**International Journal of Computer Vision manuscript No.**
(will be inserted by the editor)

# Shape and Albedo Recovery by Your Phone using Stereoscopic Flash and No-flash Photography

**Xu Cao · Michael Waechter · Boxin Shi · Ye Gao · Bo Zheng · Fumio Okura · Yasuyuki Matsushita**

Received: date / Accepted: date

**Abstract** Recovering shape and albedo for the immense number of existing cultural heritage artifacts is challenging. Accurate 3D reconstruction systems are typically expensive and thus inaccessible to many and cheaper off-the-shelf 3D sensors often generate results of unsatisfactory quality. This paper presents a high-fidelity shape and albedo recovery method that only requires a stereo camera and a flashlight, a typical camera setup equipped in many off-the-shelf smartphones. The stereo camera allows us to infer rough shape from a pair of no-flash images, and a flash image is further captured for shape refinement based on our flash/no-flash image formation model. We verify the effectiveness of our method on real-world artifacts in indoor and outdoor conditions using smartphones with different camera/flashlight configurations. Comparison results demonstrate that our stereoscopic flash and no-flash photography benefits the high-fidelity shape and albedo recovery on a smartphone. Using our method, people can immediately turn their phones into high-fidelity 3D scanners, facilitating the digitization of cultural heritage artifacts.

**Keywords** 3D reconstruction · stereo camera · flash photography · albedo · e-heritage

X. Cao, M. Waechter, F. Okura, Y. Matsushita
Osaka University, Graduate School of Information Science and Technology
E-mail: {cao.xu, okura, yasumat}@ist.osaka-u.ac.jp

B. Shi
Peking University, School of Computer Science and Institute for Artificial Intelligence & Peng Cheng Laboratory
E-mail: shiboxin@pku.edu.cn

Y. Gao, B. Zheng
Huawei Technologies Co., Ltd.

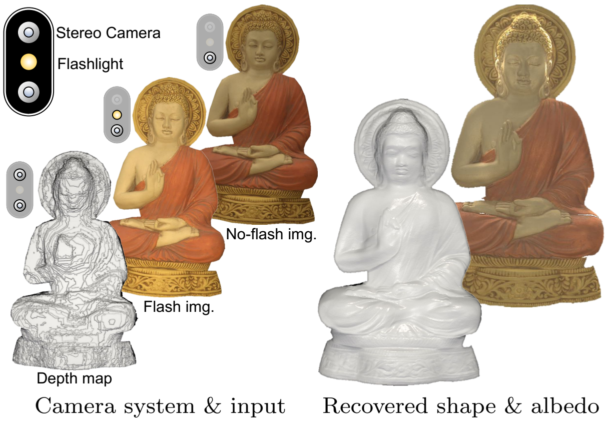

**Fig. 1** Our setup uses a stereo camera and a flashlight, which is common in modern smartphones, *e.g.*, the iPhone X from 2017. We capture a stereo image pair to infer a rough depth map and a flash/no-flash image pair to recover shape details and surface albedo.

## 1 Introduction

Recording 3D shape and surface reflectance are both invaluable for digitally archiving and analyzing cultural heritage artifacts. While the importance of digitally archiving artifacts is generally recognized, it is still not widely spread in many museums and libraries, mostly due to the complexity of the digitization process that comes with expensive specialized setups. To enable everybody to participate in digital archiving, a method that is simple to operate and only requires a commodity device is very much wanted.

With this goal in mind, this paper presents a high-fidelity shape and albedo recovery method using a simple imaging setup that is already available in widespread devices. Our method only requires a stereo camera and a flashlight as shown in Fig. 1, and takes three images in two shots from a fixed viewpoint as input: Two images in one shot by a stereo camera, and another image by one camera with a flashlight. By harnessing both geometric and photometric cues from the input images, our method recovers a fine 3D shape and a surface albedo map. Specifically, our method uses the rough shape inferred from the stereo image pair to estimate the no-flash environmental lighting. Using our flash/no-flash image formation model, the high-frequency details of the target scene are then recovered.

Unlike previous methods that rely on complex imaging setups [9, 41], our setup is minimal to introduce geometric and photometric cues. Other than an ordinary monocular camera, our method only requires one additional viewpoint (*i.e.*, a stereo camera) and lighting condition (*i.e.*, a flashlight). As will be shown later, further reducing any input significantly downgrades recovery. Fortunately, many commodity smartphones today are equipped with this imaging setup, and we will demonstrate later in this paper that our method is naturally applicable to such smartphones. With this setup, recording can be conducted outside a darkroom (*e.g.*, in an office room) and completed in a moment as it only takes two shots without any camera movement. These properties make the digitization process easy.

The key contributions of our work are as follows:

- A high-fidelity shape and albedo recovery method working with a simple, compact, and wide-spread imaging setup;
- A flash/no-flash image formation model for Lambertian surfaces with non-uniform albedos under natural lighting;
- A robust shape and albedo recovery method that harnesses both geometric and photometric cues.

This paper extends the preliminary version of our work [7] in three important aspects: First, we generalize the image formation model to flash/no-flash image pairs captured with different camera exposure settings. This generalization is crucial for the successful application using off-the-shelf devices (see Sec. 3.1). Second, we verify the effectiveness of our imaging setup and recovery method using off-the-shelf smartphones (in Sec. 4.2). Finally, we show more examples of reconstruction including outdoor objects.

## 2 Related work

Our reconstruction method is related to shading-based shape recovery and flash photography.

*Shading-based shape recovery:* Geometric shape recovery approaches such as stereo are useful for recovering a coarse shape but have fundamental limitations in recovering high-frequency details [23]. In contrast, photometric approaches can recover per-pixel surface normals using shading cues in the images. In the past, various approaches have been proposed for high-quality shape recovery by combining the strengths of both geometric and photometric approaches. For example, Ikeuchi [21] recovers the depth map from a stereo pair of normal maps, which are estimated by photometric stereo with three lights.

While photometric approaches commonly assume controlled lighting conditions without ambient lighting, when they are combined with geometric approaches this assumption is likely violated and they face more challenging lighting conditions. Basri *et al.* [3] verified that for a Lambertian surface its reflectance can be modeled as a low-dimensional linear combination of spherical harmonics. Photometric stereo under natural illumination has been shown to be feasible after this theoretical verification [4, 22]. Such approaches have been incorporated into geometric approaches. An algorithmic structure of such combinations is to estimate a coarse depth map, then estimating illumination and albedo from the coarse depth map, followed by an optimization including but not limited to depth, shading, and smoothness constraints [31, 37, 39, 40]. Estimating global spherical harmonics coefficients usually fails in local areas where cast shadows or specularities dominate the intensity. To alleviate this problem, Han *et al.* [17] split illumination into a global and a local part, Or-El *et al.* [29] handled local illumination based on first-order spherical harmonics, and Maier *et al.* [27] proposed spatially-varying spherical harmonics. Besides a single color image, photometric cues from different types of input have been used to improve the reconstruction quality, for example, from infrared images [9, 18], from RGB-D streams [36, 38], or from multiple view images [14, 28].

Our work uses a simpler setup consisting of a stereo camera and a flashlight. With two shots, our method recovers fine geometry for Lambertian objects under natural lighting.

*Flash photography:* Images taken with a flashlight have been used for various computer vision tasks. Using the light falloff property, a flash and no-flash image pair has been used for image matting [34], foreground extraction [35], and saliency detection [19]. Under low-light conditions, a flash image captures high-frequency details but changes the overall appearance of the scene, while the no-flash image captures the overall environmental ambiance but is noisy. This complementary property has been used in photography enhancement

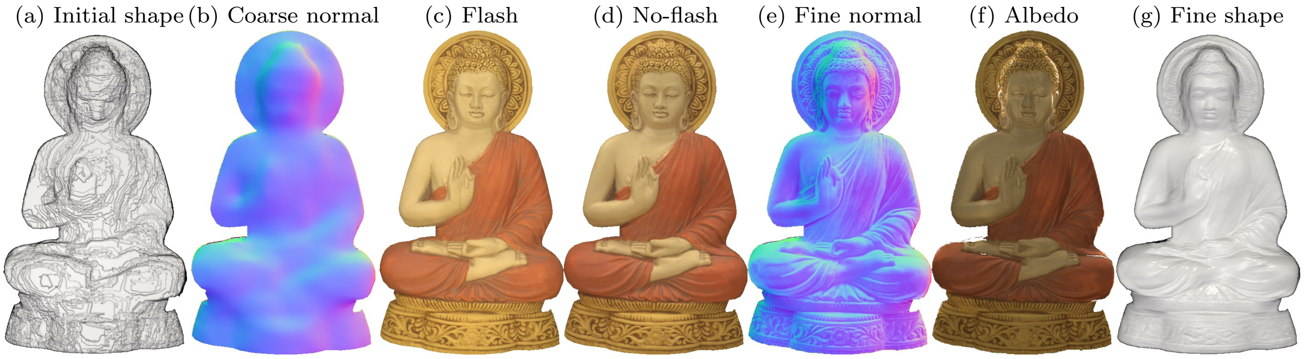

**Fig. 2** Pipeline of our method. Given (a) an initial rough shape from a stereo camera, we first estimate (b) a coarse normal map. With (c) the flash and (d) the no-flash image, we optimize for (e) a fine normal map. Finally, we compute (f) the albedo map and perform depth normal fusion to obtain (g) the fine shape. Section 3.2 details each step.

under dark illumination [12], denoising, detail transfer, or white balancing [30].

