Stereoscopic Flash and No-Flash Photography for Shape and Albedo Recovery

Xu Cao[1] Michael Waechter[1] Boxin Shi[2,3] Ye Gao[4] Bo Zheng[4] Yasuyuki Matsushita[1]
[1]Osaka University [2]Peking University
[3]Peng Cheng Laboratory [4]Huawei Technologies Co., Ltd.

Abstract

We present a minimal imaging setup that harnesses both geometric and photometric approaches for shape and albedo recovery. We adopt a stereo camera and a flashlight to capture a stereo image pair and a flash/no-flash pair. From the stereo image pair, we recover a rough shape that captures low-frequency shape variation without high-frequency details. From the flash/no-flash pair, we derive an image formation model for Lambertian objects under natural lighting, based on which a fine normal map is obtained and fused with the rough shape. Further, we use the flash/no-flash pair for cast shadow detection and albedo canceling, making the shape recovery robust against shadows and albedo variation. We verify the effectiveness of our approach on both synthetic and real-world data.

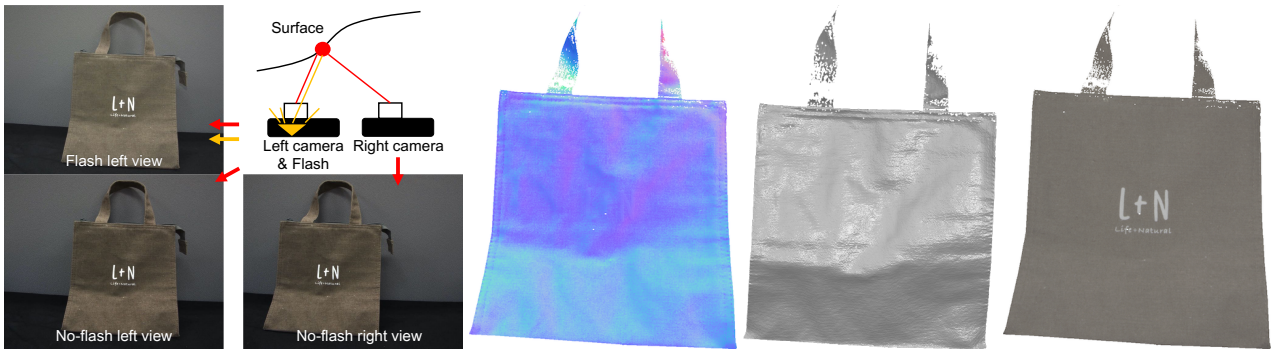
1. Introduction

Shape recovery from images is a fundamental problem in computer vision. Common methods typically fall into one of two classes: geometric or photometric approaches. Geometric approaches take images of a scene from multiple viewpoints, find point correspondences across images and establish their geometric position to recover the shape. They generally do not recover high-frequency shape details, because their patch-based stereo matching has fundamental limitations in spatial resolution [18]. On the other hand, photometric approaches recover per-pixel surface orientation using shading cues. For example, Shape from Shading (SfS) recovers per-pixel surface normal vectors from a single image taken under only one distant light from a single direction [16]; photometric stereo recovers surface normals and albedo using three or more images from the same viewpoint under different lighting conditions [29]. Photometric approaches generally only recover the first-derivative of shape instead of shape itself.

Due to the complementary properties of geometric and photometric approaches, several works have combined both for high-quality shape recovery. Based on multi-view stereo results from a set of images, Wu *et al*. [31] explored the shading cues in those images to recover high-frequency details. Zhang *et al*. [35] combined active stereo and photometric stereo for edge-preserving shape recovery.

The problem of such a combination using geometric and photometric cues, is that the imaging setup is complex: geometric approaches require images from multiple viewpoints taken by multiple cameras or one moving camera. Photometric approaches, on the other hand, require a capture setup that can generate multiple lighting conditions. Zhang *et al*. [35] mounted 3 lights around a Kinect sensor to acquire images under different lighting; Choe *et al*. [6] mounted a bulb on a Kinect and turned on the Kinect's IR projector and the light alternatively during image capture. Several works do not employ a lighting setup and only use the shading cues in the images that are used for the geometric approach. This simplification, however, either puts strict limitations on what scenes can be reconstructed, for example uniform albedo [31, 11], or it requires an additional albedo estimation step [34]. Consequently, both can introduce texture-copying artifacts on the recovered shape [24, 38] where albedo or shading variation is mistaken for shape variation.

In this paper, we present a minimal imaging setup for shape and albedo recovery of Lambertian surfaces, while benefiting from the strengths of both approaches and staying compact. A stereo camera and a flashlight are used to take three images: a left image without flash, a right image without flash and a left image with flash, as shown in Fig. 1(a). The left and right no-flash images constitute a stereo pair providing geometric cues for coarse shape recovery, and the left flash/no-flash image pair provides shading cues for high-frequency detail recovery. The setup is minimal in that, recovering surface normals only from the flash/no-flash pair is an ill-posed problem without the regularization of the coarse shape estimate from the stereo pair; three unknowns (one for albedo and two for a unit surface normal vector) cannot be uniquely solved with two shading constraints for each pixel. Similarly, high-frequency shape details cannot be recovered from the stereo pair [18] without the shading constraints introduced by the flash/no-flash pair. As a result, our setup allows high-fidelity shape and



(a) Imaging setup (b) Normal (c) Shape (d) Albedo

Figure 1: We adopt a stereo camera and a flashlight, take 3 images, and recover high-fidelity normal, shape, and albedo map.[1]

albedo recovery for Lambertian surfaces with non-uniform albedo under natural lighting.

