise depth gradient maps. Finally we get the face geometry by integrating the depth gradient maps [32]. We update the proxy model with the refined model and iterate the processing. Our method benefits from the face segmentation twofold: it is more robust to carry out lighting calibration using only initial normal in smooth facial area; we can apply denoising filter only on hairy area to maximally preserve the geometry details in bare skin regions.

3.1. Shading Model and Proxy Geometry

Under the Lambertian assumption, the intensity of a pixel is:

$$I = \rho N \cdot L \tag{1}$$

where ρ and N are the albedo and normal at the pixel, L is the light direction at the corresponding vertex.

PS [49] is an over-determined problem: for one pixel, given at least 3 intensity values I and lighting directions L, the normal N and albedo ρ can be uniquely determined. Conversely, the lighting calibration problem is under-determined: the L and ρ can NOT be figured out using I and N at only one pixel. As shown in Fig. 3, each vertex maps to a triplet of (I, ρ, N), and constrain the potential lighting directions on a conical surface. For the directional light source model, three linearly independent triplets of (I, ρ, N) can be used to figure out the light direction. For near point light model, more triplets are needed to calculate the positions of the light source. It is easy to measure I and N, but it is challenging to obtain ρ in advance. So lighting calibration generally relies on given albedo or uniform albedo assumption.

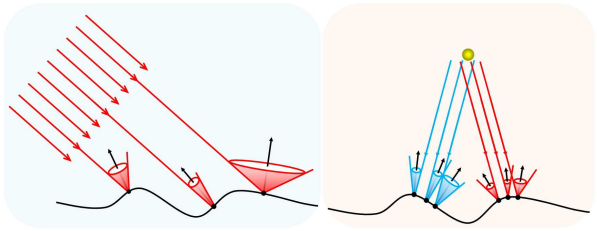

Figure 3: In traditional PS, parallel (left) and point light (right) calibrations rely on uniform albedo assumption.

In our approach, we exploit recent 3D morphable model [21] to generate a proxy face model at first. We utilize the normal map of proxy model to calibrate the light positions. We also use the proxy model to segment the photographed face into two categories of regions: smooth regions including forehead, cheekbone, inner cheek and nose bridge potentially have reliable initial normal, and hairy regions including eyebrows, eyelids, mouth surroundings, chin and outer cheek are generally noisy and require additional processing. Since proxy face models always share the same vertex topology, we can conduct coherent segmentations for different faces.

3.2. Near Point Light Calibration

We aim to replace distant directional light sources with near point light sources, to substantially reduce the cost and space requirement of the PS setup while maintaining the performance. The key challenge is to estimate relative positions from each point light to each surface vertex. In addition, illumination variations across the light sources can cause severe geometry deformation [9]. In this section, we describe a robust auto-calibration technique that conducts estimation for the positions and illumination of near point lights.

There are two classical approaches for calibrating near point lights. One resorts to specific instrument like spherical [48, 36] or planar [29, 46] probes. These light probe-

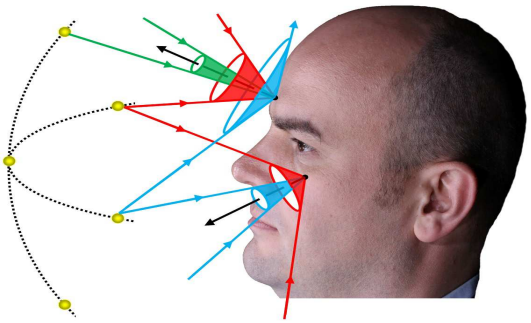

Figure 4: Our light calibration approach uses proxy model for relative light position calibrations. Assuming known surface normal, we can form over-determined systems by jointly considering constraints from multiple light sources on multiple key vertices. Cones of different colors indicate constraints imposed by different light sources.

based methods only recover the light positions in camera coordinate system. The relative positions to surface vertices, however, need extra efforts. The second utilizes the reflectance data of the Lambertian surface with known geometry/normal. For example, Zheng et al. [52] recovers the light directions from known surface normals at multiple points with uniform albedo. By further assuming that neighboring pixels have similar albedo, Mancini et al. [27] estimates the light directions at multiple vertices and subsequently the light positions. Weber et al. [47] uses two cubes covered with white paper for light position calibration. More recently, some works calibrate "in-the-wild" lighting based on spherical harmonics representation or quadratic lighting model [4, 50, 24, 45, 14, 30]. All these approaches strongly rely on uniform albedo assumption and reach a low-rank approximation of real lighting condition, while most surface, including human face, exhibits non-uniform albedo. In this paper, we employ the second approach with the proxy model to calibrate light positions and brightness without assuming uniform albedo.