Further, photometric cues introduced by a flashlight are useful in stereo matching. Feris *et al.* [13] demonstrated that the shadows cast by a flashlight along depth discontinuities help to detect half-occlusion points in stereo matching. Zhou *et al.* [42] showed the ratio of a flash/no-flash pair can make stereo matching robust against depth discontinuities. In addition, flash images are used for recovering spatially varying BRDFs (SVBRDFs). A single image captured from a flash-enabled camera, or a flash/no-flash pair [1] is used for SVBRDF and shape recovery of near-planar objects [2, 11, 24] or those with complex geometry [25].

Our work differs from the previous works in that we explicitly parameterize the image observation lit by a flashlight, and use the flash/no-flash image pair to derive an albedo-free image formation model for geometry refinement.

## 3 Proposed method

Figure 2 illustrates our method for shape and albedo recovery. The input to our method are (a) a rough depth map inferred from a stereo image pair taken by a stereo camera and (c)+(d) a flash/no-flash image pair taken by the stereo camera's reference camera. First, we compute a coarse surface normal map from the depth map as shown in (b). We then estimate the environmental lighting and refine the normal map based on our flash/no-flash image formation model as in (e). Finally, we fuse the fine normal map (e) and the coarse depth map (a) to obtain the fine shape (f) and compute the albedo map (g).

In the following, Sec. 3.1 describes our image formation model for the flash/no-flash image pair and Sec. 3.2 details the design choices of each step in our method.

### 3.1 Image formation model

Assuming Lambertian reflectance, the radiance $r \in \mathbb{R}_+$ emitted from a tiny surface patch can be modeled as

$$r = \rho\, s(\mathbf{n}), \tag{1}$$

where a shading function $s : \mathcal{S}^2 \to \mathbb{R}$ depends on the environmental lighting and is scaled by the surface albedo $\rho \in \mathbb{R}_+$. The shading function $s$ is applied to the unit surface normal $\mathbf{n} \in \mathcal{S}^2 \subset \mathbb{R}^3$.

Let $m \in \mathbb{R}_+$ be the recorded brightness of the radiance by a digital camera. Assume the camera has a linear radiometric response, say 1 for simplicity, to the radiance. The intensity $m$ is then the scene radiance $r$ scaled by the camera exposure $c \in \mathbb{R}_+$ as

$$m = cr. \tag{2}$$

The camera exposure $c$ accounts lens-aperture, ISO, and exposure time.

Now consider that a flash/no-flash image pair is taken for an object by the same camera. Assume that the viewpoint is fixed, the object is static, and the environmental lighting stays the same during the capture. A pixel at a fixed location in the flash/no-flash image pair then records the radiance from the same surface patch, scaled by possibly different camera exposures. We use the subscript "nf" and "f" to indicate the no-flash and flash images, respectively. Using Eqs. (1) and (2), we can model the intensity recorded at the same pixel location in the flash/no-flash pair as

$$\begin{cases} m_{\mathrm{nf}} = c_{\mathrm{nf}} \rho s_{\mathrm{nf}}, \\ m_{\mathrm{f}} = c_{\mathrm{f}} \rho (s_{\mathrm{nf}} + s_{\mathrm{fo}}). \end{cases} \tag{3}$$

The additional shading $s_{\mathrm{fo}}$ is introduced by the flashlight (the subscript "fo" represents flash-only), which is identical to the shading if the flashlight were the only light source in the scene. Let $\gamma = \frac{c_{\mathrm{f}}}{c_{\mathrm{nf}}}$ be the ratio of the

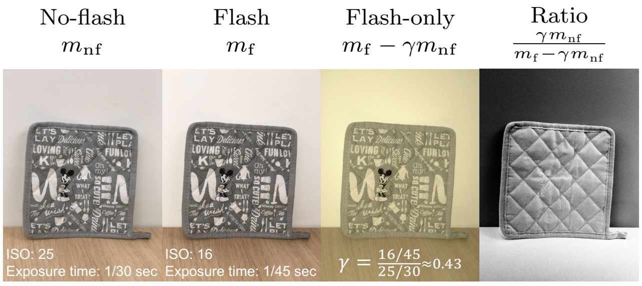
No-flash
$m_{\text{nf}}$
Flash
$m_{\text{f}}$
Flash-only
$m_{\text{f}} - \gamma m_{\text{nf}}$
Ratio
$\frac{\gamma m_{\text{nf}}}{m_{\text{f}} - \gamma m_{\text{nf}}}$
ISO: 25
Exposure time: 1/30 sec
ISO: 16
Exposure time: 1/45 sec
$\gamma = \frac{16/45}{25/30} \approx 0.43$

**Fig. 3** Subtracting the $\gamma$-scaled no-flash image from the flash image yields a virtual flash-only image. The ratio image is obtained by dividing the gray-scale no-flash image by the gray-scale flash-only image.

flash image's exposure to the no-flash image's exposure. By modifying Eq. (3), we obtain

$$\begin{cases} m_{\text{nf}} = c_{\text{nf}} \rho s_{\text{nf}}, \\ m_{\text{f}} - \gamma m_{\text{nf}} = c_{\text{f}} \rho s_{\text{fo}}. \end{cases} \tag{4}$$

The second equation implies a virtual flash-only image: The computed intensity $m_{\text{f}} - \gamma m_{\text{nf}}$ is the flash-only shading scaled by the albedo and the flash image's exposure. Figure 3 exemplifies a virtual flash-only image. Notice that the shadows caused by natural lighting disappear in the flash-only image, verifying the correctness of the subtraction.

Further taking the ratio of the two equations in Eq. (4) yields

$$\frac{\gamma m_{\text{nf}}}{m_{\text{f}} - \gamma m_{\text{nf}}} = \frac{s_{\text{nf}}}{s_{\text{fo}}}. \tag{5}$$

The division cancels out the unknown albedo $\rho$; therefore, our method can naturally handle spatially-varying albedos unlike previous methods that assume piece-wise uniform albedos [15, 16]. This albedo-free image formation model directly relates the shading to the measured intensity. The effect of this albedo canceling is illustrated in Fig. 3. While surface albedo of the mat has a complex spatial variation, only the shading information remains in the ratio image.

Explicitly modeling the camera exposure in the image formation model of Eq. (5) has practical merit. Using the identical exposure ($\gamma = 1$) in [7] is a special case of the image formation model of Eq. (5); however, in practice it causes overexposure in the flash image or underexposure in the no-flash image. Equation (5) allows us to properly expose each image in the flash/no-flash pair.

