# Photometric Stereo using Internet Images

Boxin Shi[1,2]* Kenji Inose[3] Yasuyuki Matsushita[4] Ping Tan[5] Sai-Kit Yeung[1] Katsushi Ikeuchi[3]

[1]Singapore University of Technology and Design [2]MIT Media Lab
[3]The University of Tokyo [4]Microsoft Research Asia [5]Simon Fraser University

## Abstract

*Photometric stereo using unorganized Internet images is very challenging, because the input images are captured under unknown general illuminations, with uncontrolled cameras. We propose to solve this difficult problem by a simple yet effective approach that makes use of a coarse shape prior. The shape prior is obtained from multi-view stereo and will be useful in twofold: resolving the shape-light ambiguity in uncalibrated photometric stereo and guiding the estimated normals to produce the high quality 3D surface. By assuming the surface albedo is not highly contrasted, we also propose a novel linear approximation of the nonlinear camera responses with our normal estimation algorithm. We evaluate our method using synthetic data and demonstrate the surface improvement on real data over multi-view stereo results.*

## 1. Introduction

Shape recovery is a fundamental problem in computer vision. Over the past decade, both the capturing devices and 3D reconstruction algorithms have been improved drastically that brings surface reconstruction from small scale desktop objects to large scale outdoor sculptures. Given multiple images of the same large scale object, for which Internet is an important image resource, recent progress in structure from motion (SfM) and multi-view stereo (MVS) allow reconstruction even up to city scale. There are existing works that recover sparse 3D points [6] and depth map [25] for large scale objects using Internet images. These works focus more on acquiring the rough depth using geometric constraints rather than high quality surface.

Photometric stereo, on the other hand, can recover highly detailed surface geometry at pixel-level accuracy in the form of surface normal map, by using scene radiances observed under varying lightings [21]. Recently, photometric stereo in an outdoor setting is possible by using a mirror sphere to calibrate the natural illumination [24]. However, for Internet images the natural illumination is completely unknown.

When the lighting conditions are unknown, the problem becomes *uncalibrated* photometric stereo, whose solution can only be derived up to some ambiguity, such as the generalized-bas-relief (GBR) ambiguity [3] for unknown directional lightings, or a high-dimensional linear ambiguity [2] for unknown general lightings. Besides the unknown illumination, uncontrolled sensor is another difficult issue since automatic gain control and nonlinear radiometric response deteriorate the resultant shape. Therefore in most of the previous photometric stereo approaches, sensor gains and responses are either pre-calibrated or assumed to be known; however, sensor parameters are usually inaccessible in Internet images.

In this paper, we focus on Internet images of large outdoor sculptures which are difficult to capture by Lidar or flying drone due to their size or security reason. We propose a unified approach by using the *shape prior* – coarse shape information obtained by SfM and MVS – to show its important roles in various steps throughout the whole pipeline: from preparing the organized input images for photometric stereo, resolving the ambiguity in uncalibrated photometric stereo to guiding the final surface reconstruction. We also show that the effect of nonlinear sensor responses can be approximated by a high-dimensional linear transformation applied over the illumination component except for highly contrasted albedos. This can be viewed as *pseudo multiplexing* of natural lightings which allows highly accurate shape estimation without the influence of nonlinear responses of sensors. The key contribution of this paper is to extend photometric stereo method to work with a wild setting of *unknown general illumination* and *uncontrolled sensors* on *unorganized Internet images*.

## 2. Related Work

The Internet images contain comprehensive contents for the same place of interest from various viewpoints and il-

*Part of this work was done while the first author was an intern at Microsoft Research Asia and a Ph.D. candidate at the University of Tokyo.

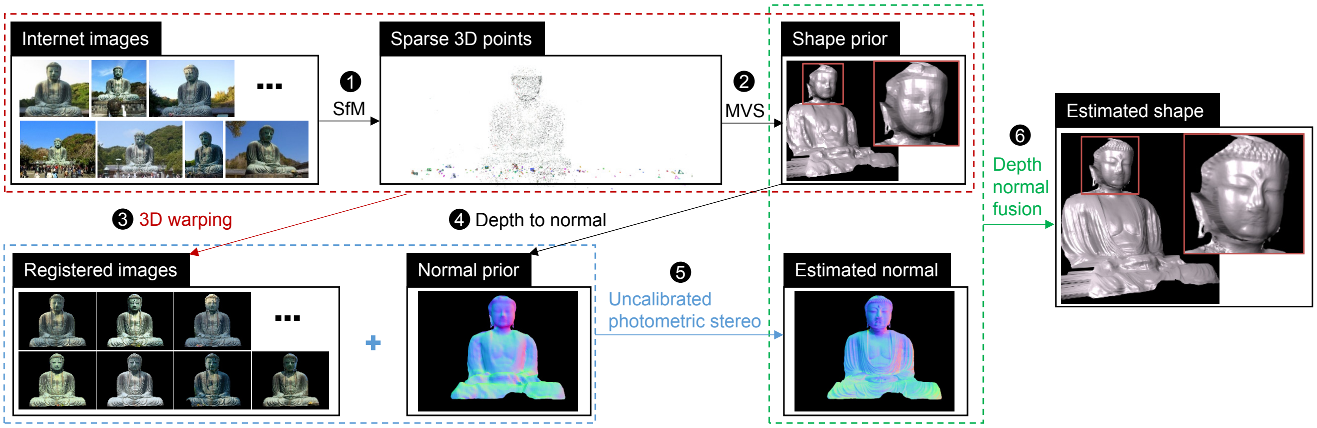


Figure 1. Pipeline of our method, which contains 6 main steps. We take unorganized Internet images as input to first generate a shape (normal) prior, which is then used to produce high quality surface reconstruction with great details.

luminations. Considering the geometric constraint from multi-view images, recent progress in SfM and MVS show that sparse yet reliable 3D points can be recovered from Internet images [20, 6, 5]. In more recent works, Shan *et al*. [17] propose a large scale system combining Internet photos from many resources (ground-level, aerial images, and street view) for realistic reconstruction, and Zheng *et al*. [25] propose a multi-view depth estimation method with consideration of pixel-level view selection. By integrating the photometric cues, Shen and Tan [18] obtain sparse normals that are useful for weather estimation. Ackermann *et al*. [1] apply MVS using Internet images to compute sparse surface normals and transfer them to images under varying lightings for estimating dense normal. By using Internet face photos as a shape prior and combining shading constraints from photometric stereo, the 3D face models can be reconstructed in the wild [12].