Our approach is instrument-free and does not make uniform albedo assumption. We first extract m key pixels from the smooth regions in the semantically segmented face image and obtain their normal $\mathbf{N} \in \mathbb{R}^{m\times 3}$ along with their corresponding key vertex positions $\mathbf{V} \in \mathbb{R}^{m\times 3}$. Recall, for non-uniform albedos, we will not be able to solve for ρ and L in Eq. (1) separately for each light. We therefore jointly solve Eq. (1) for all m key vertices and n lights as shown in Fig. 4. According to Eq. (1), we have

$$\mathbf{I} = \text{diag}(\boldsymbol{\rho})\mathbf{N}\mathbf{L}^{\mathsf{T}}, \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{m\times n}$ is the image intensity at the key pixels and $\mathbf{L} \in \mathbb{R}^{n\times 3}$ is the lighting directions. For near point lights with inconsistent illumination intensity, we replace $\mathbf{L}$ with the scaled directions $\mathbf{D}_{i,j}$ for the j^{th} light on the i^{th} key vertex, so that:

$$\mathbf{D}_{i,j} = \boldsymbol{\beta}_j \cdot \frac{1}{\|\mathbf{P}_j - \mathbf{V}_i\|_2^2} \cdot \frac{(\mathbf{P}_j - \mathbf{V}_i)}{\|\mathbf{P}_j - \mathbf{V}_i\|_2} = \frac{\boldsymbol{\beta}_j(\mathbf{P}_j - \mathbf{V}_i)}{\|\mathbf{P}_j - \mathbf{V}_i\|_2^3}, \quad \text{for } i = 1, 2, ..., m \text{ and } j = 1, 2, ..., n. \tag{3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{n\times 1}$ and $\mathbf{P} \in \mathbb{R}^{n\times 3}$ are the brightness and positions of all lights. The second term $\frac{1}{\|\mathbf{P}_j - \mathbf{V}_i\|_2^2}$ reflects the inverse square law between light brightness and distance, which is critical for near point light calibrations. The image intensity of i^{th} key pixel under the j^{th} lighting is represented as

$$\mathbf{I}_{i,j} = \boldsymbol{\rho}_i \mathbf{N}_i \mathbf{D}_{i,j}^{\mathsf{T}}. \tag{4}$$

To solve for the illumination $\boldsymbol{\beta}$ and position $\mathbf{P}$ of light sources, we formulate the problem as the following optimization:

$$\begin{aligned}\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{P}} = \underset{\boldsymbol{\rho},\boldsymbol{\beta},\mathbf{P}}{\text{argmin}} & \sum_{i=1}^{m}\sum_{j=1}^{n} \left\|\mathbf{I}_{i,j} - \boldsymbol{\rho}_i \mathbf{N}_i \mathbf{D}_{i,j}^{\mathsf{T}}\right\|_2^2 \\ & + \lambda_1 \sum_{j=1}^{n} \left\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_j\right\|_2^2 + \lambda_2 \left\|\boldsymbol{\rho}\right\|_2^2 \\ & + \lambda_3 \sum_{j=1}^{n} (\|\mathbf{P}_j\|_2 - d)_2^2,\end{aligned} \tag{5}$$

where $\bar{\boldsymbol{\beta}}$ is the mean of all elements in $\boldsymbol{\beta}$, $d \in \mathbb{R}$ is a prior of the distance between the lights and geometry proxy. The first term represents the least square error under the Lambertian surface model. The second term is based on the fact that brightness variations are relatively small in our setup.

Note that there is a scale ambiguity between $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ in Eq. (5). Therefore we append the third term to enforce the uniqueness of $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$. The last term aims to remove outliers in $\mathbf{I}$, e.g., the ones that deviate greatly from the Lambertian surface model due to noise.

We initialize $\boldsymbol{\rho}$ as the maximal image intensity of each key pixel across the light sources. We initialize $\boldsymbol{\beta}$ as vector $\mathbf{1}$. The near point light sources distribute around the subject with roughly equal distance. We roughly guess d through human perception.

Since the normal from the proxy face model may be inaccurate, the $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{P}}$ estimated from Eq. (5) are impeded by the inaccuracy of $\mathbf{N}$. To compensate for this, we further refine the estimates by iteratively performing the following two optimizations:

(1) Fix the brightness $\boldsymbol{\beta}$ and positions $\mathbf{P}$ of all lights, update albedo $\boldsymbol{\rho}$ and normal $\hat{\mathbf{N}}$ of the key pixels,

$$\min_{\boldsymbol{\rho},\hat{\mathbf{N}}} \sum_{i=1}^{m}\sum_{j=1}^{n} \left\|\mathbf{I}_{i,j} - \boldsymbol{\rho}_i \hat{\mathbf{N}}_i \mathbf{D}_{i,j}^{\mathsf{T}}\right\|_2^2 + \lambda_n \left\|\hat{\mathbf{N}} - \mathbf{N}\right\|_F^2. \tag{6}$$