*Shading model:* We now discuss how we model the no-flash shading $s_{\text{nf}}$ and the flash-only shading $s_{\text{fo}}$. Suppose a light ray in direction $\mathbf{l} \in \mathcal{S}^2 \subset \mathbb{R}^3$ with intensity $e : \mathcal{S}^2 \to \mathbb{R}$ hits a surface patch. According to the Lambert's law, the reflected light or shading, is given by

$$s(\mathbf{n}) = e(\mathbf{l}) \max(\mathbf{n}^\top \mathbf{l}, 0). \tag{6}$$

Under natural lighting, light rays reach the surface patch from infinitely many directions. The shading then becomes the integral over all possible incident directions

$$s(\mathbf{n}) = \int_{\mathcal{S}^2} e(\mathbf{l}) \max(\mathbf{n}^\top \mathbf{l}, 0) \, \mathrm{d}\mathbf{l}. \tag{7}$$

As studied in [3, 33], a Lambertian surface acts as a low-pass filter, and its shading under natural lighting is well characterized by the second-order spherical harmonics, *i.e.*, the integral in Eq. (7) can be approximated by a linear combination of the second-order spherical harmonics. Denoting the unit surface normal $\mathbf{n} = [n_1, n_2, n_3]^\top$, the spherical harmonics up to the second order can be stacked into a vector $\mathbf{h}(\mathbf{n})$ as

$$\mathbf{h}(\mathbf{n}) = [1, n_1, n_2, n_3, n_1 n_2, n_2 n_3, n_3 n_1, n_1^2 - n_2^2, 3n_3^2 - 1]^\top.$$

The shading under no-flash illumination $s_{\text{nf}}$ is then a linear combination of these spherical harmonics. Stacking the 9 coefficients into a vector $\mathbf{l}_{\text{nf}} \in \mathbb{R}^9$ yields

$$s_{\text{nf}} = \mathbf{h}(\mathbf{n})^\top \mathbf{l}_{\text{nf}}. \tag{8}$$

Note that $\mathbf{l}$ and $\mathbf{l}_{\text{nf}}$ are different; $\mathbf{l}$ is a light ray direction, and $\mathbf{l}_{\text{nf}}$ is a stack of spherical harmonic coefficients.

For the flashlight, we assume it is a point light located at the optical center of the camera. The incident light direction $\mathbf{l}$ is thus the same as the camera's viewing direction $\mathbf{v}$ for each surface patch. We further assume that the flashlight emits light uniformly in all directions and the light fall-off effect is negligible. As the flashlight is the only light source contributing to the shading $s_{\text{fo}}$, Eq. (6) can be applied. Let $e_{\text{f}}$ be the flashlight intensity. Equation (6) then reads

$$s_{\text{fo}} = e_{\text{f}} \max(\mathbf{n}^\top \mathbf{l}, 0) = e_{\text{f}} \max(\mathbf{n}^\top \mathbf{v}, 0) = e_{\text{f}} \mathbf{n}^\top \mathbf{v}. \tag{9}$$

We can drop the $\max(\cdot, 0)$ term because $\mathbf{n}^\top \mathbf{v}$ is always greater than 0 if the surface patch is visible to the camera. Inserting Eqs. (8) and (9) into Eq. (5) yields

$$\frac{\gamma m_{\text{nf}}}{m_{\text{f}} - \gamma m_{\text{nf}}} = \frac{\mathbf{h}(\mathbf{n})^\top \mathbf{l}'}{\mathbf{n}^\top \mathbf{v}}, \tag{10}$$

where $\mathbf{l}' = \mathbf{l}_{\text{nf}} / e_{\text{f}}$ is the spherical harmonics coefficient vector scaled by the flashlight intensity, and we will call $\mathbf{l}'$ global lighting vector. This final image formation model now explicitly relates surface normal and environmental lighting to the measured intensity.

### 3.2 Shape and albedo recovery

This section details the design choice for each step in our shape and albedo recovery method shown in Fig. 2.

*Obtaining coarse surface normals:* We compute the initial normal map from the depth map using PlanePCA [20]. Given the camera intrinsics, we convert the depth map into a point cloud in camera coordinates and then find each point's surface normal by fitting a plane to its nearest neighbors. Formally, given a set of points $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n \mid \mathbf{p}_i \in \mathbb{R}^3\}$, we find the coarse surface normal vector $\hat{\mathbf{n}}_i$ at $\mathbf{p}_i$ by minimizing

$$\hat{\mathbf{n}}_i = \underset{\hat{\mathbf{n}}_i}{\text{argmin}} \sum_{\mathbf{p}_j \in \mathcal{N}(\mathbf{p}_i)} (\mathbf{p}_j - \bar{\mathbf{p}}_i)^\top \hat{\mathbf{n}}_i, \tag{11}$$

where $\mathcal{N}(\mathbf{p}_i)$ is the set of $\mathbf{p}_i$'s neighbors, and $\bar{\mathbf{p}}_i$ is the mean of all $\mathbf{p}_j \in \mathcal{N}(\mathbf{p}_i)$. We search for $\mathbf{p}_i$'s neighbors by performing a ball query as

$$\mathcal{N}(\mathbf{p}_i) = \{\mathbf{p}_j \mid \|\mathbf{p}_j - \mathbf{p}_i\|_2 < r, \ \forall \mathbf{p}_j \in \mathbf{P}\}, \tag{12}$$

where $r$ is an empirically chosen ball search radius. PlanePCA robustly estimates a coarse, smooth normal map that expresses low-frequency shape which we use in the following lighting estimation step.

*Computing the global lighting vector:* Our goal now is, given the flash/no-flash image pair and a coarse normal map, to estimate the low-dimensional global lighting vector $\mathbf{l}'$ in Eq. (10). Note that solving $\mathbf{l}_{\text{nf}}$ and $e_{\text{f}}$ separately is unnecessary for shape recovery; unknown $e_{\text{f}}$ barely scales the recovered albedo map.

Suppose there are $p$ pixels in the region of interest, *i.e.*, the region of the foreground object. We stack the row vectors $\mathbf{h}(\hat{\mathbf{n}})^\top / \hat{\mathbf{n}}^\top \mathbf{v}$ for each pixel vertically into a matrix $\mathbf{N} \in \mathbb{R}^{p \times 9}$ and stack the measured $\gamma m_{\text{nf}}/(m_{\text{f}} - \gamma m_{\text{nf}})$ into a vector $\mathbf{m} \in \mathbb{R}^p$. $\mathbf{l}'$ can be obtained by solving the following over-determined system

$$\mathbf{N}\mathbf{l}' = \mathbf{m}. \tag{13}$$

Although the coarse normal map only expresses a low-frequency shape, we will show in the experiment that the estimated lighting is still as accurate as if it is estimated from a ground truth normal map.

*Refining the normal map:* We formulate the surface normal refinement as per-pixel optimization. The energy function consists of a shading constraint, a surface normal constraint, and a unit-length constraint as

$$\min_{\mathbf{n}} E_s(\mathbf{n}) + \lambda_1 E_n(\mathbf{n}) + \lambda_2 E_u(\mathbf{n}), \tag{14}$$

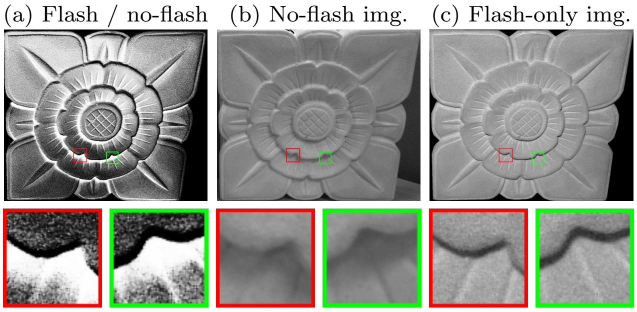

**Fig. 4** The relation between the ratio of flash to no-flash images and cast shadows. Large ratios (bright pixels in (a)) are likely caused by cast shadows under environmental lighting (b); tiny ratios (dark pixels in (a)) are caused by cast shadows under flashlight (c).

where $\lambda_1$ and $\lambda_2$ are weighting factors. The shading constraint $E_s$ measures the squared difference between the ratio image and the estimated ratio image in Eq. (10)

$$E_s(\mathbf{n}) = \left( \mathbf{h}(\mathbf{n})^\top \mathbf{l}' - \mathbf{n}^\top \mathbf{v} \frac{\gamma m_{\text{nf}}}{m_{\text{f}} - \gamma m_{\text{nf}}} \right)^2. \tag{15}$$

We multiply both sides of Eq. (10) with $\mathbf{n}^\top \mathbf{v}$ to avoid possible numerical issues.

The surface normal constraint $E_n$ forces the refined surface normal to be close to the coarse surface normal $\hat{\mathbf{n}}$, *i.e.*, their dot-product should be close to 1

$$E_n(\mathbf{n}) = (1 - \mathbf{n}^\top \hat{\mathbf{n}})^2. \tag{16}$$

Finally, $E_u$ enforces unit length of the surface normal

$$E_u(\mathbf{n}) = (1 - \mathbf{n}^\top \mathbf{n})^2. \tag{17}$$

The energy function Eq. (14) is non-convex due to the non-convex domain $\mathcal{S}^2$. We solve it with BFGS [26].