The Internet images are highly unorganized and captured in an uncontrolled setup. To apply photometric stereo on Internet images, the uncalibrated lighting needs to be estimated. For uncalibrated photometric stereo, it is well known that there exists a $3 \times 3$ linear ambiguity [8] in the recovered surface normals for general surfaces, and a three-parameter GBR ambiguity for integrable surfaces [3]. Recent works mainly focus on estimating the three unknowns to obtain final normal estimates (*e.g.*, [19]). Under a general unknown lighting, there is a $9 \times 3$ ($= 27$ unknowns) linear ambiguity in surface normals under illuminations modeled by second order spherical harmonics. Unfortunately, this high-dimensional ambiguity cannot be completely removed without additional information [2]. We propose to resolve this ambiguity by using the shape prior.

Another challenging issue for Internet images is that the sensors are uncontrollable and their parameters are inaccessible, so methods that require controlled exposure time (*e.g.*, [7]) are unsuitable to calibrate the radiometric response for linearizing Internet images. Self-calibration to radiometric response can be applied for directional lighting [19] or directional plus ambient lighting [4], but for natural lighting and uncontrolled sensors it is still an open problem. Instead of explicitly estimating the response function, we disregard it in a self-contained pipeline.

## 3. Proposed Method: Overview

The main challenges of solving photometric stereo using Internet images are uncontrolled illuminations and sensors. The first problem can be formulated as an uncalibrated photometric stereo with general unknown lightings, and the latter one is to deal with unknown exposures and radiometric responses. We tackle these two problems by taking advantages of a coarse shape prior.

The complete pipeline of our method is shown in Fig. 1. We collect Internet images of an outdoor sculpture and apply SfM [20] (Step 1) and PMVS [5] (Step 2) to obtain sparse point clouds. Then a Poisson reconstruction method [11] is used for creating a water-tight coarse depth prior. Based on the depth prior we align multiview images to the reference view via 3D warping (Step 3) to prepare the input for photometric stereo. A mask is added manually to exclude the sky. Note that in Fig. 1, the Internet images are unorganized pictures from multiple viewpoints, but the registered images contain the object from exactly the same viewpoint and varying natural illumination.

From the shape prior which is a rough depth map, we first convert it to a rough normal map that gives us the normal prior (Depth to normal, Step 4). The normal prior and registered images are then combined to solve the uncalibrated photometric stereo (Step 5) problem. Finally, the estimated normal and shape prior are integrated to produce the final 3D surface (Depth Normal Fusion, Step 6) which has more details due to the accurate normal information.

In the next section, we will explain the details of solving the uncalibrated photometric stereo problem with the shape prior, *i.e.*, Step 4-6, especially Step 5; after that, we will explain how the shape prior helps to relieve the uncontrolled camera issue in Sec. 5.

# 4. Normal Estimation from Internet Images

## 4.1. Depth to normal

Since our method works in the surface normal domain, we first convert the coarse depth prior into a surface normal prior for solving our problem (Step 4). A naïve computation of derivatives over a coarse depth map results in a noisy normal map; therefore, we use a plane principal component analysis method introduced in [13] for robustly computing the surface normal prior. Given the depth map and camera intrinsics of the reference view, the method first projects the depth map to 3D points in the world coordinate system. For each 3D point, the method groups a set of points within a short distance $d$. For the $i$-th group that contains $q_i$ 3D points, by stacking them in a matrix $\mathbf{Q} \in \mathbb{R}^{q_i \times 3}$, the surface normal is computed as

$$\tilde{\mathbf{n}} = \underset{\mathbf{n}}{\operatorname{argmin}} \|(\mathbf{Q} - \bar{\mathbf{Q}})\mathbf{n}\|_{\mathrm{F}}, \tag{1}$$

where $\bar{\mathbf{Q}} \in \mathbb{R}^{q_i \times 3}$ is a matrix containing the centroid of $\mathbf{Q}$ in all the rows. A larger $d$ produces a smoother normal estimate when the input depth contains more noise. The calculated normals are then projected back to the image plane of reference view. This step produces the normal prior $\tilde{\mathbf{N}} \in \mathbb{R}^{p \times 3}$ where $p$ is the number of foreground pixels in a registered image.

## 4.2. Uncalibrated photometric stereo

### 4.2.1 Image formation model

We begin Step 5 with a Lambertian image formation model under natural lightings. Given a scene point with Lambertian albedo $\rho$ and surface normal $\mathbf{n} = [n_x, n_y, n_z]^\top$, its radiance $r$ can be written as:

$$r = \int_\Omega \rho L(\omega) \max((\mathbf{n}^\top \omega), 0) \mathrm{d}\omega, \tag{2}$$

where $\omega \in \mathbb{R}^{3 \times 1}$ is a unit vector of spherical directions $\Omega$, and $L(\omega)$ is the light intensity from the direction $\omega$. This integration can be approximated using spherical harmonics as

$$r = \mathbf{s}^\top \mathbf{l}, \tag{3}$$

where $\mathbf{s} = [s_1, s_2, \ldots, s_k]^\top \in \mathbb{R}^{k \times 1}$ are harmonics images of surface normal $\mathbf{n}$ and albedo $\rho$, and $k$ is the number of elements determined by the order of spherical harmonics. The vector $\mathbf{l} \in \mathbb{R}^{k \times 1}$ is the $k$-dimensional lighting coefficients.