(2) Fix the albedo $\boldsymbol{\rho}$ and normal $\hat{\mathbf{N}}$ for the key pixels, update brightness $\boldsymbol{\beta}$ and positions $\mathbf{P}$ of all lights,

$$\min_{\boldsymbol{\beta},\mathbf{P}} \sum_{i=1}^{m}\sum_{j=1}^{n} \left\| \mathbf{I}_{i,j} - \boldsymbol{\rho}_i \hat{\mathbf{N}}_i \mathbf{D}_{i,j}^{\mathsf{T}} \right\|_2^2 + \lambda_\beta \sum_{j=1}^{n} \left\| \bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_j \right\|_2^2 + \lambda_P \sum_{j=1}^{n} (\|\mathbf{P}_j\|_2 - d)_2^2. \tag{7}$$

For our experiments, we empirically set $\lambda_1 = \lambda_2 = \lambda_\beta = 0.001$, $\lambda_3 = \lambda_P = 0.0001$ and $\lambda_n = 10^{-6}$.

3.3. Handling Shadow Areas

Our input images contain both cast shadow and self shadow. Image intensities in shadow areas clearly violate the PS laws, and consequently degrade normal estimations. Solutions to detect shadow areas, such as intensity-based segmentation, are sensitive to the image content, especially with non-uniform albedo.

For a pixel, albedo keeps constant under different illumination. Based on this fact, we develop a robust shadow detection method. The proxy face model provides crucial cues to detect shadows in our method.

Cast Shadow. For pixel $i \in \{1, 2, ..., m\}$ and light $j \in \{1, 2, ..., n\}$, from Eq. (4), we have

$$\hat{\boldsymbol{\rho}}_{i,j} = \frac{\mathbf{I}_{i,j}}{\mathbf{N}_i \mathbf{D}_{i,j}^{\mathsf{T}}}. \tag{8}$$

We already have $\mathbf{N}_i$ from the proxy face model and we can compute $\mathbf{D}_{i,j}^{\mathsf{T}}$ by substituting estimated $\boldsymbol{\beta}$ and $\mathbf{P}$ into Eq. (3). Therefore, we can get $\hat{\boldsymbol{\rho}}_{i,j}$ as the estimated albedo of pixel i under light j. Theoretically, pixel i is in shadow under light j if $\hat{\boldsymbol{\rho}}_{i,j} = 0$. In reality, however, $\hat{\boldsymbol{\rho}}_{i,j}$ may be nonzero even when pixel i lies in shadow due to calibration errors, inter-reflections, subsurface scattering, etc.

In experiments, we find that under lightings that produce no shadow, the values of $\hat{\boldsymbol{\rho}}_{i,j}$ are similar. On the other hand, under lightings that do produce shadow, the values of $\hat{\boldsymbol{\rho}}_{i,j}$ are very small. Eq. (9) reveals that, for each pixel i, we can first calculate the mean albedo $\bar{\boldsymbol{\rho}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\boldsymbol{\rho}}_{i,j}$ of this pixel and obtain the set $\mathcal{S}_i$ including $\hat{\boldsymbol{\rho}}_{i,j}$ higher than $\bar{\boldsymbol{\rho}}_i$. And we calculate the mean value μ_i of the set $\mathcal{S}_i$. We then deem the pixel i out of shadow under light j if the estimated albedo $\hat{\boldsymbol{\rho}}_{i,j}$ is higher than $(1-\tau)\mu_i$, where τ is set as 0.4 in our experiments. Consequently, we deem the lights in $\mathcal{L}_i$ possibly reliable for the normal estimation at pixel i.

$$\begin{cases} \mathcal{S}_i = \{\hat{\boldsymbol{\rho}}_{i,j} \mid \hat{\boldsymbol{\rho}}_{i,j} > \bar{\boldsymbol{\rho}}_i\} \\ \mu_i = mean(\mathcal{S}_i) \\ \mathcal{L}_i = \{j \mid \hat{\boldsymbol{\rho}}_{i,j} > (1-\tau)\mu_i\} \end{cases} \tag{9}$$

Self Shadow. We further only deem lights whose incident lighting direction is smaller than 90° valid, as shown in Eq. (10).

$$\mathcal{A}_i = \{j \mid \mathbf{N}_i \mathbf{D}_{i,j}^{\mathsf{T}} > 0\} \tag{10}$$

Considering both cast and self shadow, for pixel i, we only use valid light sources in $\mathcal{V}_i = \mathcal{L}_i \cap \mathcal{A}_i$ to estimate the normal at this pixel.

3.4. Denoising Hairy Regions

Hairy regions of the face such as shaggy beards and bushy eyebrows contain very complex geometric and shading effects where the Lambertian model fails. Under sparse lightings, normal estimations in these regions are particularly noisy. Thabo Beeler [6] detected hairs and then employed a hair-synthesis method to create hair fibers that plausibly match the image data. We detect the hairy area based on the semantic segmentation and reconstruct high-fidelity realistic hair micro-geometry.