The contributions of our work are as follows:

1. We present a compact imaging setup for shape and albedo recovery that uses both geometric and photometric cues.
2. We derive an image formation model for flash/no-flash image pairs that works for Lambertian surfaces with non-uniform albedo under natural lighting, based on which high-frequency shape details can be recovered.
3. We propose a robust shape recovery framework by taking advantage of a ratio image from the flash/no-flash pair for global lighting estimation and handling of cast shadows.

2. Related Work

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

Shape Recovery Shape recovery by geometric approaches has fundamental limitations in recovering high-frequency details [18]. Contrarily, photometric approaches recover per-pixel surface normals using shading cues in images. Various approaches have been proposed for high-quality shape recovery by combining the strengths of both geometric and photometric approaches.

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*. [4] 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 [3, 17]. Such approaches have been incorporated into geometric approaches. A 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 [26, 31, 33, 34]. 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*. [11] split illumination into a global and a local part, Or-El *et al*. [24] handled local illumination based on first-order spherical harmonics, and Maier *et al*. [21] proposed spatially varying spherical harmonics.

Beside intensity maps captured in the visible spectrum, setups that, capture infrared (IR) images for shading cues have also been explored [6, 12]. To address dynamic scenes, Wu *et al*. [30, 32] proposed approaches that are suitable for refining RGB-D streams. Furthermore, volumetric signed distance functions have been used to encode geometry information [5, 21, 37].

Incorporating photometric cues into multi-view shape recovery is also an active research problem. Gallardo *et al*. [10] used shading cues in non-rigid structure-from-motion, and Maurer *et al*. [22] optimized over shape containing both geometry and photometric constraints.

Flash Photography Different aspects of images taken with a flash have been explored to aid with various computer vision tasks. Due to light intensity falloff, objects close to the flashlight have a stronger change in appearance than distant objects, when comparing a flash and a no-flash image. This has been used in image matting [28], foreground extraction [27], and saliency detection [13]. 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

[1]The brightness of all images in this paper is adjusted for better visualization.

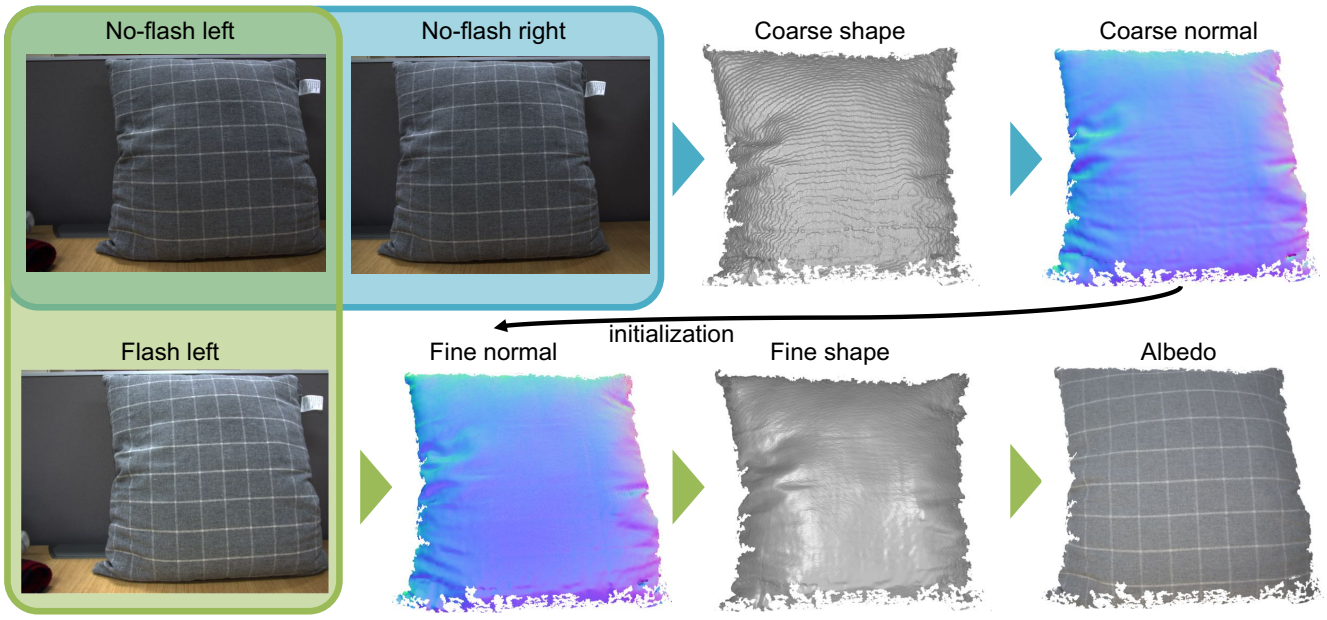


Figure 2: Pipeline of our approach. Coarse shape and the normal map are first acquired from the stereo pair. The flash/no-flash pair is then used to optimize a fine normal map, given the coarse normal map as initialization. Fine shape is obtained by fusing the fine normal map and the coarse shape. An albedo map can also be computed with the fine normal map from our image formation model.

photography enhancement under dark illumination [8], denoising, detail transfer, or white balancing [25].

Further, photometric cues introduced by a flashlight are useful in stereo matching. Feris *et al*. [9] demonstrated that the shadows cast by a flashlight along depth discontinuities help detecting half-occlusion points in stereo matching. Zhou *et al*. [36] showed the ratio of a flash/no-flash pair can make stereo matching robust against depth discontinuities.

In addition, flash images are used in spatially varying BRDF (SVBRDF) recovery. A single image captured from a flash-enabled camera, or a flash/no-flash pair [2] is used for SVBRDF and shape recovery of near-planar objects [1, 7, 19] or those with complex geometry [20].