After optimizing the normal map we can compute the albedo map: According to Eq. (3) and Eq. (8),

$$\rho = \frac{m_{\text{nf}}}{c_{\text{nf}} \mathbf{h}(\mathbf{n})^\top \mathbf{l}_{\text{nf}}} = \frac{e_{\text{f}} m_{\text{nf}}}{c_{\text{nf}} \mathbf{h}(\mathbf{n})^\top \mathbf{l}'}. \tag{18}$$

A global scalar ambiguity remains in the albedo due to the camera exposure $c_{\text{nf}}$ and flashlight intensity $e_{\text{f}}$.

*Handling cast shadows (optional):* The spherical harmonics-based image formation model of Eq. (6) can handle attached shadows but not cast shadows [3]. Our method is thus likely to break down and produce artifacts in regions dominated by cast shadows. In such regions, instead of refining normals using our shading constraints, the initial normal vector estimated from the depth map is more reliable. To this end, we heuristically introduce a confidence term $\omega$ into the energy function's shading constraint as

$$\min_{\mathbf{n}} \omega E_s(\mathbf{n}) + \lambda_1 E_n(\mathbf{n}) + \lambda_2 E_u(\mathbf{n}), \tag{19}$$

where $\omega$ is defined as

$$\omega = \exp\Big(-\frac{(r-\mu)^2}{2\sigma^2}\Big). \tag{20}$$

$r$ is the ratio of the flash to no-flash intensities, and $\mu$ and $\sigma$ are the mean and the standard deviation of the ratio in the object region. This definition is based on the observation that cast shadows strongly deviates the ratio $r$ from the mean ratio. From Eq. (5), once the pixel intensity is distorted by cast shadow under environmental light or flashlight, the numerator or denominator becomes close to zero, yielding too small or too large ratio values. This phenomenon is shown in Fig. 4. When environmental light causes shadows, the ratio of flash to no-flash becomes high (bright pixels in Fig. 4(a)); when flashlight causes shadows, the ratio becomes low (dark pixels in Fig. 4(a)).[1]

The above observation leads to the choice of $\omega$ in Eq. (20). For pixels where the ratio deviates too much from the mean ratio, the shading constraint in Eq. (19) is unlikely reliable. The weight $\omega$ should be small according to Eq. (20) so that the shading constraint contributes less to the normal refinement. As a result, the normal vector stays close to the initial one.

*Fusing the normal and the depth map:* Finally, we fuse the fine normal map with the coarse shape to obtain the fine shape. To this end, we minimize the weighted sum of normal integration and depth terms.

For the normal integration term, we follow the inverse plane fitting method [8] to minimize the sum of plane fitting residuals as

$$E_n(\mathbf{z},\mathbf{d}) = \sum_i \sum_{j\in\mathcal{N}(i)} (z_j \mathbf{n}_i^\top \mathbf{K}^{-1}\tilde{\mathbf{u}}_j + d_i)^2, \tag{21}$$

where $\mathbf{z}$ and $\mathbf{d}$ are the vectorized depth map and plane distances to the coordinate origin, respectively. $\mathcal{N}(i)$ is the pixel $i$ and its four neighborhoods; $z_j$, $\mathbf{n}_i$, $\mathbf{u}_j$, and $d_i$ are the $j$-th entry in $\mathbf{z}$, the normal vector at pixel $i$, the homogeneous coordinates of pixel $j$, and the $i$-th entry in $\mathbf{d}$, respectively. $\mathbf{K} \in \mathbb{R}^{3\times 3}$ is the perspective camera intrinsic matrix. The inner term of Eq. (21) measures the distance of the 3D point $z_j\mathbf{K}^{-1}\tilde{\mathbf{u}}_j$ to the plane, which is parameterized by its normal direction $\mathbf{n}_i$ and its distance $d_i$ to the coordinate origin. For the depth term, we force the estimated depth $\mathbf{z}$ to be close to the initial depth $\hat{\mathbf{z}}$

$$E_d(\mathbf{z}) = \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 . \tag{22}$$

The whole objective now reads

$$\min_{\mathbf{z},\mathbf{d}} E_n(\mathbf{z},\mathbf{d}) + \lambda_d E_d(\mathbf{z}), \tag{23}$$

where $\lambda_d$ is a weighting factor to be tuned. Equation (23) can be formed as a sparse linear system, and we use a multigrid method [6] to find its solution.

## 4 Experiments

This section evaluates our shape and albedo recovery results quantitatively on synthetic images and qualitatively using real-world images captured with iPhones.

### 4.1 Experiments using synthetic images

*Data generation:* We rendered two publicly available 3D mesh models, the Stanford BUNNY and a STATUE[2] with the physically-based renderer Mitsuba[3]. For the no-flash image, we put each object under an environment map lighting[4]. We then simulate the flashlight by placing an additional directional light source in the same scene. We obtain the objects' ground truth shape, depth maps, and normal maps from the 3D models. To simulate the coarse shape from a stereo camera, we apply the quantization on the ground truth depth map. For the ground truth albedo, we use a texture image. To visualize the refinement of the estimated albedo map, we also compute the initial albedo according to Eq. (18) using the coarse normal map.

*Baselines:* Although our setup combining a depth measurement with flash/no-flash image pairs is new and has no direct comparison methods, we assess our shape reconstruction results with the recent depth refinement methods by Han *et al.* [17] and Yan *et al.* [39]. Unlike ours, both baseline methods refine the initial shape using a single color image (*i.e.*, without flash/no-flash image pairs). We therefore aim to verify the effectiveness of our use of flash/no-flash pairs via this comparison.

We implemented [17] as their source code is not publicly available. For Yan *et al.*'s method [39], we used a trained convolutional neural network provided by the

[1] The flashlight can cause shadows because in practice its location is non-identical to the camera's optic center.

[2] "The Getty Caligula" by CosmoWenman / CC BY 4.0. `https://sketchfab.com/3d-models/the-getty-caligula-6bd927a5779d479e83303635c79f81ac`, last accessed on April 1, 2021

[3] Mitsuba Renderer. `https://www.mitsuba-renderer.org/index_old.html`, last accessed on April 1, 2021

[4] High-Resolution Light Probe Image Gallery. `http://vgl.ict.usc.edu/Data/HighResProbes/`, last accessed on April 1, 2021

