

## Shading-based Shape Refinement of RGB-D Images

Lap-Fai Yu<sup>1\*</sup> Sai-Kit Yeung<sup>2</sup>

<sup>1</sup> University of California, Los Angeles

<sup>3</sup> Korea Advanced Institute of Science and Technology

Yu-Wing Tai<sup>3</sup> Stephen Lin<sup>4</sup>

<sup>2</sup> Singapore University of Technology and Design

<sup>4</sup> Microsoft Research Asia

### Abstract

We present a shading-based shape refinement algorithm which uses a noisy, incomplete depth map from Kinect to help resolve ambiguities in shape-from-shading. In our framework, the partial depth information is used to overcome bas-relief ambiguity in normals estimation, as well as to assist in recovering relative albedos, which are needed to reliably estimate the lighting environment and to separate shading from albedo. This refinement of surface normals using a noisy depth map leads to high-quality 3D surfaces. The effectiveness of our algorithm is demonstrated through several challenging real-world examples.

### 1. Introduction

Shape-from-shading (SfS) is a challenging problem because of the considerable ambiguity in its solution. For the simplest case of Lambertian reflectance and known albedo, the derived solution suffers from bas-relief ambiguity [6, 22, 4]. When albedo is unknown, the range of possible solutions expands significantly. To resolve these ambiguities, an obvious solution is to utilize a set of input images under different lighting conditions, which transforms the SfS problem into that of photometric stereo [18, 21]. However, such additional input data is often inconvenient to obtain in practise. Recent techniques for SfS [8, 12] estimate shape from a single input image under natural illumination, but deal with uniform-albedo objects and require a special calibration target to measure lighting.

In this paper, we propose a shading-based shape refinement algorithm that utilizes Microsoft Kinect to address the ambiguities that exist among lighting, normals and albedo. The Kinect records each RGB image together with a depth map. Although the depth map is noisy and typically contains holes<sup>1</sup>, we present a method that effectively

\*This work was done while Lap-Fai Yu was a visiting student at MSRA and at SUTD.

<sup>1</sup>Depth map holes result from scene areas in a Kinect depth image outside the depth sensing range or occluded from the infrared light projections, since the infrared projection and sensing directions are not the same.

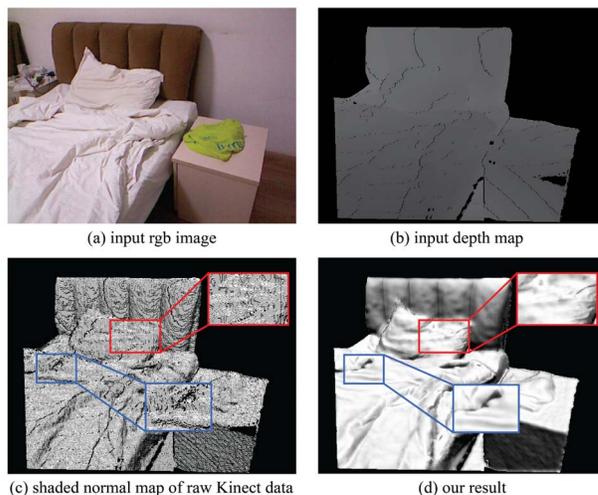


Figure 1. Our shading-based shape refinement deals with the shape and reflectance ambiguities of SfS while effectively enhancing surface normals computed from the raw, noisy depth data of Kinect.

tively utilizes this information to improve the performance of SfS for scenes with unknown reflectance variation and lighting. The depth information not only helps to resolve bas-relief ambiguity, but also aids in clustering pixels with similar normal directions. Such grouping allows us to effectively estimate relative albedos and the environment illumination in terms of spherical harmonics. To handle the holes in a depth map, we use edges from the RGB image to guide a structure-preserving hole filling process and create a reliable depth map proxy for our shading-based shape refinement algorithm. The utilization of a noisy, incomplete depth map in our approach leads to high-quality 3D scene reconstruction, as exemplified in Figure 1.