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

3. Image Formation Model

Figure 2 illustrates the general pipeline of our method for shape and albedo recovery. First we obtain coarse shape from the stereo pair by stereo matching [14], then we use the flash/no-flash pair to optimize the coarse shape to obtain high-quality shape and albedo. The coarse-to-fine shape recovery is possible due to our image formation model derived from the flash/no-flash pair. In this section, we derive our image formation model. In the next section we then describe how we optimize the fine normal map and how we make it robust against cast shadows.

For an image of a surface with Lambertian reflectance, the observed intensity $m \in \mathbb{R}$ at a pixel can be modeled as a shading function $s : \mathcal{S}^2 \to \mathbb{R}$ scaled by an albedo $\rho \in \mathbb{R}$:

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

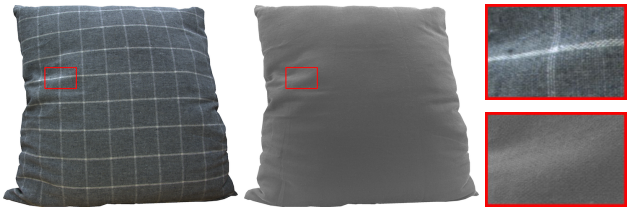
The shading function s is applied to the surface normal $\mathbf{n} \in \mathcal{S}^2 \subset \mathbb{R}^3$ and depends on the environmental lighting.

Now consider a flash/no-flash image pair which both follow Eq. (1). We assume that the flash/no-flash image pair is taken from the same viewpoint, the scene is static, and the environmental lighting stays the same. The pixel at the same location in the flash/no-flash pair then records the intensity of the same surface patch, which we denote as m_{f} and m_{nf}, respectively. The observed intensity difference $m_{\mathrm{f}} - m_{\mathrm{nf}}$ is caused by the additional shading $s_{\mathrm{fo}}(\mathbf{n})$ introduced by the flashlight only, which is also scaled by the albedo ρ. We thus have two equations:

$$\begin{cases} m_{\mathrm{nf}} = \rho s_{\mathrm{nf}}(\mathbf{n}) \\ m_{\mathrm{f}} - m_{\mathrm{nf}} = \rho s_{\mathrm{fo}}(\mathbf{n}), \end{cases} \tag{2}$$

Dividing the two equations yields:

$$\frac{m_{\mathrm{nf}}}{m_{\mathrm{f}} - m_{\mathrm{nf}}} = \frac{s_{\mathrm{nf}}(\mathbf{n})}{s_{\mathrm{fo}}(\mathbf{n})}. \tag{3}$$



(a) No-flash (b) Ratio (c) Closeup

Figure 3: Computing the ratio image according to Eq. (3) cancels the flash/no-flash image pair's albedo variation but the shading variation remains, as shown in the closeup.

Note that the division cancels the albedo out, which means we do not have to put any assumption on ρ, *e.g.*, uniformity. This results in an albedo-free image formation model which describes only the relation between shading and observed intensity. The effect of this albedo canceling is illustrated in Fig. 3. The cushion in Fig. 3(a) has clear stripes with a different albedo than the rest of the cushion, the intensity variation is therefore caused by a joint effect of albedo and shading variation. By taking the ratio of the flash/no-flash pair, our image formation model cancels the albedo variation as shown in Fig. 3(b). In the following, we call Eq. (3) the *ratio image* of a flash/no-flash image pair.

Shading model We now discuss how we model the two shading functions $s_{\text{nf}}(\mathbf{n})$ and $s_{\text{fo}}(\mathbf{n})$. According to Lambert's law, at a surface patch with normal $\mathbf{n}$, the reflected light caused by one ray of light in a direction $\mathbf{l} \in \mathcal{S}^2 \subset \mathbb{R}^3$ with intensity $e : \mathcal{S}^2 \to \mathbb{R}$ is given by:

$$s(\mathbf{n}) = e(\mathbf{l}) \max(\mathbf{n}^\top \mathbf{l}, 0), \tag{4}$$

If light rays reach the surface patch from multiple directions, the reflected light becomes the integral over all possible incident directions:

$$s(\mathbf{n}) = \int_{\mathcal{S}^2} e(\mathbf{l}) \max(\mathbf{n}^\top \mathbf{l}, 0)\, \mathrm{d}\mathbf{l}, \tag{5}$$

As studied by Basri and Jacobs [4], a Lambertian surface acts as a low-pass filter and the shading can be approximated by a low-dimensional model, like spherical harmonics [4] or quadric functions [17]. We use second-order spherical harmonics to parameterize the shading function under no-flash conditions. While spherical harmonics are basis functions defined on $\mathcal{S}^2$, they can be denoted in Cartesian coordinates. Denote $\mathbf{n} = [n_1, n_2, n_3]^\top$, second-order spherical harmonics can be stacked into a vector $\mathbf{h}(\mathbf{n})$ as:

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

The shading under no-flash illumination $s_{\text{nf}}(\mathbf{n})$ 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:

$$s_{\text{nf}}(\mathbf{n}) = \mathbf{h}(\mathbf{n})^\top \mathbf{l}_{\text{nf}}. \tag{6}$$

Note that $\mathbf{l}$ and $\mathbf{l}_{\text{nf}}$ differ; $\mathbf{l}$ is a light ray direction and $\mathbf{l}_{\text{nf}}$ is a stack of linear combination coefficients.