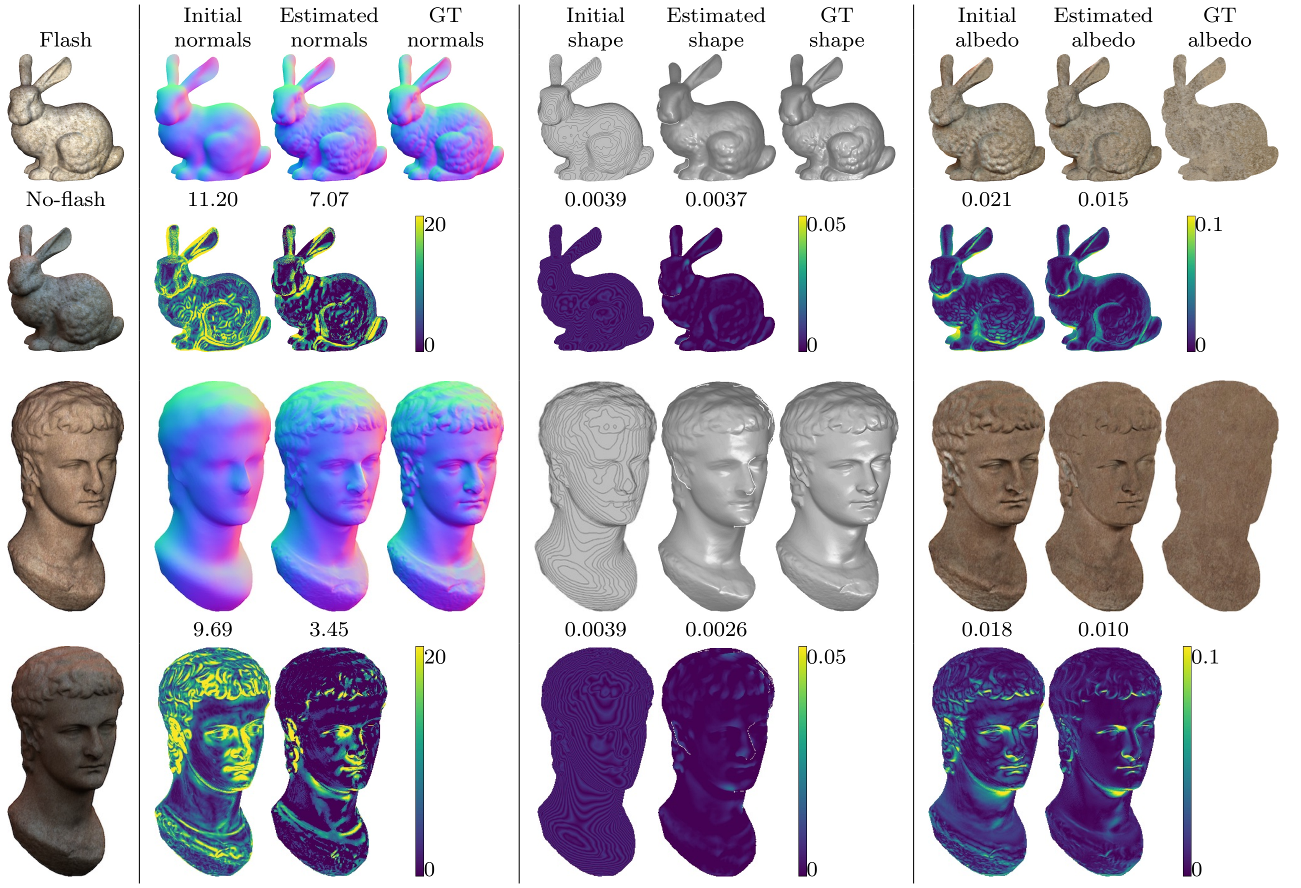

**Fig. 5** Shape and albedo recovery results on the synthetic Bunny and Statue datasets. The first column shows the rendered flash/no-flash pair. The even rows display the error map. The numbers above the error maps are the mean angular error (MAngE) of normal maps and the mean absolute error (MAbsE) of shape and albedo maps. Our method recovers high-frequency shape details.

**Table 1** MAbsE of the depth maps recovered by different methods. Two objects, Bunny and Statue, are rendered under three environmental lighting maps. "w/ conf." means using Eq. (19) for optimization.

| Env. map | Method | Bunny | Statue |
|---|---|---|---|
| Pisa | Han *et al.* [17] | 3.56e-3 | 3.59e-3 |
| | Yan *et al.* [39] | 4.02 | 1.26 |
| | Ours | 3.43e-3 | **2.42e-3** |
| | Ours w/ conf. | **3.39e-3** | 2.48e-3 |
| Doge | Han *et al.* [17] | 3.66e-3 | 3.68e-3 |
| | Yan *et al.* [39] | 4.02 | 1.26 |
| | Ours | 3.54e-3 | 3.09e-3 |
| | Ours w/ conf. | **3.44e-3** | **2.98e-3** |
| Glacier | Han *et al.* [17] | 3.65e-3 | 3.64e-3 |
| | Yan *et al.* [39] | 4.02 | 1.26 |
| | Ours | 3.45e-3 | 3.69e-3 |
| | Ours w/ conf. | **3.41e-3** | **3.64e-3** |

authors[5]. For a fair comparison, we use the uniform albedo maps for all objects, since the baseline methods assume the uniformness while our method is capable of spatially-varying albedos. We also use the same initial normal map and shape for all three methods. We measured the mean absolute error (MAbsE) between the estimated and the ground truth shape.

*Results:* Table 1 summarizes the results of the quantitative comparison with the two baseline methods (Han *et al.* [17] and Yan *et al.* [39]). Our method using flash/no-flash image pairs achieves the lowest MAbsE among all methods. Further, the confidence term $\omega$ in the energy function improves the results by our method in most cases, which verifies the effectiveness of our strategy for handling cast shadows.

Figure 5 shows shape and albedo recovery results by our method along with their coarse initializations and the ground truth. We also show the mean angular error (MAngE) of normal maps and MAbsE of shape and albedo maps. While the coarse normal maps contain only low-frequency content, our method recovers

[5] `https://github.com/neycyanshi/DDRNet`, last accessed on April 1, 2021

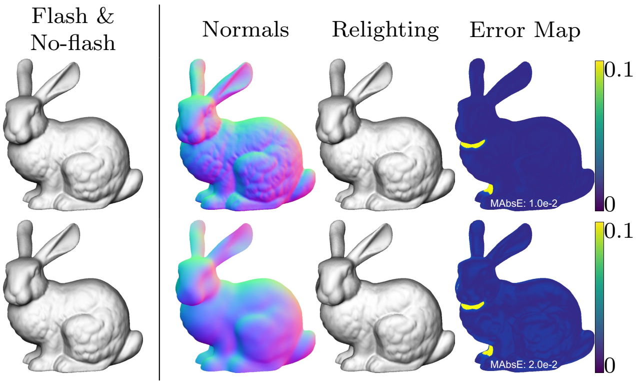

**Fig. 6** Lighting estimation from synthetic flash/no-flash images. Both relighting images are computed using GT normals and spherical harmonic coefficients, estimated from (the first row) GT normals or (the second row) coarse normals. The major approximation error exists in the cast shadow. Estimating spherical harmonic coefficients from coarse normals achieves a comparable relighting result, verifying the correctness of our lighting estimation using coarse normals.

high-frequency details and yields lower errors than the initializations. This verifies the effectiveness of the optimization Eq. (14) based on our flash/no-flash image formation model Eq. (10). After the depth normal fusion, the shape also reflects the recovered details. The albedo map still appears to have shading components left due to the approximation error of the second-order spherical harmonics and the estimation error introduced by cast shadow in practice. But the error of the estimated albedo is smaller than that of the initial albedo. This quantitative evaluation justifies our shape and albedo recovery pipeline.

Figure 6 shows lighting estimation results on synthetic data. We render the flash/no-flash images of the Stanford Bunny with uniform albedo. To verify that estimating spherical harmonic coefficients $\mathbf{l}_{\text{nf}}$ from coarse normals is reliable, we compare the relighting images using coefficients estimated from ground truth and coarse normals. We estimate the flashlight intensity scaled coefficients $\mathbf{l}'$ by Eq. (13), use the coefficients to compute the relighting images by Eq. (8), and compute the absolute error maps between the relighting and no-flash images. For both relighting images, we compute the spherical harmonic bases $\mathbf{h}(\mathbf{n})$ from ground truth normals. We cancel the scale ambiguity between $\mathbf{l}'$ and $\mathbf{l}_{\text{nf}}$ using the rendered no-flash image when visualizing the relighting images and computing the absolute error maps. As the spherical harmonics approximates the shading and assumes no cast shadow, the absolute error map shows that the approximation error is inevitable and mainly exists in cast shadow regions. The comparable relighting results verify that using initial coarse geometry for spherical harmonics estimation is reliable.

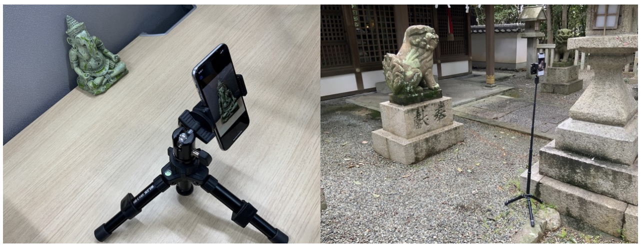

**Fig. 7** Indoor and outdoor image capturing with phones.

## 4.2 Experiments using smartphones

The camera system we require has become standard in modern smartphones. For example, iPhone models support stereo-based depth capture since the iPhone X released in 2017. This section describes shape and albedo recovery results from images captured by iPhones. To verify our method in practical scenarios, we captured small statues indoors as well as outdoor stone statues in an old shrine. Figure 7 shows the scenes of our image capture in indoor and outdoor environments using an iPhone X. Our method is handy to use as the recording only requires mounting a smartphone on a tripod.

*Image capturing and preprocessing:* We implemented a custom iOS application to control the image capture pipeline. Instead of capturing a stereo image pair and performing stereo matching by ourselves, we directly acquire the depth map via Apple's API[6]. Due to API limitations, when the stereo camera is used for depth map capture, raw image delivery is unsupported. We instead take a no-flash image one more time to acquire a raw image. In summary, one scene capture using an iPhone required three shots

- a depth map associated with the intrinsic parameters from the stereo camera,
- a raw flash image from the reference camera, and
- a raw no-flash image from the reference camera.

The flash/no-flash images are taken in auto-exposure mode, and the exposure ratio $\gamma$ is computed from the EXIF tags.