Given N_x, N_y, N_z as the x, y, z components of the normal, we first compute depth gradient maps G_x, G_y as

$$G_x = -\frac{N_x}{N_z}, \quad G_y = -\frac{N_y}{N_z}. \tag{11}$$

Our goal is to denoise the gradient maps. However, traditional denoising filters also remove high-frequency geometry. We adopt a simple yet effective bidirectional extremum filter in Eq. (12) to eliminate the singular values in the gradient maps while preserving high-frequency geometry. Specifically, we first compute a transformed gradient map G^t (for both G_x and G_y) as $G^t = \left|G - \bar{G}\right|$, where $\bar{G}$ is the mean value. For each pixel, if its transformed gradient $G^t(u,v)$ exceeds the mean of the transformed gradient map $\bar{G^t}$ scaled by a factor σ, we update the original gradient $G(u,v)$ with the median gradient in neighboring pixels.

$$G'(u,v) = \begin{cases} median(G \in win(u,v)), & G^t(u,v) > \sigma\bar{G^t} \\ G(u,v), & G^t(u,v) \le \sigma\bar{G^t} \end{cases} \tag{12}$$

where $win(u,v)$ is the neighboring window around the pixel at (u,v). In our experiments, we set σ as 5 and neighborhood window as 10×10, and apply the filter only to the hairy regions.

3.5. Iterative Optimization

In our experiments, face proxy from different methods can greatly affect the light calibration accuracy and then reconstruction quality. We adopt an iterative framework to significantly reduce the sensitivity to the initial rough model. Once we obtain the face model after all the steps mentioned above, we can substitute proxy model with our reconstructed high quality model and repeat the entire process for further refinement, as shown in Fig. 2. The process stops when the change of estimation is rather small. For all experiments in this paper, we iteratively conduct the process

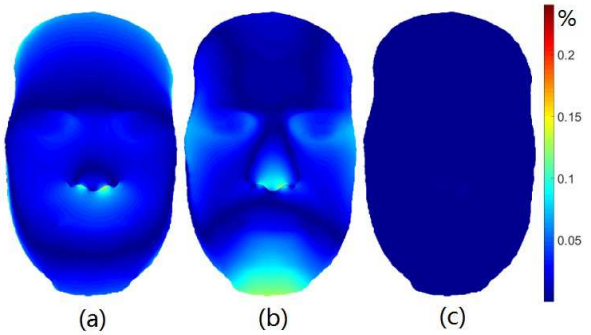

Figure 5: Reconstruction error comparisons. (a) Parallel lighting assumption with known albedo and normal. (b) Using matrix factorization [30]. (c) Our approach. Error is measured in terms of the ratio between the depth deviation and the ground truth depth.

for no more than 10 times and the results are already highly accurate.

4. Experiments

We have conducted comprehensive experiments on both publicly available datasets and our own captured data.

4.1. Synthetic Data

For synthetic experiments, we use the face models reconstructed from the Light Stage [26] as the ground truth. The model contains highly accurate low-frequency geometry and high-frequency details. Using the Lambertian surface model and point light source model, we render 5 images of the model illuminated by different near point light sources on a sphere surrounding the face. The radius of the sphere is set to be equal to the length from the forehead to chin. We use the rendered data to compare the accuracy of various reconstruction schemes.

We first test the parallel light assumption. Specifically, we analyze two scenarios: 1) using the ground truth albedo and normal to calibrate parallel light directions, and then using the light directions to calculate normals, and 2) using the matrix factorization-based method [30] to simultaneously solve for parallel light directions and normals. For point light model, we use a proxy face model predicted from one of the rendered image for lighting calibration and use the results to obtain per-pixel lighting direction and normal.

To apply [30], we use the normal from proxy model as prior. To measure the reconstruction error, we align the reconstructed face models with ground truth model under the same scale and then calculate the reconstruction error as the sum of per-pixel absolute depth error normalized by the depth range of ground truth model. Fig. 5 shows that face models reconstructed using parallel light model yield noticeable geometric deformations while the face model from our method produces much smaller error. Notice that all

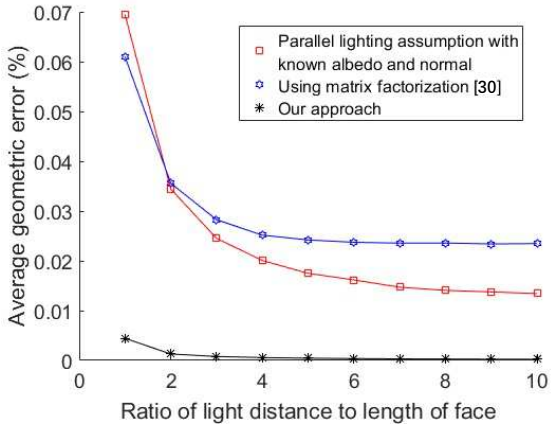

Figure 6: Reconstruction errors under different light distances using our technique vs. the state-of-the-art. Unit distance corresponds to the face length (the distance between forehead and chin).