For the flashlight, we assume it is a point light at infinity, its direction is aligned with the camera's principal axis and it points towards the camera, *i.e.* the flashlight direction is $[0, 0, -1]^\top$. Further, we denote the flashlight intensity as e_{f}. As the flashlight is the only light source bringing the shading $s_{\text{fo}}(\mathbf{n})$, Eq. (4) can be applied and it reads:

$$s_{\text{fo}}(\mathbf{n}) = e_{\text{f}} \max([n_1, n_2, n_3][0, 0, -1]^\top, 0) = -e_{\text{f}} n_3. \tag{7}$$

For simplicity, we drop the $\max(\cdot, 0)$ term as the surface normal would in general point towards the camera if the surface patch is visible to the camera. Exceptions occur only when the normal is nearly perpendicular to the ray of light under perspective projection.

Substituting Eqs. (6) and (7) into Eq. (3) yields:

$$\frac{\mathbf{h}(\mathbf{n})^\top \mathbf{l}'}{-n_3} = \frac{m_{\text{nf}}}{m_{\text{f}} - m_{\text{nf}}}, \tag{8}$$

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

4. Shape and Albedo Recovery

In this section, we detail our shape and albedo recovery given a flash/no-flash pair and coarse shape from semi-global stereo matching [14]. For now, we assume the *global lighting vector* $\mathbf{l}'$ in Eq. (8) and the *coarse normal map* is available, and there are no *cast shadows* in the scene. We will describe how to obtain these later in this section.

We formulate the surface normal recovery as per-pixel energy function optimization with a shading constraint, a surface normal constraint, and a unit-length constraint:

$$\min_{\mathbf{n}} E_s(\mathbf{n}) + \lambda_1 E_n(\mathbf{n}) + \lambda_2 E_u(\mathbf{n}), \tag{9}$$

where λ_1 and λ_2 are two weighting factors. The shading constraint minimizes the squared difference between the ratio image and the estimated ratio image in Eq. (8):

$$E_s(\mathbf{n}) = \left(\mathbf{h}(\mathbf{n})^\top \mathbf{l}' + n_3 \frac{m_{\text{nf}}}{m_{\text{f}} - m_{\text{nf}}} \right)^2. \tag{10}$$

Both sides in Eq. (8) are multiplied by n_3 to avoid possible numerical issues.

With the surface normal constraint we enforce the refined surface normal to be close to the coarse surface normal $\mathbf{n}^{(0)}$, *i.e.*, the dot-product should be close to 1:

$$E_n(\mathbf{n}) = (1 - \mathbf{n}^\top \mathbf{n}^{(0)})^2. \tag{11}$$

Finally, we enforce unit length of the surface normal vector:

$$E_u(\mathbf{n}) = (1 - \mathbf{n}^\top \mathbf{n})^2. \tag{12}$$

Due to the non-linearity of the spherical harmonics image formation model, this optimization becomes a non-linear least squares problem which we solve with BFGS.

After the normal map is optimized, we fuse it with the coarse shape following Nehab's approach [23]. The albedo map can be computed, according to Eq. (2), up to a global scale e_f as:

$$\rho = \frac{m_\mathrm{nf}}{\mathbf{h}(\mathbf{n})^\top \mathbf{l}_\mathrm{nf}} = \frac{e_\mathrm{f} m_\mathrm{nf}}{\mathbf{h}(\mathbf{n})^\top \mathbf{l}'}. \tag{13}$$

We now detail how we obtain the coarse normal map, compute the global lighting vector, and handle cast shadows.

Obtaining coarse surface normals We compute the normal map using the PlanePCA method [15]. Given camera intrinsics, all pixels with valid values in a depth map are projected to camera coordinates. For each point, its nearest neighbor points are searched, and the surface normal for the point is found by fitting a plane to its neighbor points. Formally, given a set of points $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$, $\mathbf{p}_i \in \mathbb{R}^3$, the coarse surface normal vector $\mathbf{n}_i^{(0)}$ at $\mathbf{p}_i$ is found by minimizing the following objective:

$$\mathbf{n}_i^{(0)} = \underset{\mathbf{n}_i^{(0)}}{\operatorname{argmin}} \sum_{\mathbf{p}_j \in \mathcal{N}(\mathbf{p}_i)} (\mathbf{p}_j - \bar{\mathbf{p}}_i)^\top \mathbf{n}_i^{(0)}, \tag{14}$$

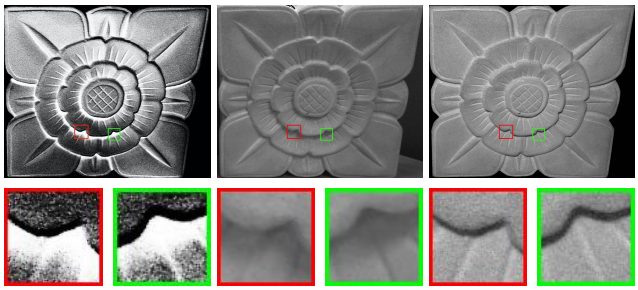
where $\mathcal{N}(\mathbf{p}_i)$ is the set of $\mathbf{p}_i$'s neighborhood points, and $\bar{\mathbf{p}}_i$ is the mean of all $\mathbf{p}_j \in \mathcal{N}(\mathbf{p}_i)$. We measure the nearness of points by Euclidean distance and perform a ball query to search for the neighborhood points of $\mathbf{p}_i$:

$$\mathcal{N}(\mathbf{p}_i) = \{\mathbf{p}_j \mid \left\|\mathbf{p}_j - \mathbf{p}_i\right\|_2 < r,\ \forall \mathbf{p}_j \in \mathbf{P}\}, \tag{15}$$

where r is an empirically chosen ball search radius. Note that $\mathbf{p}_i$ is included in its neighborhood set and different points may have a different number of neighbors. PlanePCA robustly estimates a coarse, smooth normal map, expressing low-frequency shape which is used in the following lighting estimation step.