### 2. Related Work

Our work is related to SfS and depth map enhancement. Recent advances in SfS aim to relax strict assumptions about lighting and reflectance. In [8], Johnson and Adelson show that the inherent complexity of natural illumination actually benefits shape estimation instead of introducing greater ambiguity. Their work uses a reference sphere

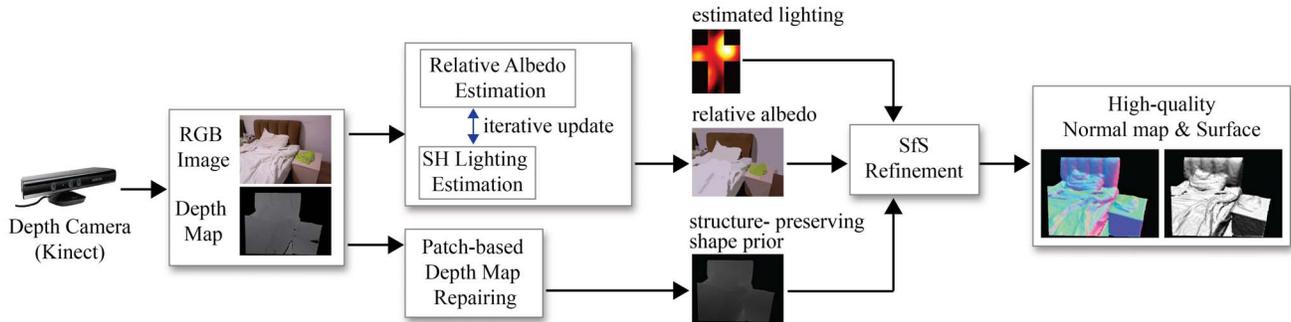


Figure 2. Flowchart of our approach.

with the same reflectance properties as the target object to model the object’s shading under the environment illumination [16]. Instead of assuming Lambertian reflectance, Oxholm and Nishino consider arbitrary isotropic BRDFs [12], with the illumination environment acquired using a reflective sphere. Recently, Barron and Malik [1] proposed an SfS approach that enforces multiple priors on shape, albedo and illumination in estimating those properties. Our approach differs from these recent techniques in that it employs an RGB-D camera, but does not require a calibration target, an assumption of uniform albedo, or reliance on smoothness and entropy constraints which may be unsuitable for the given scene. This makes our approach more general and robust in practice. For a more extensive review of SfS, we refer readers to various surveys [6, 22, 4].

Apart from single image approaches, another direction is to use a shape prior to constrain the solution space of SfS. Huang and Smith [7] interpolate the boundary normals of an object to obtain a rough shape prior to constrain SfS. Wu *et al.* [19] use multiview stereo to obtain rough but reliable geometry and use it to resolve the local ambiguity of SfS. After that, the SfS solution is used to enhance the multiview stereo geometry by integrating subtle details from SfS. Such a solution, however, cannot be directly applied with an RGB-D image that contains substantial noise and holes. Their method also does not handle objects with reflectance variation.

With regard to depth map enhancement, recent advances use an additional RGB image to denoise and upsample a depth map [20, 3, 13]. With an RGB image that has a higher resolution and signal-to-noise ratio than the depth map, a direct approach is to apply a joint bilateral filter [20, 3] using the RGB image to define a neighborhood smoothness term. In [13], Park *et al.* formulate this as an optimization problem and show that with a small amount of user interaction, the depth map can be greatly improved. But while these depth map enhancement methods can reduce noise and increase resolution, they also lose fine depth details during the smoothing process. By contrast, our approach recovers fine depth details even if they are not captured in the initial

noisy depth map, by making greater use of the RGB image through an analysis of its shading.

### 3. Depth-assisted SfS Approach

To facilitate SfS, our approach utilizes partial depth information to separate shading from albedo, aid illumination estimation, and resolve surface normal ambiguity. No assumptions are made on the incident illumination or surface geometry, while the reflectance in the scene is taken to be Lambertian.

#### 3.1. Overview

Figure 2 displays a flowchart of our algorithm. From the input RGB image and depth map, our method first computes a normal map from the captured depth map and segments the RGB image into regions of piecewise smooth color. Through alternating optimization (AO), the relative albedos among the different regions are calculated, and the environment illumination is estimated from the albedo-normalized image. After that, we estimate normals over the whole image using SfS with the help of a normal map computed from Kinect as a shape prior to resolve bas-relief ambiguity. For regions that lack depth map values from Kinect, we use a constrained texture synthesis to fill in the missing depth values prior to applying our normal estimation algorithm. As shown in Figure 1, the output of our method is a refined normal map without the shape and reflectance ambiguities of SfS nor the noise and holes of the Kinect range data.