Figure 7: We constructed an acquisition system composed of 5 point light sources and a single DSLR camera.

three face models uniformly incur larger errors around the forehead and lower edge of nose tip. This is because at such spots, N_z approaches 0, and according to Eq. (11), a small disturbance in normal incurs large errors in G_x and G_y and subsequently the depth estimation.

We further test how parallel lighting assumption impact reconstruction accuracy when the point lights are positioned farther away. We vary the distance between the light sources and the face, ranging from one unit of the length between the forehead and chin to ten units, as shown in Fig. 6. For both parallel and point light source models, the error decreases as the distance increases. However, our method outperforms the other two with a significant margin.

4.2. Real Data

For real data, we have constructed a sparse photometric capture system composed of 5 LED near point light sources and an entry-level DSLR camera (Canon 760D) as illustrated in Fig. 7. The distance between the light sources and photographed face is about 1 meter. To eliminate specular reflectance, both light sources and camera are mounted with polarizers, where the polarizers on light sources are orthogonal to the one on the camera. Each acquisition captures 5 images (1 light source per image) at a resolution of 6000 $\times$ 4000. The process takes less than 2 seconds.

We acquire faces of people with different gender, race and age. Fig. 10 shows our reconstruction of four faces,

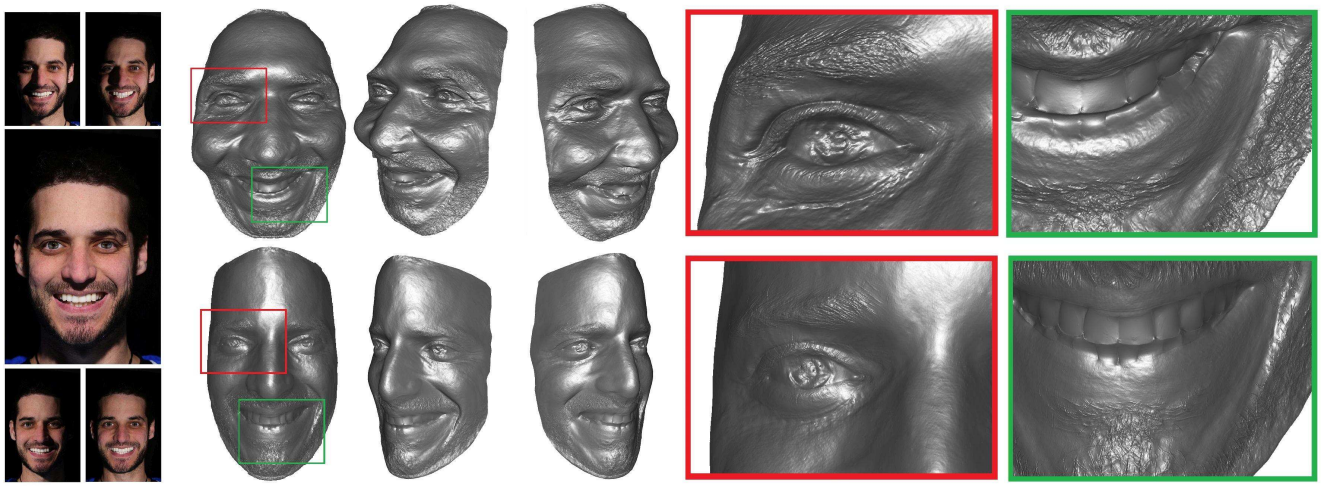

Figure 8: Reconstruction results of [30] (top row) vs. ours (bottom row). [30] causes low-frequency geometrical deformation and high-frequency geometrical noise when using a sparse set of images. Our approach is able to faithfully reconstruct face geometry without deformation and at the same time recover fine details.

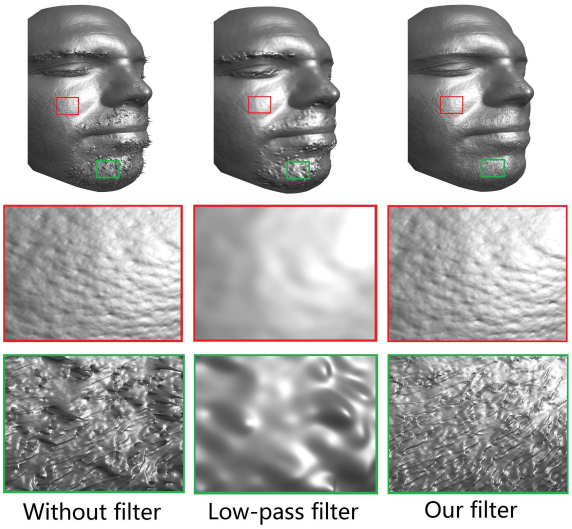

Figure 9: Comparisons of different denoising filters in hairy regions. Notice that spiked artifacts in beards are removed by both filters. However, low-pass filter smooths out the high-frequency geometry of hair while our filter preserves such details.