The dimensions of acquired depth maps and flash/no-flash images are $768 \times 576$ and $4032 \times 3024$, respectively. To close the resolution gap, we unify their dimensions to $1008 \times 756$ by rescaling. Specifically, we upsample the depth map with bi-cubic interpolation and downsampled the flash/no-flash images with inter-area interpolation. The intrinsic camera parameters (focal

[6] AVDetphData. `https://developer.apple.com/documentation/avfoundation/avdepthdata`, last accessed on April 1, 2021.

length and principal point coordinates) are scaled accordingly. As an implementation detail, we found that the depth map from the stereo camera and the color images from the reference camera are misaligned. Fortunately, we empirically found the misalignment was a simple fixed offset, therefore shifted the pixels in the flash/no-flash image pairs to align with the depth map.

*Baselines:* In addition to the quantitative comparisons by the synthetic dataset, we also compare our results visually with two shape and reflectance estimation methods by Haefner *et al.* [15] and Boss *et al.* [5]. Our method takes as input flash/no-flash images and a depth map, while the baseline methods do not use all the cues. We simulate Haefner *et al.*'s setup [15], which uses a color image and a depth map, by removing the no-flash image from our input. Boss *et al.*'s [5] setup, which uses a flash/no-flash image pair, was simulated by removing the depth map from our input.

We used the implementations released by the authors[7]. For Haefner *et al.*'s method [15], we followed their default parameter settings and used a 1008 × 756 flash image and a 768 × 576 depth map as input. Since their method does not directly output a normal map, we computed normal maps [32] from the estimated depth maps. For Boss *et al.*'s method [5], we used their trained neural network. To fit the 256 × 256 input image dimension, we cropped and downsampled our flash/no-flash images. As Boss *et al.* [5] estimates Cook-Torrance model parameters [10] as diffuse, roughness, and specular, we show the estimated diffuse maps and treat them as albedo maps for notational simplicity.

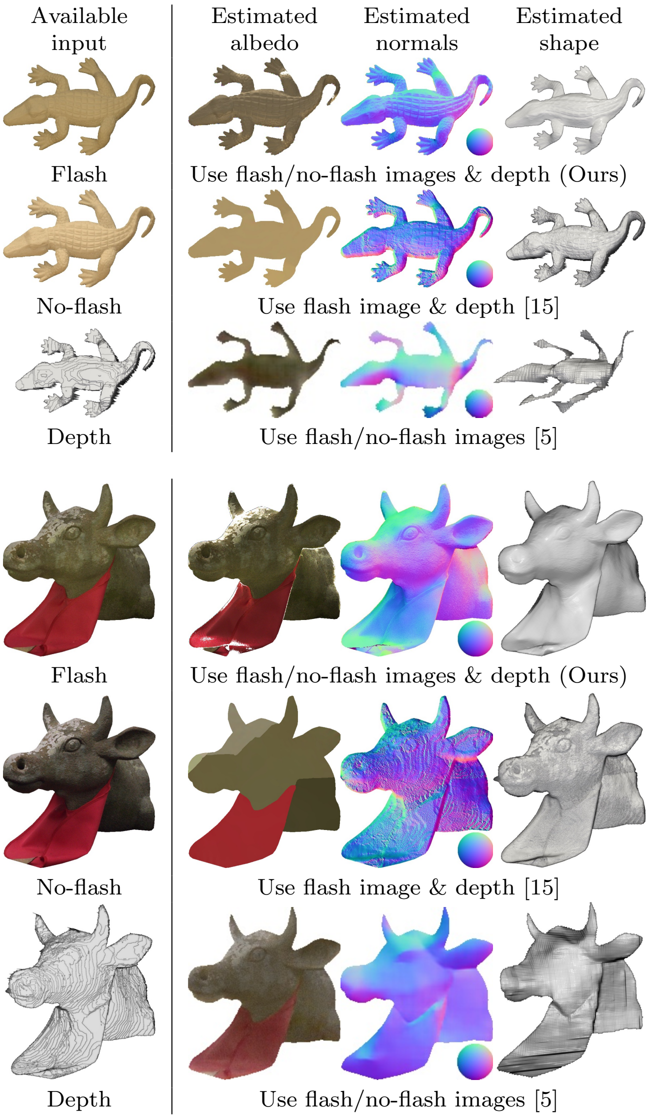
Available input
Estimated albedo
Estimated normals
Estimated shape
Flash
Use flash/no-flash images & depth (Ours)
No-flash
Use flash image & depth [15]
Depth
Use flash/no-flash images [5]
Flash
Use flash/no-flash images & depth (Ours)
No-flash
Use flash image & depth [15]
Depth
Use flash/no-flash images [5]

**Fig. 8** Visual comparison on an iPhone's input. We use all three input images: Flash/no-flash images and a depth map. Removing the no-flash image leads to Haefner *et al.*'s setup [15], which assumes piece-wise constant albedo and is not suitable for surfaces with complex albedo variation. Removing the depth map leads to Boss *et al.*'s setup [5], which is ill-posed and results in distorted shape estimation. Stereoscopic flash and no-flash photography is key for high-fidelity shape and albedo recovery via a smartphone.

*Results:* Figure 8 shows a visual comparison using the input from an iPhone X. Overall, our setup combining flash/no-flash imaging and a rough depth map yields the high-fidelity shape and albedo recovery. Haefner *et al.*'s method [15] assumes the piece-wise constant albedo. We thus observe noises on the estimated shape when the surface albedo has a complex spatial variation (see the stone cow in Fig. 8). Boss *et al.*'s method [5] explores shading information from only two images, which is inherently ill-posed. As a consequence, the estimated shapes are distorted; for example, concave surfaces can be wrongly estimated as convex, which can be seen in the stone cow's ear.

Figure 9 displays visual results by our method for cultural heritage artifacts. The first three objects are about 10 cm high and were captured in an office room (Fig. 7, left). Although there is no access to the ground truth, our method qualitatively recovers the fine details that are absent in the initial shape derived from the stereo camera despite of the complex albedo. The last three rows of Fig. 9 show stone statues in an old shrine, which are fixed in place outdoors and impossible to move. With our stereoscopic flash/no-flash photography, we can recover fine shapes of such outdoor objects with a commodity smartphone without requiring special lighting equipment or a darkroom.

[7] DepthSRfromShading. `https://github.com/BjoernHaefner/DepthSRfromShading`; Two-shot-BRDF-shape. `https://github.com/NVlabs/two-shot-brdf-shape`, last accessed on April 1, 2021.