Given $p$ pixels observed under $q$ different illuminations, we store all these $p \times q$ radiance values into a radiance matrix $\mathbf{R} \in \mathbb{R}^{p \times q}$. By a row-wise stacking of $p$ transposed harmonics images $\mathbf{s}^\top$ in a shape matrix $\mathbf{S} \in \mathbb{R}^{p \times k}$ and a column-wise stacking of $q$ lighting coefficients $\mathbf{l}$ in a lighting matrix $\mathbf{L} \in \mathbb{R}^{k \times q}$, Eq. (3) can be written in a matrix form as:

$$\mathbf{R} = \mathbf{S}\mathbf{L}. \tag{4}$$

We further include the effect of sensor gains and responses in the image formation model. Under varying illumination, the exposure time for each image is likely different for an uncontrolled sensor. Each exposure time corresponds to a scaling of its lighting coefficient $\mathbf{l}$, which is one column in the lighting matrix $\mathbf{L}$. For simplicity of notations, we still use $\mathbf{L}$ to represent the scaled lighting coefficient matrix. In addition, a nonlinear response function transforms the radiance $\mathbf{R}$. Let us denote the camera's radiometric response as $f$. For now, we assume the response function $f$ is the same for all images. The registered images are vectorized and stacked together in a column-wise manner to form the observation matrix $\mathbf{I} \in \mathbb{R}^{p \times q}$. $\mathbf{I}$ can be expressed using the response function $f$, which is applied in an element-wise manner using an operator $\circ$, as:

$$\mathbf{I} = \mathbf{R} \circ f = (\mathbf{S}\mathbf{L}) \circ f. \tag{5}$$

Our method approximates the nonlinear response function $f$ using a high-dimensional linear transformation $\mathbf{F} \in \mathbb{R}^{q \times q}$ as

$$\mathbf{I} = (\mathbf{S}\mathbf{L}) \circ f \approx \mathbf{S}\mathbf{L}\mathbf{F}. \tag{6}$$

The transformation $\mathbf{F}$ varies with the response function $f$ and radiance $\mathbf{R}$. We will explain and verify the appropriateness of this approximation in Sec. 5. Since our goal is to estimate the shape component $\mathbf{S}$, we rewrite Eq. (6) as $\mathbf{I} = \mathbf{S}\mathbf{L}_F$ by $\mathbf{L}_F = \mathbf{L}\mathbf{F}$ so that the illumination component embeds the transformation caused by response functions.

### 4.2.2 Normal estimation algorithm

Similar to previous approaches [8, 2], we perform the singular value decomposition (SVD) on the observation matrix $\mathbf{I}$ to estimate the shape matrix $\mathbf{S}$ up to a linear ambiguity $\mathbf{B} \in \mathbb{R}^{k \times k}$. In other words, the ambiguous $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}_F$ are related to their ground truths $\mathbf{S}$ and $\mathbf{L}_F$ by $\tilde{\mathbf{S}}\mathbf{B} = \mathbf{S}$ and $\mathbf{B}^{-1}\tilde{\mathbf{L}}_F = \mathbf{L}_F$, respectively. As discussed in [2], the surface normal is encoded in the second to fourth columns of $\tilde{\mathbf{S}}$. Therefore, a $k \times 3$ matrix $\mathbf{A}$ is sufficient for computing normal from $\tilde{\mathbf{S}}$ as $\tilde{\mathbf{S}}\mathbf{A}$.

Given the coarse normal prior $\tilde{\mathbf{N}}$ from Step 4, we can estimate $\mathbf{A}$ to remove the ambiguity:

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmin}} \|\tilde{\mathbf{S}}\mathbf{A} - \tilde{\mathbf{N}}\|_{\mathrm{F}}, \tag{7}$$

**Algorithm 1** Normal estimation with shape prior

```
Decompose observation matrix $\mathbf{I}$ as $\mathbf{I} = \tilde{\mathbf{S}}\tilde{\mathbf{L}}_F$;
Solve the linear equations for $\hat{\mathbf{A}}$ using Eq. (7);
Nonlinear refinement to obtain $\mathbf{A}^*$ using Eq. (8);
Compute normal by $\mathbf{N}^* = O(\tilde{\mathbf{S}}\mathbf{A}^*)$.
```

By applying $\hat{\mathbf{A}}$ to the original $\tilde{\mathbf{S}}$, we obtain disambiguated normals $\hat{\mathbf{N}}$ by $\hat{\mathbf{N}} = O(\tilde{\mathbf{S}}\hat{\mathbf{A}})$, where $O$ is a row-wise normalization operator forcing each row of the matrix to be a unit vector. In practice, we apply Gaussian smoothing to both $\tilde{\mathbf{N}}$ and the ambiguous shape matrix $\tilde{\mathbf{S}}$ before solving for $\hat{\mathbf{A}}$.

Solving Eq. (7), however, can only provide a correct solution if the object's albedo is uniform. When a scene contains variant albedos, the norm of rows of $\tilde{\mathbf{S}}\hat{\mathbf{A}}$ varies, while $\tilde{\mathbf{N}}$ only contains unit normal vectors. To explicitly handle the albedo variations, we further optimize $\mathbf{A}$ using the following objective function:

$$\mathbf{A}^* = \underset{\mathbf{A}}{\operatorname{argmin}} \|O(\tilde{\mathbf{S}}\mathbf{A}) - \tilde{\mathbf{N}}\|_{\mathrm{F}}. \tag{8}$$

The above optimization problem is highly nonlinear, but we can use the linear solution $\hat{\mathbf{A}}$ as an initial guess to solve for $\mathbf{A}$. The optimization is solved using a Matlab build-in function "fminsearch". While the global optimum is not guaranteed, in our experiments this nonlinear refinement works well because of the good initialization. The final surface normal is computed by $\mathbf{N}^* = O(\tilde{\mathbf{S}}\mathbf{A}^*)$.

The complete normal estimation method (Step 5) is summarized in Algorithm 1.