where the first column shows the proxy models using [21]. The proxy models are reasonable but lack geometric details. Our reconstruction reduces geometric deformations and reveals compelling high-frequency geometric details. We compare our technique with [30] in Fig. 8. Note that neither method requires additional instruments for calibration or 3D scanning. The result from [30] exhibits noisy normals and contains bumpy artifacts as well as geometry deformation over the entire face. This is mainly because the images contain large areas of shadows that generate significant amount of outliers. The outliers are detrimental to the reconstruction especially with only 5 input images. In contrast, our reconstruction exhibits very high quality and low noise, largely attributed to our optimization techniques together with shadow and hairy region detection schemes, as shown in Fig. 8.

In Fig. 9, we demonstrate the importance and effectiveness of our denoising filter on hairy regions. Without denoising, we observe a large amount of spiking artifacts at the beard and eyebrow regions. Direct low-pass filtering reduces the noise but at the same time over-smooths the geometry. Notice that the beards become smoothed after low-pass filtering. Our bidirectional extremum filter, instead, simultaneously removes noise while preserving geometric details. We use the facial region segmentation from Section 3.1 and only apply our denoising filter on the hairy regions.

5. Conclusions and Future Work

We have presented a novel sparse photometric stereo technique for reconstructing very high quality 3D faces with fine details. At the core of our approach is to use base geometric model obtained from 3D morphable faces as geometry proxy for robustly and accurately calibrating the light sources. We have shown our joint optimization strategy is capable of calibration under non-uniform albedo. Finally, we have exploited semantic segmentation techniques for separating hairy vs. bare skin regions where we use bidirectional extremum filter for handling the hairy regions.

Although our paper exploits the 3D morphable face models, we can also potentially use the recent learning-based approaches [44, 17] that can produce plausible 3D face models from a single image. In our experiments, we found that the initial result from [44], although visually pleasing, still deviates from the ground truth too much for reliable lighting