#### 3.2. Relative Albedo and Lighting Estimation

The input from Kinect consists of an RGB image  $I = \{I_i\}$ ,  $I_i = [I_{i,r}, I_{i,g}, I_{i,b}]^T$  where  $i$  is the pixel index, and a depth map. From the point cloud determined from the depth map, we calculate a rough normal map  $N = \{n_i\}$ , where  $n_i = [n_{i,x}, n_{i,y}, n_{i,z}]^T$  is the unit normal at pixel  $i$ , obtained by a simple cross-product of the neighboring points. For pixels with missing depth values, or whose neighboring pixels have any missing depth values, no initial normal is computed.

### 3.2.1 Relative Albedo from Common Normals

We first perform a mean-shift clustering on the RGB image, with a minimum region size of 500 pixels. Suppose this forms a set of  $S$  clusters  $\mathbb{C} = \{C_u, u = 1, \dots, S\}$ . Each cluster contains a set of pixels and a corresponding set of normals. Under consistent environment lighting, any two pixels  $a$  and  $b$  with same normal direction in two different clusters have the same shading, and thus the differences between their pixel values are due only to differences in their relative albedos,  $p_{a,k}$  and  $p_{b,k}$ :

$$\frac{I_{a,k}}{I_{b,k}} = \frac{p_{a,k}}{p_{b,k}}$$

where  $k = 1, 2, 3$  respectively index the RGB channels. With this property, we solve for the relative albedos between different clusters using pixel-pairs of common normals from different clusters. We note that intensity ratios have also been used as an illumination invariant for object recognition [5, 11].

### 3.2.2 Data Structure

To facilitate normal direction comparisons among clusters, we quantize all possible normal directions to vertices on an icosahedron, which provides a uniformly-distributed set of  $T = 642$  normal directions over a sphere. The normals in an image are stored in a data structure  $B_{u,j,k}$  which we refer to as *bins*, where  $u = 1, \dots, S$  denotes the cluster index,  $j = 1, \dots, T$  denotes the normal directions as sampled from the icosahedron, and  $k = 0, \dots, 3$  with  $B_{u,j,0}$  as an indicator bit of whether the  $j$ -th normal direction exists within cluster  $C_u$ , and  $[B_{u,j,1}, B_{u,j,2}, B_{u,j,3}]$  store the RGB values corresponding to the  $j$ -th normal direction in cluster  $C_u$ .

All the bin values are initialized to zeros. Then, for each cluster  $C_u$ , each normal  $n$  falls into a bin  $B_{u,t,k}$ , where  $n$  has the smallest dot-product with the  $t$ -th normal direction among all the  $T$  normal directions on the icosahedron. We set  $B_{u,t,0} = 1$  to indicate that this bin is utilized. Then we fill in  $B_{u,t,k}$ , where  $k = 1, 2, 3$ , with the RGB values of the pixel with normal  $n$ . If there are multiple pixels having normals that fall into the same bin  $B_{u,t,k}$ , the median of their RGB values is used.

### 3.2.3 Graph Representation

After the data structure is built, we represent the common-normal-direction relationships between different clusters as a graph,  $\mathbb{G} = \{\mathbb{C}, \mathbb{E}\}$ . Each cluster  $C_u$  is represented as a node. An edge  $E_{u_1, u_2}$  exists between cluster  $C_{u_1}$  and  $C_{u_2}$  only if there are more than  $\lambda$  common normal directions between clusters  $C_{u_1}$  and  $C_{u_2}$ , with  $\lambda = 20$  in our experiments. The edge is given a score equal to the number of

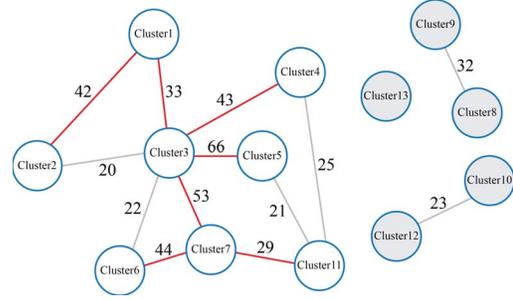


Figure 3. Graph of common-normal-direction relationships among clusters, with the maximum spanning tree indicated by red edges.

common normal directions. Refer to Figure 3 for an example graph.