### 4.3. Depth normal fusion

The shape prior is beneficial not only for surface normal estimation, but also for surface reconstruction (Step 6) by serving as anchor points for the surface recovery from the normal map [16, 10, 9]. To estimate the optimal depth $\mathbf{Z}^* \in \mathbb{R}^{p\times 1}$ by combining the estimated surface normal $\mathbf{N}^*$ (Step 5) and a vectorized noisy depth map $\mathbf{Z} \in \mathbb{R}^{p\times 1}$ (Step 1 and 2), we can form a linear system of equations as [16] to reconstruct the surface:

$$\begin{bmatrix} \lambda \mathbf{I}_d \\ \nabla^2 \end{bmatrix} [\mathbf{Z}^*] = \begin{bmatrix} \lambda \mathbf{Z} \\ \partial \mathbf{N}^* \end{bmatrix}, \tag{9}$$

where $\nabla^2$ is a Laplacian operator, $\mathbf{I}_d$ is an identity matrix and $\lambda$ is a weighting parameter controlling the contribution of depth constraint. $\partial\mathbf{N}^*$ is the stacks of $-\frac{\partial}{\partial x}\frac{n_x}{n_z} - \frac{\partial}{\partial y}\frac{n_y}{n_z}$ for each normal $\mathbf{n} \in \mathbf{N}^*$. While it forms a large linear system of equations, because the left matrix is sparse, it can be efficiently solved using existing sparse linear solvers (*e.g.*, QR decomposition based solvers), or multigrid techniques.

## 5. Linear Approximation of Sensor Responses

A useful byproduct of our pipeline is that by using the shape prior, our method naturally ignores the nonlinear response functions embedded in Internet images. This is another challenging issue of using Internet images for photometric stereo, due to that for each input image the nonlinear response function is unknown and arbitrary. Further, the uncontrolled cameras used for recording Internet images bring difficulty in performing radiometric calibration using conventional methods [15]. We address this challenge using a practical approximation by aligning the shape estimates with the shape prior and encoding the unknown responses to a linearly multiplexed lighting component.

### 5.1. Intuition

The shape estimation method in Sec. 4 relies on a high-dimensional linear approximation of nonlinear responses (Eq. (6)), which allows us to separate the effects of unknown sensor gains and responses from the shape estimation as $\mathbf{I} = \mathbf{S}(\mathbf{LF})$. The resulting lighting component $\mathbf{L}_F(= \mathbf{LF})$ becomes different from the actual $\mathbf{L}$. However, it is a linear combination of the original lightings, and it can be viewed as *pseudo multiplexing* of natural lightings, which allows us to effectively account for unknown sensor responses. Intuitively speaking, for each image, the pseudo multiplexing can be explained as such a process: A uniform surface is illuminated by a natural illumination $\mathbf{l}$ and captured via a nonlinear response function $f$ that maps $\mathbf{r} = \mathbf{Sl}$ to $\mathbf{i} = \mathbf{r} \circ f$; the observed image is approximately equal to that of the same surface illuminated by $\mathbf{l}_F$ (one column of $\mathbf{L}_F$) and captured with a linear camera, *i.e.*, $\mathbf{i} = \mathbf{Sl}_F$.

Qualitatively, the linear approximation becomes less accurate when a surface contains more diverse albedos. For example, consider two surface points that have the same normal $\mathbf{n}$ but different albedos $\rho_1$ and $\rho_2$ ($\rho_1 \neq \rho_2$). With a little bit abuse of notations, we use $\mathbf{s} = \rho\mathbf{n}$ for simplicity [1]. The radiance at these two points are $r_1 = \mathbf{s}_1^\top \mathbf{l} = \rho_1 \mathbf{n}^\top \mathbf{l}$ and $r_2 = \mathbf{s}_2^\top \mathbf{l} = \rho_2 \mathbf{n}^\top \mathbf{l}$, respectively. A camera response function maps these radiance values to $f(r_1)$ and $f(r_2)$. Since $f$ is a nonlinear function, generally, the ratio $f(\rho_1 \mathbf{n}^\top \mathbf{l}) : f(\rho_2 \mathbf{n}^\top \mathbf{l})$ becomes different from $\rho_1 : \rho_2$. However, the linear approximation is limited in representing this nonlinear effect, and this error becomes more obvious as the difference between $\rho_1$ and $\rho_2$ becomes larger.

For our method, it is not necessary to estimate the multiplexing matrix $\mathbf{F}$; however, the approximation power of the linear transformation $\mathbf{F}$ is of interest because it is related to the shape estimation accuracy. We therefore assess the appropriateness of the approximation using the

[1] Strictly speaking, $s_1 = \rho, s_2 = \rho n_x, s_3 = \rho n_y, s_4 = \rho n_z, s_5 = \rho(3n_z^2 - 1), s_6 = \rho n_x^2, s_7 = \rho n_x n_z, s_8 = \rho n_y n_z, s_9 = \rho(n_x^2 - n_y^2)$ for a second order spherical harmonics representation.

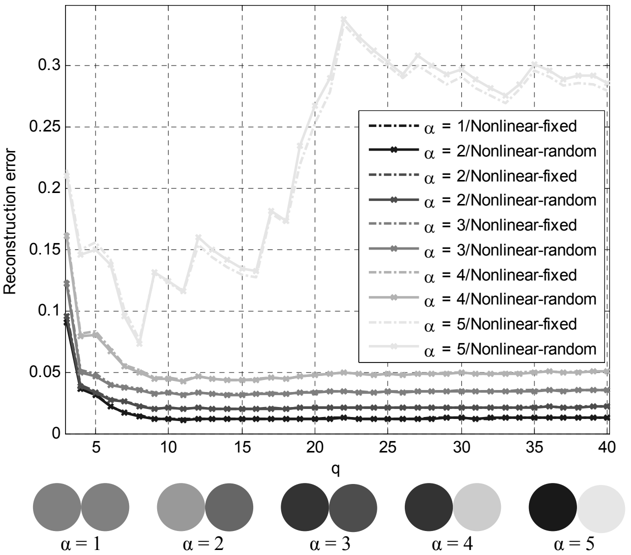


Figure 2. Reconstruction errors of Eq. (6) w.r.t. varying numbers of images ($q$) for scenes containing two spheres with different albedos. $\alpha = \{1, 2, 3, 4, 5\}$ indicate that left/right spheres have albedo values of $\{0.5/0.5, 0.4/0.6, 0.3/0.7, 0.2/0.8, 0.1/0.9\}$.

database of measured response functions [7]. Fortunately, as we will see below, the approximation error is consistently and sufficiently small for the real world response functions even for variant albedos, due to the high regularity of real response functions and good approximation capability of high-dimensional linear multiplexing.