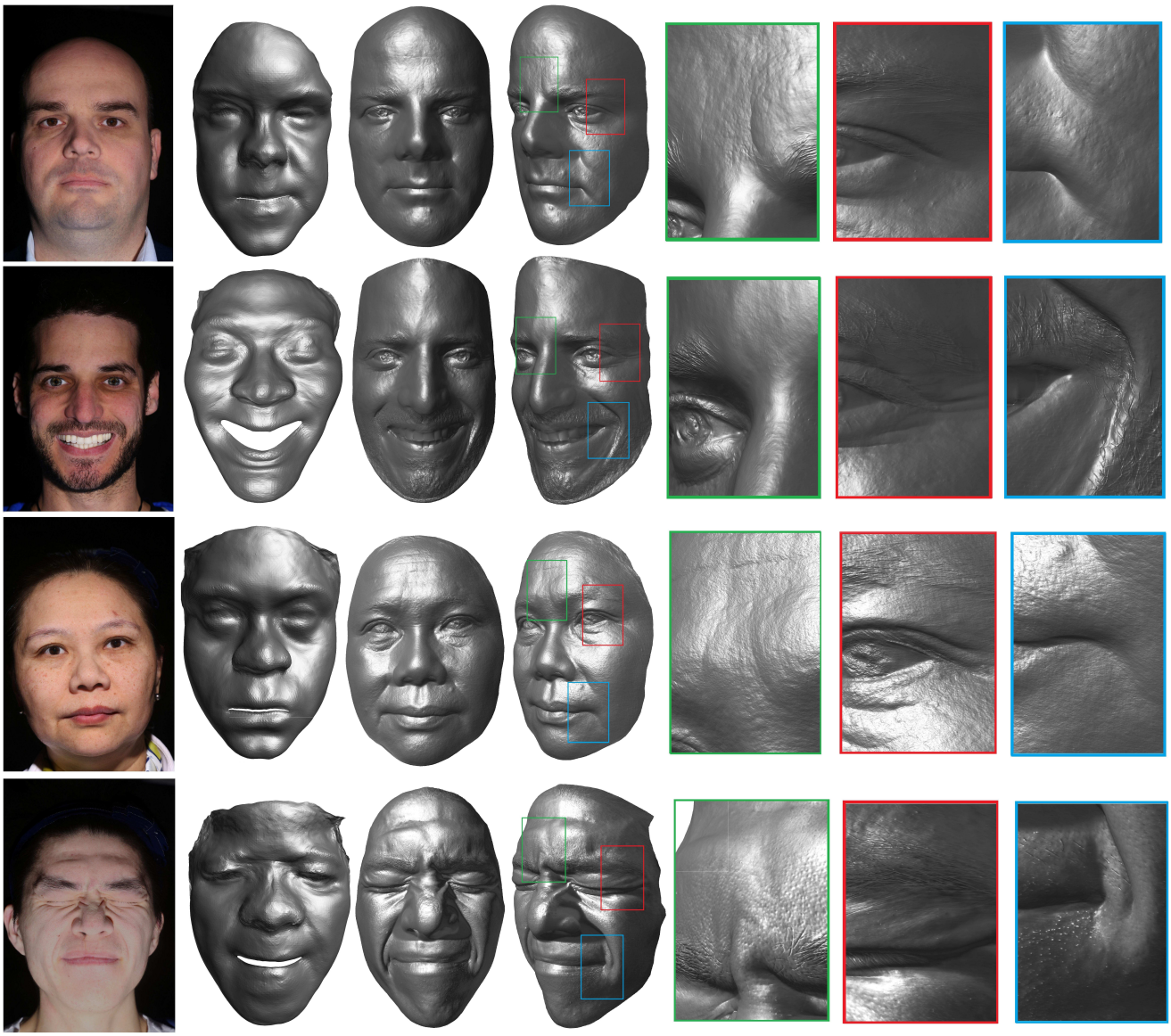

Figure 10: Our reconstruction results across gender, race and age. From left to right, we show one of the 5 input images, the proxy face model, and our final reconstruction results. Closeup views of the eyes and mouth regions illustrate fine geometric details recovered by our technique. Additional results can be found in the supplementary materials.

estimation (see supplementary materials). Our immediate next step therefore is to see how to integrate the shading information into their network framework to produce similar quality results.

There is also an emerging trend of combining semantic labeling with stereo or volumetric reconstruction [16, 11]. In our work, we have only used a small set of labels. In the future, we plan to explore more sophisticated semantic labeling technique that can reliably separate a face into finer regions, e.g., eye region, cheek, mouth, teeth, forehead, etc, where we can handle each individual region based on their characteristics. A more interesting problem is how to simultaneously recover multiple faces (of different people) under the photometric stereo setting. For example, if each face exhibits a different pose, a single shot under directional lighting will produce appearance variations across these faces that are amenable for PS reconstruction.

Acknowledgement

The authors would like to thank Sören Schwertfeger, Laurent Kneip, Andre Rosendo, Qilei Jiang, Mario Eduardo Villanueva for cooperation of capturing face data. Thanks Wenguang Ma for assistance in building the prototype.

References

[1] http://raytrix.de/. 1

[2] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. Creating a photoreal digital actor: The digital emily project. pages 176–187, 2009. 1

[3] A. Bartoli and P. Sturm. Structure-from-motion using lines: Representation, triangulation, and bundle adjustment. *CVIU*, 100(3):416–441, 2005. 1

[4] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007. 4

[5] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *ACM Transactions on Graphics*, 29(4):1–9, 2010. 1

[6] T. Beeler, B. Bickel, G. Noris, P. Beardsley, S. Marschner, R. W. Sumner, and M. Gross. Coupled 3d reconstruction of sparse facial hair and skin. *Acm Transactions on Graphics*, 31(4):1–10, 2012. 5

[7] T. Bolkart and S. Wuhrer. A robust multilinear model learning framework for 3d faces. In *IEEE Conference on CVPR*, pages 4911–4919, 2016. 2

[8] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniahy, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *IEEE Conference on CVPR*, pages 5543–5552, 2016. 2

[9] S. Buyukatalay, O. Birgul, and U. Hahc. Effects of light sources selection and surface properties on photometric stereo error. In *Signal Processing and Communications Applications Conference*, pages 336–339, 2010. 1, 3

[10] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *Acm Transactions on Graphics*, 34(4):46, 2015. 2

[11] I. Cherabier, C. Häne, M. R. Oswald, and M. Pollefeys. Multi-label semantic 3d reconstruction using voxel blocks. In *International Conference on 3d Vision*, pages 601–610. IEEE, 2016. 8

[12] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *IEEE Conference on CVPR*, 2017. 2

[13] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conference on CVPR*, 2017. 2

[14] P. Garrido, L. Valgaert, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *Acm Transactions on Graphics*, 32(6):1–10, 2013. 4

[15] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. In *SIGGRAPH Asia Conference*, page 129, 2011. 1, 2

[16] C. Häne, C. Zach, A. Cohen, and M. Pollefeys. Dense semantic 3d reconstruction. *IEEE TPAMI*, 39(9):1730–1743, 2017. 8

[17] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni. Pooling faces: template based face recognition with pooled face images. In *Proceedings of the CVPR Workshops*, pages 59–67, 2016. 1, 7

[18] S. Herbort and C. Whler. An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods. *3d Research*, 2(3):1–17, 2011. 2