Computing the global lighting vector Here, our goal is to estimate the low-dimensional global lighting vector $\mathbf{l}'$ in Eq. (8), given two observations and a coarse normal map. Note that solving $\mathbf{l}_\mathrm{nf}$ and e_f separately is unnecessary for shape recovery; unknown e_f barely scales the recovered albedo map. Suppose there are p pixels in the region of interest, for each pixel we stack the row vector $\mathbf{h}(\mathbf{n})^\top/(-n_3)$ vertically into a matrix $\mathbf{N} \in \mathbb{R}^{p\times 9}$ and stack all observed $m_\mathrm{nf}/(m_\mathrm{f} - m_\mathrm{nf})$ into a vector $\mathbf{m} \in \mathbb{R}^p$, yielding an overdetermined system

$$\mathbf{N}\mathbf{l}' = \mathbf{m}. \tag{16}$$



(a) Ratio img. (b) No-flash img. (c) Flash-only img.

Figure 4: Relation between ratio image and cast shadows. Bright and dark ratio image regions correspond to cast shadows caused by the environmental lighting and the flashlight, respectively.

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

Handling cast shadows The image formation model based on spherical harmonics described by Eq. (4) handles attached shadows but is unable to model cast shadows [4]. Our image formation model thus breaks down in regions dominated by cast shadows and the optimization in Eq. (9) would produce artifacts. To tackle this we introduce a confidence term ω into the energy function's shading constraint:

$$\min_{\mathbf{n}} \omega E_s(\mathbf{n}) + \lambda_1 E_n(\mathbf{n}) + \lambda_2 E_u(\mathbf{n}), \tag{17}$$

where ω denotes the confidence of the image formation model at a pixel. We define it as:

$$\omega = \exp\Big(-\frac{(r-\mu)^2}{2\sigma^2}\Big), \tag{18}$$

where r is the ratio between the flash and the no-flash intensity, and μ and σ are the mean and the standard deviation of the ratio in the region of interest. This definition is based on the observation that cast shadows strongly increase the ratio r and make it deviate far from the mean ratio.

This phenomenon is shown in Fig. 4. The no-flash image Fig. 4(a), exhibits shadows from environmental light which correspond to bright regions (*i.e.*, high ratios) in the ratio image Fig. 4(b). Contrarily, the flash-only image Fig. 4(c), *i.e.*, flash minus no-flash image, exhibits shadows from the flashlight only, which correspond to dark regions (*i.e.*, low ratios) in the ratio image. Shadows from the flashlight in the flash-only image only occur if the flashlight's light direction and the camera's principal axis are not perfectly aligned.

The above two observations lead to the design of ω in Eq. (18). For pixels where the ratio deviates too much