### 5.2. Verification

We use synthetic images [2] to assess the approximation ability. We simulate the imaging process where $f$ is applied to $\mathbf{R}$ in two different manners: 1) "Nonlinear-fixed": the same response is applied to all images under varying illuminations. This case corresponds to a scenario with an uncontrolled camera. And, 2) "Nonlinear-random": each image under one lighting condition is distorted by a randomly selected response function in the database. This case more closely mimics Internet images, where each image is recorded via a distinct unknown and nonlinear response. We average the results over all 201 response functions in "Nonlinear-fixed" case, and 201 random trials are performed and averaged for "Nonlinear-random" case. The test scene consists of two spheres with different albedos, whose values are shown at bottom of Fig. 2.

To assess the approximation ability, we evaluate the reconstruction error of Eq. (6). Given $\mathbf{R}$ and $\mathbf{I}$, we solve for $\mathbf{F}$ by linear least squares as $\hat{\mathbf{F}} = \mathbf{R}^{+}\mathbf{I}$, where $\mathbf{R}^{+}$ is the pseudo-inverse of $\mathbf{R}$. Then, a reconstruction of $\mathbf{R}_F$ is computed as $\mathbf{R}_F = \mathbf{R}\hat{\mathbf{F}}$. The reconstruction error is evaluated as the mean of $|i - r_F|/i$, where $i$ is a pixel observation of $\mathbf{I}$ and $r_F$ is the corresponding element in $\mathbf{R}_F$. This is a relative error (percentage) defined for each observation. We show the reconstruction errors with respect to the varying numbers of input images $q$ and albedo contrast in Fig. 2. The errors are pretty low (about $1\%$) when the number of input images $q$ becomes $q \geq 9$ for the case of uniform albedo ($\alpha = 1$). On the other hand, as the albedo contrast becomes greater, the errors increase accordingly. Except for the extreme case ($\alpha = 5$) [3], the reconstruction errors are consistently low (below $5\%$). Therefore, the approximation generally works well, except for scenes that exhibit significantly high contrast. The high correlation of different radiometric response functions makes our method works for both "Nonlinear-fixed" and "Nonlinear-random" cases (their reconstruction errors are always similar), *i.e.*, our approximation is valid for Internet images.

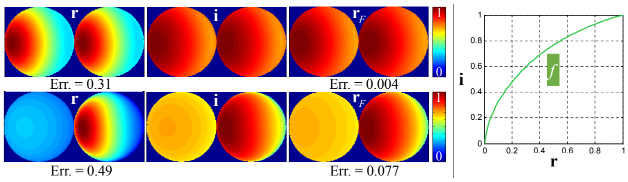


Figure 3. Input radiance $\mathbf{r}$ is transformed by a nonlinear response $f$ (shown on rightmost) to $\mathbf{i}$ as $\mathbf{i} = \mathbf{r} \circ f$. Top row: two spheres with the same albedo ($\alpha = 1$); bottom row: two spheres with greatly different albedo ($\alpha = 5$). Our linearly approximated $\mathbf{r}_F$ shows very close appearance to $\mathbf{i}$. The errors below show mean of relative differences of $\mathbf{r}$ and $\mathbf{r}_F$ from $\mathbf{i}$, respectively. Color encoded images are used for better visualization.

As an intuitive example, we show linearly approximated images of uniform abledo ($\alpha = 1$, top row) and strong contrast case ($\alpha = 5$, bottom row) in Fig. 3. Note that $\mathbf{i}$ and $\mathbf{r}_F$ have very small difference visually, especially for uniform albedo case, which shows the validity of our linear approximation.

## 6. Experiments

### 6.1. Quantitative evaluation

We use a synthetic scene, CAESAR, to quantitatively evaluate our method. The data is synthesized in the same way as the simulation test did in Sec. 5.2. We fix the number of distinct lightings $q = 40$ for this test. We evaluate how the varying albedo contrast and coarseness of shape prior influence the normal estimates given nonlinear images, by applying various real-world response functions [7] to input images. We found that our method works well for most albedo variations except for the highly contrasted one, and even a severely contaminated shape prior is quite useful in estimating accurate normal. We provide the complete anal-

[2] We use a 9D spherical harmonics expansion of both normal and lighting to create these images.

[3] $\alpha = 5$ mimics a scene of almost black and white spheres.

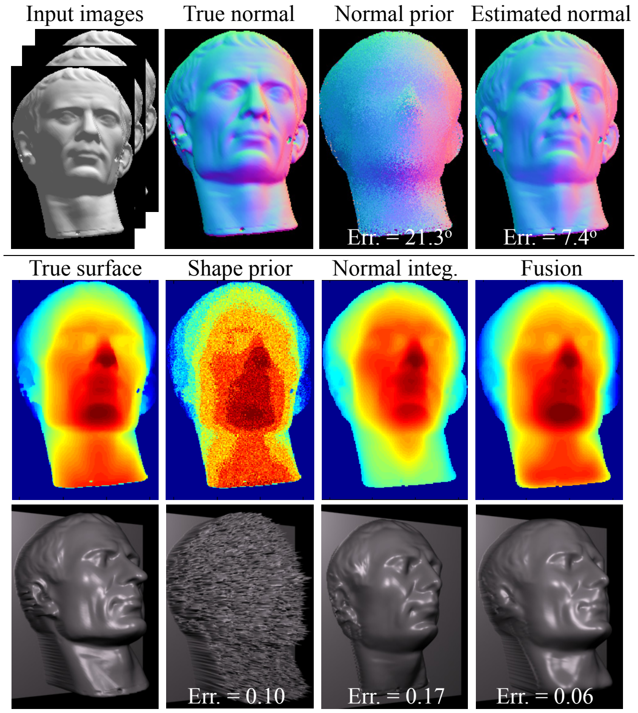


Figure 4. Surface normal estimation (top row) and surface reconstruction result (depth and surface in two bottom rows) using synthetic data. The numbers on normal maps show the angular errors and the numbers on surface show the reconstruction errors.

ysis in the supplementary material.