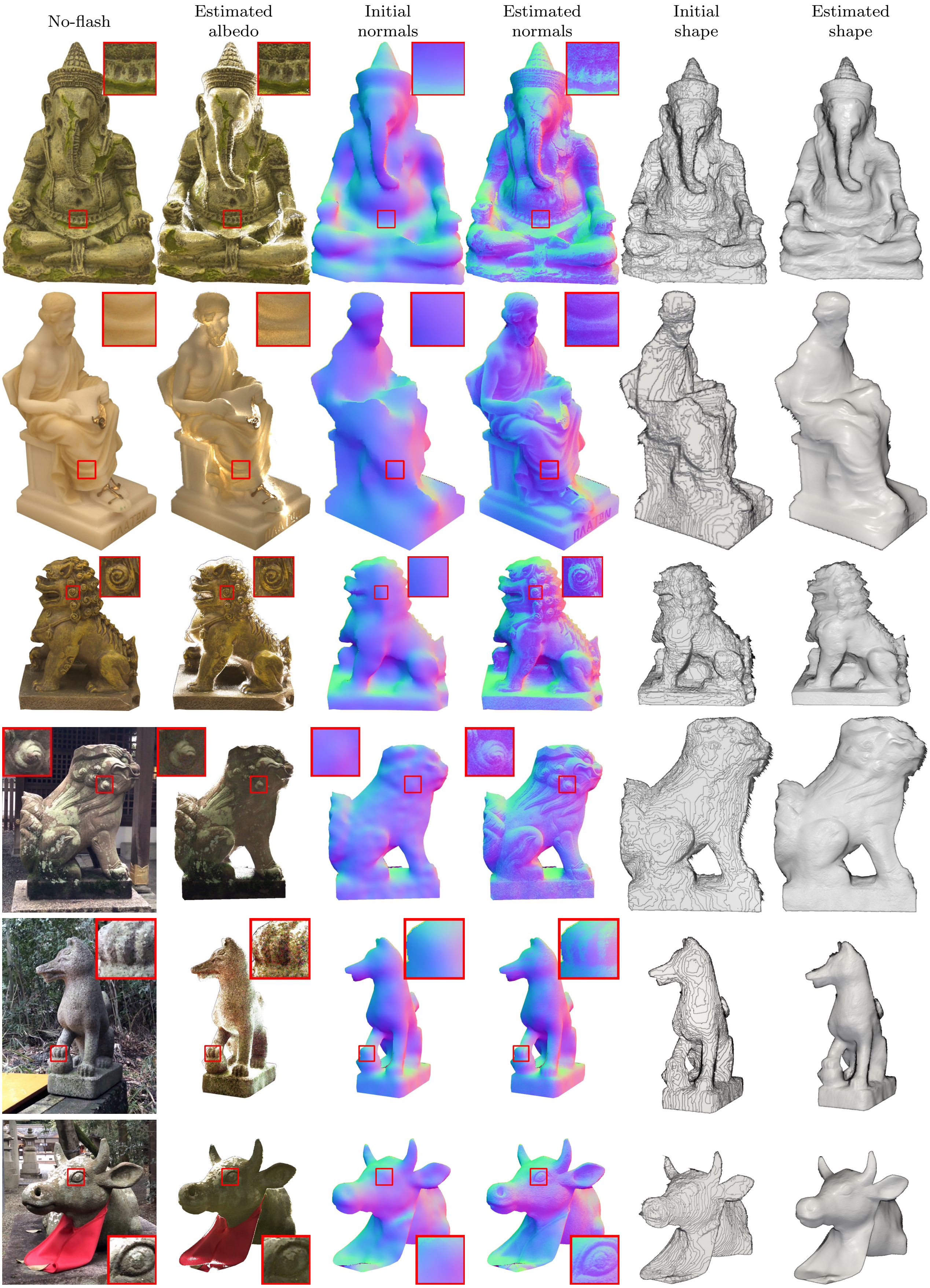

**Fig. 9** Shape and albedo recovery results from an iPhone X; see Fig. 10 for its camera system. The objects in the first three rows are about 10 cm in height and placed in an office room. The last three rows display outdoor stone statues in an old shrine. Our method is able to recover shape details and surface albedo for both indoor and outdoor objects.

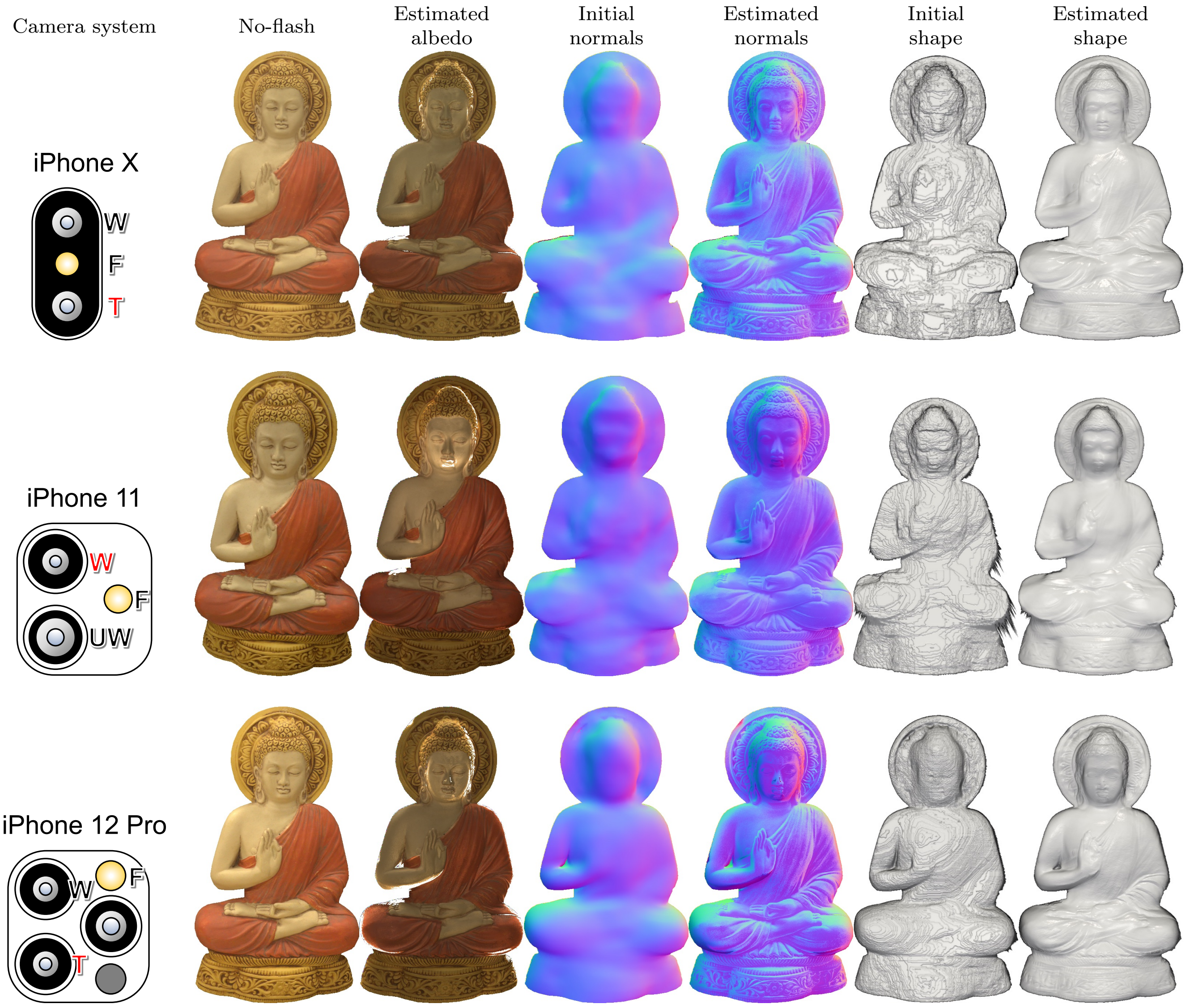

**Fig. 10** Reconstruction of the same object using smartphone models with different camera/flashlight configurations. The first column depicts the camera systems of the iPhone X, 11, and 12 Pro. “UW”,“W”,“T”, and “F” are short for ultra-wide, wide angle, telephoto camera, and flashlight, respectively. The reference camera in the stereo camera is colored red. Our method generates stable results across different camera/flashlight configurations.

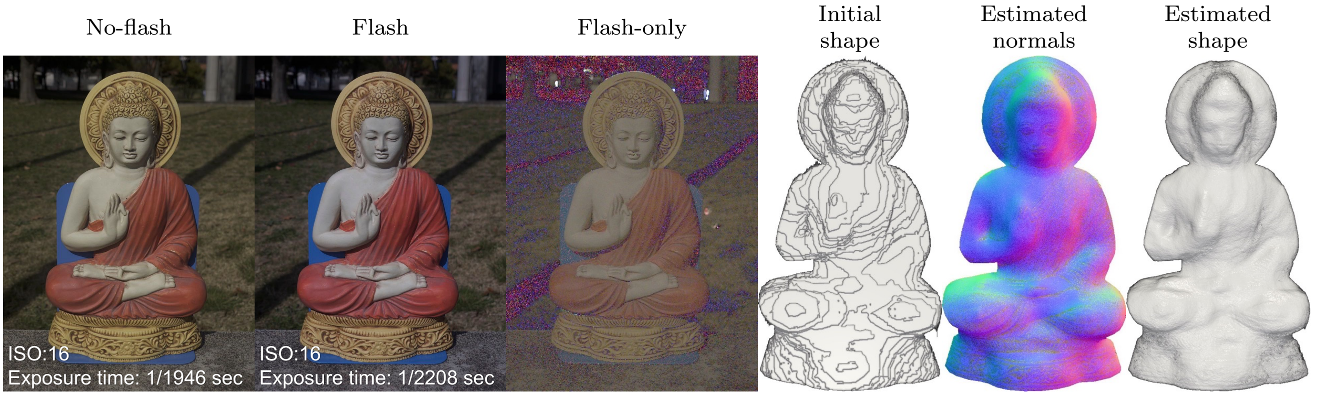

**Fig. 11** Our method breaks down under direct sunlight due to the relatively weak flashlight. The virtual flash-only image (enhanced for visibility) obtained via Eq. (4) hardly provides additional photometric cues, leading to unsatisfactory recovery.