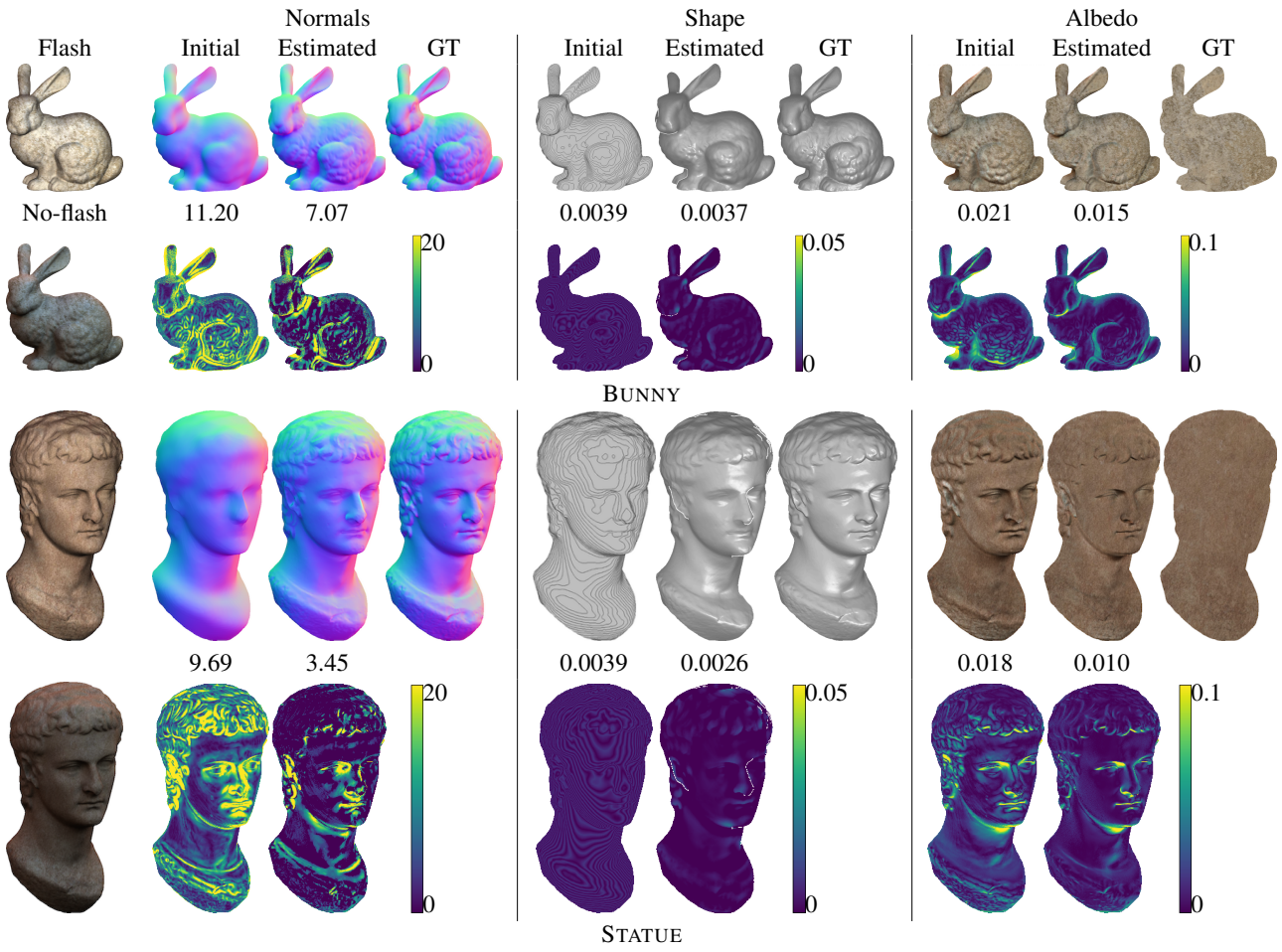


Figure 5: Shape and albedo recovery results on synthetic data. The first column shows the rendered flash/no-flash pair. The even rows display the error map. The numbers above error maps are MAngE of normal map, and MAbsE of shape as well as albedo map. Our method recovers high-frequency shape details.

from the mean ratio, our image formation model cannot reliably explain the observations and we trust the coarse surface normal from the stereo pair more.

5. Experiments

In this section we show experiments on both synthetic and real-world data to evaluate our method.

5.1. Synthetic Data

Experimental setup To quantitatively evaluate the shape and albedo recovery accuracy, we rendered synthetic data with the physically based renderer Mitsuba.[2] Two publicly available 3D objects, Stanford BUNNY and a STATUE,[3] are rendered under orthographic projection. Regarding the no-flash image, we put the objects under three different environment maps: PISA, DOGE and GLACIER.[4] We then simulated the flashlight by putting an additional directional light source in the same scene. We assumed there to be no inter-reflection. In addition, we obtained the scenes' ground truth shape, depth maps, and normal maps. To simulate the coarse shape and normal map, we quantized the ground truth depth map with 128 levels and applied PlanePCA with a ball query radius of 0.07 to obtain coarse normal maps. The ground truth albedo was set as a texture image, and the initial albedo was obtained according to Eq. (13) using the coarse normal map. Note that this initial albedo map only serves as a comparison.

Results Figure 5 shows shape and albedo recovery results along with their coarse initializations and ground truth, under the lighting condition PISA. We use mean angular error

[2] Mitsuba Renderer

[3] "The Getty Caligula" by CosmoWenman / CC BY 4.0

[4] High-Resolution Light Probe Image Gallery

Table 1: Qualitative comparison of recovered shape between different methods. We report the MAbsE of two objects under three lighting condition. w/ conf. means using Eq. (17) for optimization.

Env. map	Method	BUNNY	STATUE
PISA	Han *et al*. [11]	3.56e-3	3.59e-3
	Yan *et al*. [33]	4.02	1.26
	Ours	3.43e-3	**2.42e-3**
	Ours (w/ conf.)	**3.39e-3**	2.48e-3
DOGE	Han *et al*. [11]	3.66e-3	3.68e-3
	Yan *et al*. [33]	4.02	1.26
	Ours	3.54e-3	3.09e-3
	Ours (w/ conf.)	**3.44e-3**	**2.98e-3**
GLACIER	Han *et al*. [11]	3.65e-3	3.64e-3
	Yan *et al*. [33]	4.02	1.26
	Ours	3.45e-3	3.69e-3
	Ours (w/ conf.)	**3.41e-3**	**3.64e-3**

(MAngE) and mean absolute error (MAbsE), respectively, to evaluate normal maps and shape as well as albedo maps. While the coarse normal map contains only low-frequency content, our optimization based on our albedo-free image formation model recovers high-frequency shape details and exhibits a lower error than the initializations. The recovered fine-grained details are also reflected in the shape after the depth-normal fusion. Further, compared with the coarse albedo, errors in the final albedo are decreased thanks to the recovered high-fidelity shape. As the optimization over the normal map is non-convex, error the optimized normals can get stuck in non-global minima when the initial normal deviates too much from the ground truth.

We compare our method with Han *et al*.'s [11] and Yan *et al*.'s [33] and report MAbsE of two objects in Table 1. The albedo maps are uniform for all objects for a fair comparison, as our method uses the ratio of a flash/no-flash pair to eliminate the effect of albedo variations while the other two methods assume a uniform albedo. Our method performs best among all comparison methods. Further, our method's results are generally improved using the confidence term in the energy function, which verifies the effectiveness of our strategy for handling cast shadows.

5.2. Real-world Data

In this section, we show qualitative results of real-world data from our imaging setup.

Experimental setup Figure 6 shows our imaging setup. We used two FLIR machine vision cameras, which have linear radiometric response and a 12-bit ADC, accompanied with two $8.5\,\mathrm{mm}$-lens. The baseline between the two cameras was roughly $70\,\mathrm{mm}$ and the focal length was 2500 pixels. As flashlight we used the flash of a smartphone mounted above the left camera.

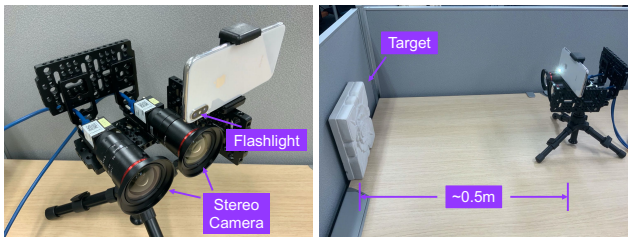


Figure 6: Stereo imaging setup and capture environment.