Here, we show a typical example in Fig. 4. In this example, a severely contaminated depth map, quantized to $4$ bits with zero-mean Gaussian of standard deviation $0.04$ being added, is used as shape prior. The normal estimation accuracy is evaluated using angular difference (degree) w.r.t. the true normal. With a noisy normal prior of angular error $21.3°$, we obtain a normal map with much smaller errors, which is about $14°$ smaller than the prior. The surface reconstruction error is defined as the mean value of $|z_0 - z^*|/z_0$ across pixels, where $z_0$ is the true depth and $z^*$ is the depth estimate. The original rough depth (Shape prior) is noisy, but it still provides useful positional information. On the other hand, direct integrating a surface from the normals results in a distorted reconstruction with a larger bias (Normal integ.). By fusing the normal and depth information, a more accurate surface can be reconstructed (Fusion), as pointed out by previous work [16].

### 6.2. Result using Internet images

In addition to the KAMAKURA BUDDHA data shown in Fig. 1, we show results of three more scenes named MOTHERLAND CALLS, STATUE OF LIBERTY, and MOUNT RUSHMORE in Fig. 5. These four datasets contain 200, 109, 320, and 128 downloaded images respectively, which are used for SfM and MVS. We use $163$, $109$, $63$, and $42$ images in each dataset which roughly have the same viewpoints for 3D warping and normal estimation. We model the natural illumination using the third order spherical harmonics ($k = 16$) for all the experiments. We compare our method to [1] by using the same input for reference. Our normal estimates show more meaningful shape information than the results from [1], because of the capability of handling natural lightings and variations of camera responses. For example, in the pedestal of STATUE OF LIBERTY, our result shows consistent normals for plane structures and clearer details of the bricks. We also show the reconstructed surfaces by fusing our estimated normal and the shape prior from MVS (as baseline for surface reconstruction comparison) in Fig. 6, where more details can be observed thanks to the refined surface normals by photometric stereo.

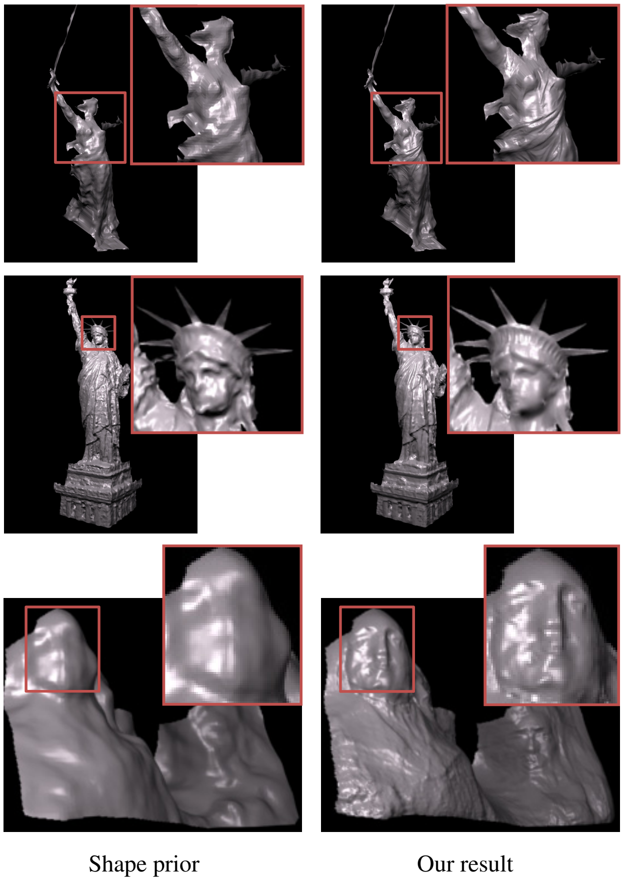


Figure 6. 3D reconstruction results using Internet images. Close-up views are indicated by red rectangles.

When the Internet images have almost the same viewpoints, SfM and MVS will produce degenerated results. But if a scene contains (partly) regular shapes, we can directly use this knowledge as the shape prior. We show such an example of TAJ MAHAL where the shape of the dome has a comprehensive structure. We manually assign a hemisphere