In estimating a globally consistent set of relative albedos, we utilize the maximum spanning tree (MST) algorithm to determine a cycle-free set of links that maximize the number of common normal directions between clusters. As the graph may be disconnected, a forest of trees may be formed. After the MST is found, we calculate the relative albedos between all of its clusters in a depth-first search order along the tree.

The relative albedo between two clusters is computed by first determining the common bins (corresponding to common normals) utilized in clusters  $C_{u_1}$  and  $C_{u_2}$ , denoted by  $Q = \{q : B_{u_1,q,0} = 1 \text{ and } B_{u_2,q,0} = 1\}$ . Then we obtain an estimate of relative albedo of  $C_{u_2}$  over  $C_{u_1}$  for each of the RGB channels:

$$p_{u_2,k} = \frac{B_{u_2,q,k}}{B_{u_1,q,k}}$$

Among all common bins, we run RANSAC to obtain the relative albedo estimates in a manner robust against outliers. Pseudocode of this relative albedo estimation procedure is provided in the supplementary material.

### 3.2.4 Lighting Estimation

The estimated relative albedos are highly useful. By normalizing the albedos in different regions, we can then jointly use their rich variety of normal directions to more reliably estimate the environment lighting.

Suppose there are  $R$  pixels whose relative albedos are estimated from the MST, and let  $\hat{n}_i = [n_i^T \ 1]^T$ . We estimate the lighting in terms of 2nd order spherical harmonics (SH) for each RGB channel  $k = 1, 2, 3$ :

$$\hat{n}_i^T M_k \hat{n}_i = \frac{I_{i,k}}{p_{i,k}} \quad (1)$$

where  $i = 1, \dots, R$  and  $M_k$  depends on the SH coefficients for the  $k$ -th RGB channel [16]. Using the RGB image  $I$  and initial normal map  $N$  computed from Kinect,  $M_k$  in (1) can be estimated up to a scale factor by linear least-squares minimization.



Figure 4. Relative albedos estimated by our alternating optimization. (a) Input image, (b) & (c) Relative albedos estimated at the 1st and 5th iteration, (d) & (e) Corresponding shading images at the 1st and 5th iteration. Clusters without relative albedos in the 1st iteration are simply filled by original RGB values in (b).

### 3.2.5 Refinement by Alternating Optimization

With the estimated lighting, we refine the relative albedos and calculate the relative albedos of those clusters not yet estimated. For each cluster, an estimate of relative albedo for each RGB channel  $k$  is obtained for *each normal*  $n_i$  in the cluster as:

$$p_{i,k} = \frac{I_{i,k}}{\hat{n}_i^T M_k \hat{n}_i}. \quad (2)$$

RANSAC is again run on these estimates to obtain an updated relative albedo for each cluster. Using the updated relative albedos of the MST clusters, we re-estimate the SH coefficients by (1). This alternating optimization process is repeated until the change falls below a small value. In practice, convergence is obtained in 3-5 iterations. An example of the improvements gained through iterative optimization is shown in Figure 4. We note that despite the noisy normals of the depth map, the relative albedos between two regions can be reliably determined when they have many normals in common, as is the case for connected nodes in the MST. Moreover, the environment lighting can also be dependably recovered when the number and range of noisy normals is large, as again is the case with the MST.

## 3.3. Geometry Estimation

### 3.3.1 Structure-preserving Shape Prior

Shape-from-shading on a single image is an ill-posed problem that suffers from bas-relief ambiguity [2] (see Figure 5) unless a shape prior is enforced. In our work, we exploit the Kinect RGB-D data to obtain a structure-preserving shape prior, in the form of prior normals to be used later in a normal refinement step.