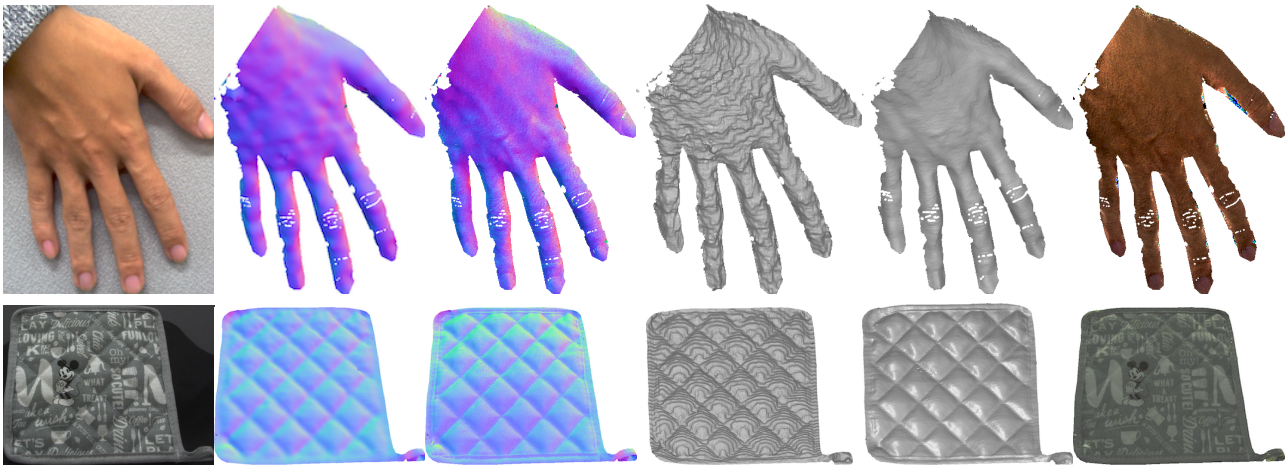
We used the PySpin SDK to control both cameras. When capturing flash/no-flash pairs, we kept the exposure time identical and turned off all non-linear image processing like gamma correction to ensure linear images. Each image was taken 5 times and averaged to suppress imaging noise.

We obtained coarse depth maps with OpenCV's semi-global matching [14], applied median filtering to remove outliers, and closed small holes by solving a Laplacian equation with Dirichlet boundary conditions. Missing values close to depth boundaries were not considered. We then project the depth map to camera coordinate, and applied PlanePCA with a ball search radius of $5\,\mathrm{mm}$ to obtain coarse normal maps. The weighting factors λ_1 and λ_2 in Eq. (9) were both set to 0.1 for all objects.

Results Figures 1, 2, and 7 show qualitative results of shape and albedo recovery on real data. While the objects have a complex albedo, our approach successfully recovered high-quality details in the normal maps based on our image formation model with one single lighting. Benefiting from the high-quality normal map, we recovered the high-frequency details that were missing in the coarse shape. Further, the albedo map looks reasonable despite the surface's shading variation.

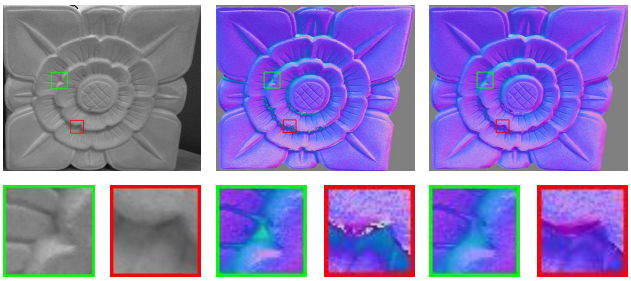
Figure 8 shows a comparison of surface normal recovery results with and without the confidence ω of Eq. (17). While all image formation models based on spherical harmonics would fail at regions dominated by cast shadows, our confidence term ω largely reduces artifacts in the recovered normal map.

Figure 9 shows a qualitative comparison of shape recovery results between Han *et al*. [11], Yan *et al*. [33], and our approach. For a fair comparison, all three started with the same initial normal map and shape as our approach. From Fig. 9 we can see that our approach successfully recover high-frequency details with fewer outliers. Our usage of a flash/no-flash pair avoids texture-copy artifacts, which exists in Yan's method. As the two methods adopt a single no-flash image for shape recovery, the comparison verifies



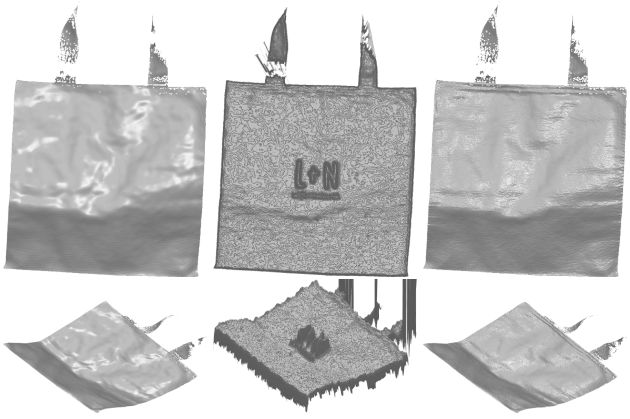
(a) (b) (c) (d) (e) (f)

Figure 7: Shape and albedo recovery on real-world data. From left to right: (a) left camera's no-flash image, (b) coarse normal map obtained by applying PlanePCA on the coarse shape, (c) fine normal map obtained through our surface normal optimization, (d) coarse shape from stereo matching, (e) fine shape fused from fine surface normal and coarse shape, and (f) recovered albedo map.



(a) No-flash (b) w/o confidence (c) w/ confidence

Figure 8: Comparison of normal map recovery without and with the confidence term ω in Eq. (17). The use of ω reduces artifacts in the surface normal vectors, as shown in (c).