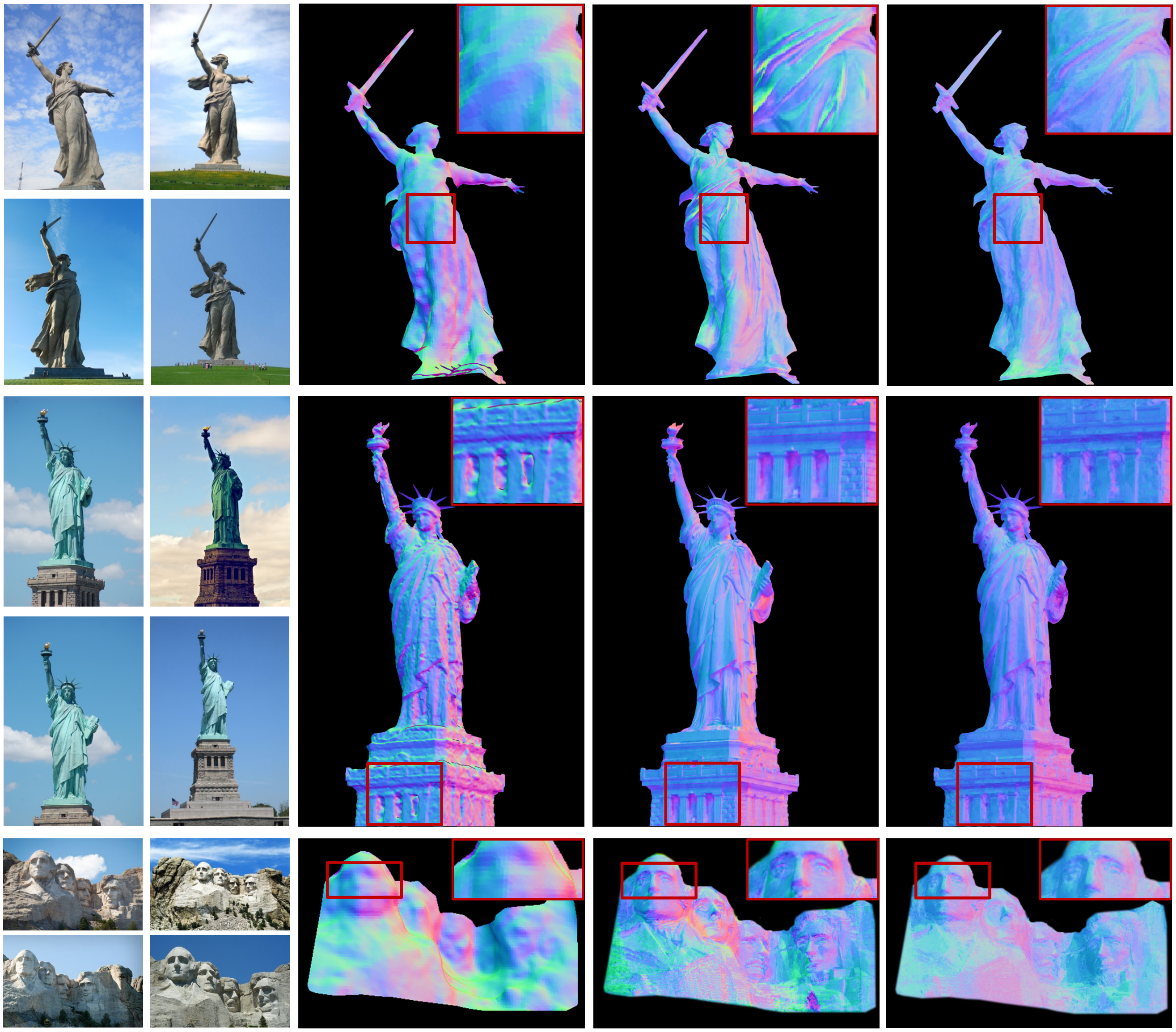


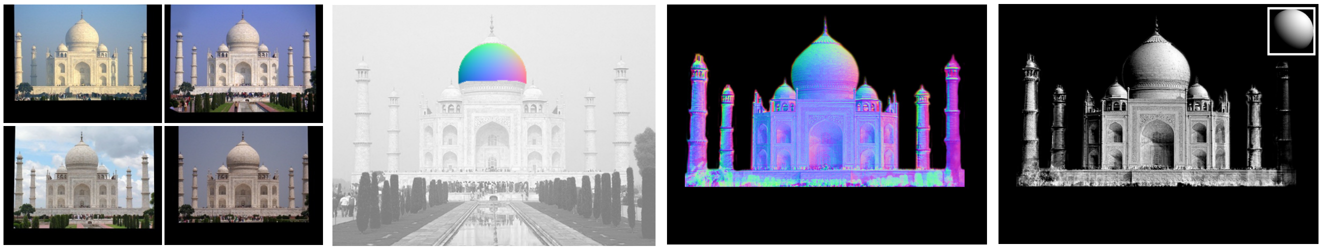
Figure 5. Surface normal estimation results using Internet images. Four representative images from the input dataset are shown in the left column; the image on top left is the reference image to which other images are registered. Close-up views of estimated normal maps are indicated by red rectangles.

surface normal map to the dome part as the shape prior. The Internet images of TAJ MAHAL are registered to a reference view via homography using SIFT [14] features in this case, as the 3D shape information is unavailable. The result using 66 images is shown in Fig. 7. The rendered Lambertian shading using the estimated normal under a distant lighting $[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]^\top$ is shown on the rightmost for verification.

## 7. Conclusion

We present a photometric stereo method that works with unorganized Internet images, captured under general unknown illumination, with uncontrolled sensors. We suggest using shape priors from SfM and MVS to fully remove the ambiguity in uncalibrated illumination setting, to guide the normal to surface integration, and to avoid the effect of uncontrolled sensor. The proposed method shows high-quality 3D modeling over existing MVS method.

**Limitations** In our current solution, cast shadows are not handled. Due to the shape-light ambiguity, it is difficult to explicitly calculate the visibility map like [22]. We have investigated some robust algorithm [23] to handle cast shad-



Registered images Normal prior Estimated normal Relighting result

Figure 7. Surface normal estimation result using Internet images and the known shape as a prior.

ows as outliers by forcing the input matrix to be rank-$k$. However, the result showed almost no improvement in our context, because the ideal rank-$k$ matrix is seldom observed for Internet images. Properly modeling cast shadows in our pipeline is left as our future work. Combining our method with recent works of MVS reconstruction using Internet images, which consider large scale data [17] and view selection [25], is also an interesting direction.

## Acknowledgement

Boxin Shi is supported by SUTD-MIT joint postdoctoral fellowship. Sai-Kit Yeung is supported by SUTD Start-Up Grant ISTD 2011 016, SUTD-MIT International Design Center Grant IDG31300106, and Singapore MOE Academic Research Fund MOE2013-T2-1-159.

## References

[1] J. Ackermann, M. Ritz, A. Stork, and M. Goesele. Removing the example from example-based photometric stereo. In *Proc. ECCV Workshop RMLE*, 2010.

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

[3] P. Belhumeur, D. Kriegman, and A. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999.

[4] M. Diaz and P. Sturm. Radiometric calibration using photo collections. In *Proc. ICCP*, 2011.

[5] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[6] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. In *Proc. ICCV*, 2007.

[7] M. Grossberg and S. Nayar. Modeling the space of camera response functions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(10):1272–1282, 2004.

[8] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of the Optical Society of America*, 11(11):3079–3089, 1994.

[9] T. Higo, Y. Matsushita, N. Joshi, and K. Ikeuchi. A hand-held photometric stereo camera for 3-D modeling. In *Proc. ICCV*, 2009.

[10] N. Joshi and D. Kriegman. Shape from varying illumination and viewpoint. In *Proc. ICCV*, 2007.

[11] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proc. Eurographics Symposium on Geometry processing (SGP)*, pages 61–70, 2006.

[12] I. Kemelmacher-Shlizerman and S. Seitz. Face reconstruction in the wild. In *Proc. ICCV*, 2011.