[19] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2008. 1

[20] L. Hu, H. Li, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, and Y. C. Chen. Avatar digitization from a single image for real-time rendering. *Acm Transactions on Graphics*, 36(6):1–14, 2017. 2

[21] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 1, 2, 3, 7

[22] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *IEEE Conference on CVPR*, 2017. 2

[23] I. Kemelmacher-Shlizerman. Internet based morphable model. In *IEEE ICCV*, pages 3256–3263, 2013. 2

[24] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE TPAMI*, 33(2):394–405, 2011. 4

[25] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, pages 1746–1753, 2011. 2

[26] 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 *Eurographics Conference on Rendering Techniques*, pages 183–194, 2007. 1, 2, 6

[27] T. A. Mancini and L. B. Wolff. 3 d shape and light source location from depth and reflectance. In *IEEE Conference on CVPR*, pages 707–709. IEEE, 1992. 4

[28] R. Mecca, A. Wetzler, A. M. Bruckstein, and R. Kimmel. Near field photometric stereo with point light sources. *Siam Journal on Imaging Sciences*, 7(4):2732–2770, 2014. 1

[29] J. Park, S. N. Sinha, Y. Matsushita, Y. W. Tai, and I. S. Kweon. Calibrating a non-isotropic near point light source using a plane. In *IEEE Conference on CVPR*, pages 2259–2266, 2014. 1, 3

[30] J. Park, S. N. Sinha, Y. Matsushita, Y. W. Tai, and I. S. Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE TPAMI*, 39(8):1591–1604, 2017. 2, 4, 6, 7

[31] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *IEEE Conference on CVPR*, pages 3418–3427, 2016. 2

[32] Y. Queau and J. D. Durou. Edge-preserving integration of a normal field weighted least-squares, tv and l1 approaches. In *Scale Space and Variational Methods for Computer Vision*, pages 576–588, 2015. 3

[33] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *International Conference on 3d Vision*, pages 460–469, 2016. 2

[34] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *IEEE Conference on CVPR*, pages 5553–5562, 2017. 2

[35] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *IEEE Conference on CVPR*, pages 4197–4206, 2016. 2

[36] D. Schnieders and K.-Y. K. Wong. Camera and light calibration from reflections on a sphere. *CVIU*, 117(10):1536–1547, 2013. 1, 3

[37] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on CVPR*, pages 519–528, 2006. 1

[38] T. Simchony, R. Chellappa, and M. Shao. Direct analytical methods for solving poisson equations in computer vision problems. *IEEE TPAMI*, 12(5):435–446, 1990. 2

[39] A. Sinha, A. Unmesh, Q. Huang, and K. Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *IEEE Conference on CVPR*, pages 791–800, 2017. 2

[40] L. Smith and M. Smith. The virtual point light source model the practical realisation of photometric stereo for dynamic surface inspection. In *International Conference on Image Analysis and Processing*, pages 495–502, 2005. 1

[41] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, pages 796–812, 2014. 2

[42] T. Takai, A. Maki, K. Niinuma, and T. Matsuyama. Difference sphere: an approach to near light source estimation. *CVIU*, 113(9):966–978, 2009. 1

[43] G. Trigeorgis, P. Snape, I. Kokkinos, and S. Zafeiriou. Face normals "in-the-wild" using fully convolutional networks. In *IEEE Conference on CVPR*, 2017. 2

[44] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *IEEE Conference on CVPR*, 2017. 2, 7

[45] L. Valgaerts, C. Wu, C. Theobalt, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *Acm Transactions on Graphics*, 31(6):187, 2012. 4

[46] M. Visentini-Scarzanella and H. Kawasaki. Simultaneous camera, light position and radiant intensity distribution calibration. In *Pacific-Rim Symposium on Image and Video Technology*, pages 557–571. Springer, 2015. 1, 3

[47] M. Weber and R. Cipolla. A practical method for estimation of point light-sources. In *Proceedings BMVC*, pages 1–10, 2001. 4

[48] K.-Y. K. Wong, D. Schnieders, and S. Li. Recovering light directions and camera poses from a single sphere. In *ECCV*, pages 631–642. Springer, 2008. 1, 3

[49] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):1–22, 1980. 2, 3

[50] C. Wu, K. Varanasi, Y. Liu, and H. P. Seidel. Shading-based dynamic shape refinement from multi-view video under general illumination. In *ICCV*, pages 1108–1115, 2011. 4

[51] R. Yu, S. Saito, H. Li, D. Ceylan, and H. Li. Learning dense facial correspondences in unconstrained images. *ICCV*, pages 4733–4742, 2017. 2

[52] Q. Zheng and R. Chellappa. Estimation of illuminant direction, albedo, and shape from shading. In *IEEE Conference on CVPR*, pages 540–545. IEEE, 1991. 4