Han *et al*. [11] Yan *et al*. [33] Ours

Figure 9: Shape recovery comparison on real-world data among Han *et al*. [11], Yan *et al*. [33], and our approach. The second row shows side views.

the effectiveness of our use of flash/no-flash pairs.

6. Conclusion

We presented a simple imaging setup for effective shape and albedo recovery by a stereo camera and a flash. We demonstrated that this setup can benefit from the strengths of both geometric and photometric approaches while sticking to the minimal setup. Fine shape details and the albedo map can be recovered based on our image formation model derived from the flash/no-flash pair. Comparison experiments with methods using single no-flash image verified the effectiveness of our usage of flash/no-flash pairs. One limitation of our approach is that albedo recovery may fail in shadowed region. Future works include improving the albedo recovery and adapting the setup for dynamic scene capture and recovery.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant No. JP19H01123, JSPS postdoctoral fellowship (No. JP17F17350), National Natural Science Foundation of China Grant No. 61872012, National Key R&D Program of China (2019YFF0302902), and Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Miika Aittala, Timo Aila, and Jaakko Lehtinen. Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 35(4):1–13, 2016. 3

[2] Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. Two-shot SVBRDF capture for stationary materials. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 34(4):110–1, 2015. 3

[3] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision (IJCV)*, 72(3):239–257, 2007. 2

[4] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(2):218–233, 2003. 2, 4, 5

[5] Erik Bylow, Robert Maier, Fredrik Kahl, and Carl Olsson. Combining depth fusion and photometric stereo for fine-detailed 3D models. In *Proc. of Scandinavian Conference on Image Analysis*, 2019. 2

[6] Gyeongmin Choe, Jaesik Park, Yu-Wing Tai, and In So Kweon. Exploiting shading cues in Kinect IR images for geometry refinement. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[7] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image SVBRDF capture with a rendering-aware deep network. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 37(4):1–15, 2018. 3

[8] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2004. 3

[9] Rogerio Feris, Ramesh Raskar, Longbin Chen, Kar-Han Tan, and Matthew Turk. Discontinuity preserving stereo with small baseline multi-flash illumination. In *Proc. of International Conference on Computer Vision (ICCV)*, 2005. 3

[10] Mathias Gallardo, Toby Collins, and Adrien Bartoli. Dense non-rigid structure-from-motion and shading with unknown albedos. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 3884–3892, 2017. 2

[11] Yudeog Han, Joon-Young Lee, and In So Kweon. High quality shape from a single RGB-D image under uncalibrated natural illumination. In *Proc. of International Conference on Computer Vision (ICCV)*, 2013. 1, 2, 7, 8

[12] Sk. Mohammadul Haque, Avishek Chatterjee, and Venu Madhav Govindu. High quality photometric reconstruction using a depth camera. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[13] Shengfeng He and Rynson W.H. Lau. Saliency detection with flash and no-flash image pairs. In *Proc. of European Conference on Computer Vision (ECCV)*, 2014. 2

[14] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, 2007. 3, 4, 7

[15] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. Surface reconstruction from unorganized points. In *Proc. of ACM SIGGRAPH*, 1992. 5

[16] Berthold K.P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, Massachusetts Institute of Technology, 1970. 1

[17] Micah K. Johnson and Edward H. Adelson. Shape estimation in natural illumination. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2011. 2, 4

[18] Ronny Klowsky, Arjan Kuijper, and Michael Goesele. Modulation transfer function of patch-based stereo systems. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2

[19] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 72–87, 2018. 3

[20] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 37(6):1–11, 2018. 3

[21] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proc. of International Conference on Computer Vision (ICCV)*, 2017. 2

[22] Daniel Maurer, Yong Chul Ju, Michael Breuß, and Andrés Bruhn. 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, 2018. 2

[23] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2005. 5

[24] Roy Or-El, Guy Rosman, Aaron Wetzler, Ron Kimmel, and Alfred M. Bruckstein. RGBD-Fusion: Real-time high precision depth recovery. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[25] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2004. 3

[26] Yvain Quéau, Jean Mélou, Fabien Castan, Daniel Cremers, and Jean-Denis Durou. A variational approach to shape-from-shading under natural illumination. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2017. 2

[27] Jian Sun, Sing Bing Kang, Zong-Ben Xu, Xiaoou Tang, and Heung-Yeung Shum. Flash cut: Foreground extraction with flash and no-flash image pairs. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[28] Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. Flash matting. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2006. 2

[29] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980. 1

[30] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. of International Conference on Computer Vision (ICCV)*, 2011. 2

[31] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2011. 1, 2

[32] Chenglei Wu, Michael Zollhöfer, Matthias Nießner, Marc Stamminger, Shahram Izadi, and Christian Theobalt. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 33(6):200, 2014. 2

[33] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu. DDRNet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *Proc. of European Conference on Computer Vision (ECCV)*, 2018. 2, 7, 8

[34] Lap-Fai Yu, Sai-Kit Yeung, Yu-Wing Tai, and Stephen Lin. Shading-based shape refinement of RGB-D images. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2013. 1, 2

[35] Qing Zhang, Mao Ye, Ruigang Yang, Yasuyuki Matsushita, Bennett Wilburn, and Huimin Yu. Edge-preserving photometric stereo via depth fusion. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[36] Changyin Zhou, Alejandro Troccoli, and Kari Pulli. Robust stereo with flash and no-flash image pairs. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[37] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 34(4):96, 2015. 2

[38] Xinxin Zuo, Sen Wang, Jiangbin Zheng, and Ruigang Yang. Detailed surface geometry and albedo recovery from RGB-D video under natural illumination. In *Proc. of International Conference on Computer Vision (ICCV)*, 2017. 1