[13] K. Klasing, D. Althoff, D.Wollherr, and M. Buss. Comparison of surface normal estimation methods for range sensing applications. In *Proc. ICRA*, 2011.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] T. Mitsunaga and S. Nayar. Radiometric self calibration. In *Proc. CVPR*, 1999.

[16] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. In *Proc. SIGGRAPH (ACM Trans. on Graphics)*, pages 536–543, 2005.

[17] Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. Seitz. The visual turing test for scene reconstruction. In *Proc. 3DV*, 2013.

[18] L. Shen and P. Tan. Photometric stereo and weather estimation using Internet images. In *Proc. CVPR*, 2009.

[19] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *Proc. CVPR*, 2010.

[20] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *Proc. SIGGRAPH (ACM Trans. on Graphics)*, pages 835–846, 2006.

[21] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[22] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proc. CVPR*, 2011.

[23] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Proc. ACCV*, 2010.

[24] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin. Depth-assisted shape-from-shading. In *Proc. CVPR*, 2013.

[25] E. Zheng, V. Jojic, E. Dunn, and J.-M. Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proc. CVPR*, 2014.

# Photometric Stereo using Internet Images – Supplementary Material

Boxin Shi[1,2] Kenji Inose[3] Yasuyuki Matsushita[4] Ping Tan[5] Sai-Kit Yeung[1] Katsushi Ikeuchi[3]

[1]Singapore University of Technology and Design [2]MIT Media Lab
[3]The University of Tokyo [4]Microsoft Research Asia [5]Simon Fraser University

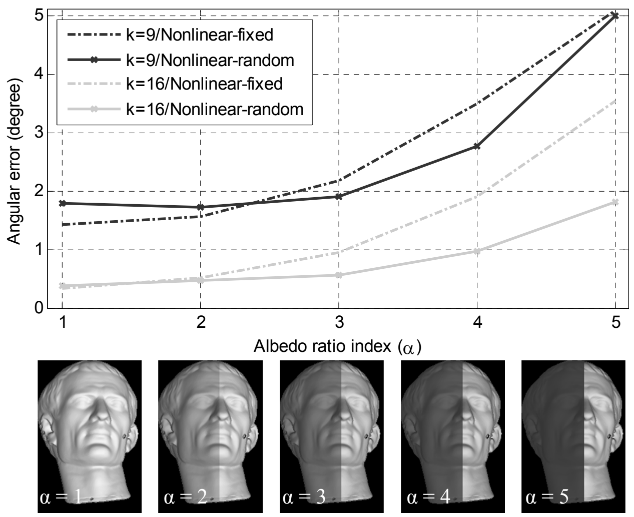


Figure 1. Normal estimation accuracy (angular error in degrees) w.r.t. different albedo contrasts. $\alpha = \{1, 2, 3, 4, 5\}$ indicate that left/right half of the object have albedo values of $\{0.5/0.5, 0.4/0.6, 0.3/0.7, 0.2/0.8, 0.1/0.9\}$. Two different dimensions of lighting coefficients $k$ are evaluated.

This document shows the complete quantitative evaluation in Sec. 6.1.

**Effect of albedo contrast** We evaluate the effect of albedo contrast to normal estimation accuracy. To exclude other factors except for nonlinear sensor responses, in this we use the ground truth normal as $\tilde{\mathbf{N}}$ to remove the ambiguity. Figure 1 shows the normal estimation accuracy with respect to varying albedo contrast $\alpha$ for different dimensions of lighting coefficients $k$. As we have observed in Sec. 5, the accuracy is affected by the greater albedo contrast in general, and the errors become smaller with a larger $k$. This indicates that the higher-order lightings make the pseudo multiplexing more effective. The "Nonlinear-fixed" cases show larger errors than "Nonlinear-random" cases due to large errors caused by some response functions in unusual shapes that are difficult to approximate.

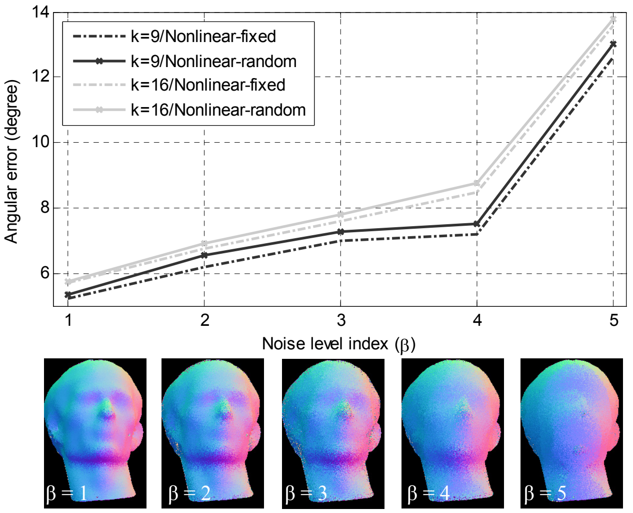


Figure 2. Normal estimation accuracy (angular error in degrees) w.r.t. varying noise levels in shape priors. $\beta = \{1, 2, 3, 4, 5\}$ are labels to represent corruptions where the clean depth maps are quantized to $\{8, 6, 5, 4, 3\}$ bits, with zero-mean Gaussian noise of standard deviations $\{0.02, 0.03, 0.04, 0.04, 0.05\}$ added. Two different dimensions of lighting coefficients $k$ are evaluated.

**Effect of noise in shape priors** Figure 2 shows the variation of normal estimation errors with different noise levels in shape priors. The input depth values are quantized to 3 bits in the worst case, and Gaussian noise with standard deviations up to 0.05 is further added in order to simulate the real-world shape priors. The computed surface normal priors have errors from about 13° to 25°. In this test, the albedo is set as uniform to remove the effect from albedo variations. The errors increase with the roughness of the shape priors. Except for the extreme case ($\beta = 5$), the normal estimation accuracy is consistently high even with nonlinear responses. Under severe noise, a large $k$ lowers the normal estimation accuracy, because it allows too much freedom in the ambiguity matrix, which makes the solution sensitive to noise. In practice, $k$ should be adjusted according to the tradeoff between nonlinearity of responses (prefers a larger $k$) and the coarseness of shape prior (prefers a smaller $k$).