To verify that our method is suitable for different camera and flashlight configurations, we captured images of the same object using an iPhone X, 11, and 12 Pro. From the results in Fig. 10, we can see that our method produces stable results on devices with different camera systems, implying that our method is applicable on a fairly large number of smartphones.

Regarding runtime, each object took about 30 s on a 2.3 GHz Intel i9 CPU. The computational bottlenecks are the fine normal optimization of Eq. (14) and the depth normal fusion of Eq. (23).

## 5 Conclusions

We presented a simple imaging setup for high-fidelity shape and albedo recovery using a stereo camera and flashlight. This setup can be naturally applied to two-shot images from smartphones with a stereo camera, which has become common today. Quantitative evaluation using synthetic images justifies our high-fidelity shape and albedo recovery pipeline. Qualitative results using images captured by a smartphone demonstrate our method's effectiveness in real scenarios. The comparison with related methods shows that our setup is the minimal setup to recover high-fidelity shape and surface albedo via a smartphone.

*Practical implications:* We have verified our method for digitizing cultural heritage artifacts using images captured by off-the-shelf smartphones. This implies that people can immediately turn their smartphones into high-fidelity 3D scanners using our setup and method. We believe that our method is useful in a scenario of crowd-sourced digital archiving, which accelerates the digitization of the world's cultural heritages.

*Limitation:* Our method breaks down if the object is directly lit by strong environmental lighting, such as sunlight; see Fig. 11 for an example. In this scenario, compared with the sunlight the flashlight is too weak to provide additional photometric cues. This problem might be alleviated if smartphones adopt flashlights of stronger intensity in the future. For now, we recommend capturing outdoor objects on cloudy days or around sunrise or sunset. Further, we require the object to be close to the camera due to flashlight falloff in practice.

*Future work:* Our shape and albedo recovery method is based on images shot from a single viewpoint. A practical extension would be to use multi-view images for recovering complete objects.

## 6 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19H01123, JSPS postdoctoral fellowship (JP17F17350), and National Natural Science Foundation of China under Grant Number 62136001 and 61872012.

## References

1. Aittala M, Weyrich T, Lehtinen J (2015) Two-shot SVBRDF capture for stationary materials. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH) 34(4):110–1
2. Aittala M, Aila T, Lehtinen J (2016) Reflectance modeling by neural texture synthesis. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH) 35(4):1–13
3. Basri R, Jacobs DW (2003) Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 25(2):218–233
4. Basri R, Jacobs D, Kemelmacher I (2007) Photometric stereo with general, unknown lighting. International Journal of Computer Vision (IJCV) 72(3):239–257
5. Boss M, Jampani V, Kim K, Lensch H, Kautz J (2020) Two-shot spatially-varying brdf and shape estimation. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
6. Brandt A (1977) Multi-level adaptive solutions to boundary-value problems. Mathematics of computation 31(138):333–390
7. Cao X, Waechter M, Shi B, Gao Y, Zheng B, Matsushita Y (2020) Stereoscopic flash and no-flash photography for shape and albedo recovery. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
8. Cao X, Shi B, Okura F, Matsushita Y (2021) Normal integration via inverse plane fitting with minimum point-to-plane distance. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
9. Choe G, Park J, Tai YW, So Kweon I (2014) Exploiting shading cues in Kinect IR images for geometry refinement. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
10. Cook RL, Torrance KE (1982) A reflectance model for computer graphics. ACM Trans Graphic
11. Deschaintre V, Aittala M, Durand F, Drettakis G, Bousseau A (2018) Single-image SVBRDF capture with a rendering-aware deep network. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH) 37(4):1–15

12. Eisemann E, Durand F (2004) Flash photography enhancement via intrinsic relighting. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH)
13. Feris R, Raskar R, Chen L, Tan KH, Turk M (2005) Discontinuity preserving stereo with small baseline multi-flash illumination. In: Proc. of the International Conference on Computer Vision (ICCV)
14. Gallardo M, Collins T, Bartoli A (2017) Dense non-rigid structure-from-motion and shading with unknown albedos. In: Proc. of the International Conference on Computer Vision (ICCV), pp 3884–3892
15. Häfner B, Quéau Y, Möllenhoff T, Cremers D (2018) Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
16. Häfner B, Peng S, Verma A, Quéau Y, Cremers D (2019) Photometric depth super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 42(10):2453–2464
17. Han Y, Lee JY, So Kweon I (2013) High quality shape from a single RGB-D image under uncalibrated natural illumination. In: Proc. of the International Conference on Computer Vision (ICCV)
18. Haque SM, Chatterjee A, Madhav Govindu V (2014) High quality photometric reconstruction using a depth camera. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
19. He S, Lau RW (2014) Saliency detection with flash and no-flash image pairs. In: Proc. of the European Conference on Computer Vision (ECCV)
20. Hoppe H, DeRose T, Duchamp T, McDonald J, Stuetzle W (1992) Surface reconstruction from unorganized points. In: Proc. of the ACM SIGGRAPH
21. Ikeuchi K (1987) Determining a depth map using a dual photometric stereo. The International journal of robotics research 6(1):15–31
22. Johnson MK, Adelson EH (2011) Shape estimation in natural illumination. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
23. Klowsky R, Kuijper A, Goesele M (2012) Modulation transfer function of patch-based stereo systems. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
24. Li Z, Sunkavalli K, Chandraker M (2018) Materials for masses: SVBRDF acquisition with a single mobile phone image. In: Proc. of the European Conference on Computer Vision (ECCV), pp 72–87
25. Li Z, Xu Z, Ramamoorthi R, Sunkavalli K, Chandraker M (2018) Learning to reconstruct shape and spatially-varying reflectance from a single image. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH) 37(6):1–11
26. Liu DC, Nocedal J (1989) On the limited memory bfgs method for large scale optimization. Mathematical programming 45(1-3):503–528
27. Maier R, Kim K, Cremers D, Kautz J, Nießner M (2017) Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In: Proc. of the International Conference on Computer Vision (ICCV)
28. Maurer D, Ju YC, Breuß M, Bruhn A (2018) Combining shape from shading and stereo: A joint variational method for estimating depth, illumination and albedo. International Journal of Computer Vision (IJCV) 126(12):1342–1366
29. Or-El R, Rosman G, Wetzler A, Kimmel R, Bruckstein AM (2015) RGBD-Fusion: Real-time high precision depth recovery. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
30. Petschnigg G, Szeliski R, Agrawala M, Cohen M, Hoppe H, Toyama K (2004) Digital photography with flash and no-flash image pairs. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH)
31. Quéau Y, Mélou J, Castan F, Cremers D, Durou JD (2017) A variational approach to shape-from-shading under natural illumination. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition
32. Quéau Y, Durou JD, Aujol JF (2018) Normal integration: a survey. Journal of Mathematical Imaging and Vision 60(4):576–593
33. Ramamoorthi R, Hanrahan P (2001) A signal-processing framework for inverse rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp 117–128
34. Sun J, Li Y, Kang SB, Shum HY (2006) Flash matting. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH)
35. Sun J, Kang SB, Xu ZB, Tang X, Shum HY (2007) Flash cut: Foreground extraction with flash and no-flash image pairs. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
36. Wu C, Varanasi K, Liu Y, Seidel HP, Theobalt C (2011) Shading-based dynamic shape refinement from multi-view video under general illumination. In: Proc. of the International Conference on Computer Vision (ICCV)
37. Wu C, Wilburn B, Matsushita Y, Theobalt C (2011) High-quality shape from multi-view stereo and shading under general illumination. In: Proc. of Computer Vision and Pattern Recognition (CVPR)

38. Wu C, Zollhöfer M, Nießner M, Stamminger M, Izadi S, Theobalt C (2014) Real-time shading-based refinement for consumer depth cameras. ACM Transactions on Graphics (Proc of the ACM SIGGRAPH) 33(6):200
39. Yan S, Wu C, Wang L, Xu F, An L, Guo K, Liu Y (2018) DDRNet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In: Proc. of the European Conference on Computer Vision (ECCV)
40. Yu LF, Yeung SK, Tai YW, Lin S (2013) Shading-based shape refinement of RGB-D images. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
41. Zhang Q, Ye M, Yang R, Matsushita Y, Wilburn B, Yu H (2012) Edge-preserving photometric stereo via depth fusion. In: Proc. of Computer Vision and Pattern Recognition (CVPR)
42. Zhou C, Troccoli A, Pulli K (2012) Robust stereo with flash and no-flash image pairs. In: Proc. of Computer Vision and Pattern Recognition (CVPR)