Kinect depth maps, however, frequently contain holes where there is no depth information for directly computing surface normals. Rather than filling the holes by smooth interpolation, which tends to lose sharp edges and corners

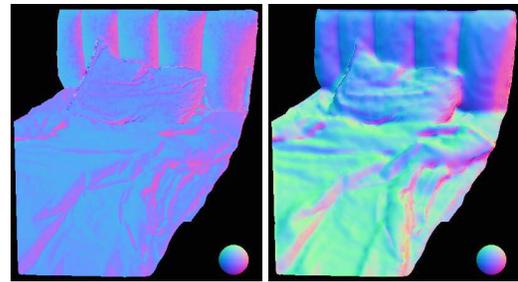


Figure 5. Effect of prior normals on handling bas-relief ambiguity. Left: without prior normals, the bed is roughly co-planar with the backboard. Right: by accounting for prior normals, the bed normals are correctly pointing upward.

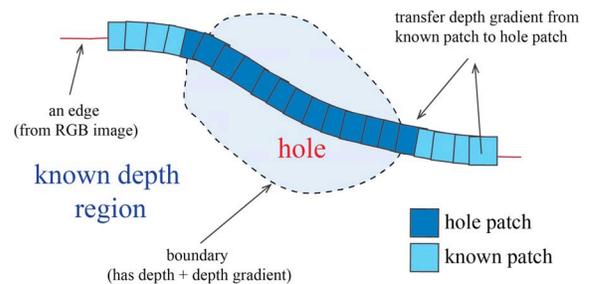


Figure 6. Illustration of patch-based repairing of a structural hole.

(see Figure 8), we estimate the missing data in a structure-preserving manner, similar in spirit to [17] but with different considerations due to our different problem setting.

Though holes may exist in the depth image, they do not appear in the corresponding RGB image. We thus take advantage of the RGB image as a guide for depth completion in the hole region. First, a Canny edge detector is applied to the RGB image. We then identify RGB edges that pass through a hole, referred to as a *structural hole*, in the depth image. Along the edge, we generate *hole patches* which contain hole pixels whose depths need to be obtained, and *known patches* which contain no hole and are used for repairing the hole patches. Figure 6 shows an illustration.

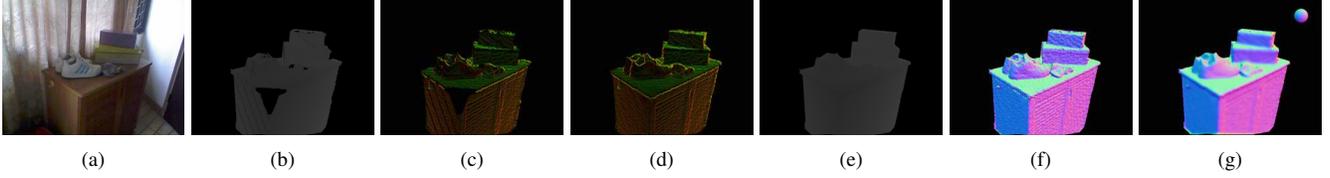


Figure 7. Example of repairing a depth map hole. (a) Input RGB, (b) Input depth, (c) Depth gradient map, (d) Depth gradient map after patch repair, (e) Depth map after patch repair and poisson integration, (f) Prior normal map, (g) Resulting normal map after SfS.

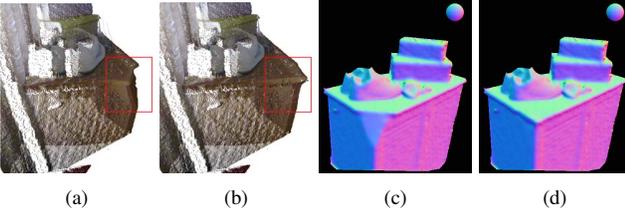


Figure 8. Example of patch-based repairing versus smoothing to obtain a structure-preserving shape prior for the example scene in Figure 7. Patch-based repairing allows propagation of existing structure to the hole region. (a) Shape prior using poisson smoothing. (b) Shape prior using patch-based repairing. (c) & (d) Resultant normals using shape prior from (a) & (b).

The goal is to transfer the depth gradients from the known patches to the hole patches, after which the depth of the hole can be filled in by poisson integration while preserving the structure along the edge. This structural propagation is formulated as an MRF which is solved by belief propagation [14]. The MRF total cost function for the set of hole patches  $\mathbb{H}$  is defined as:

$$\mathcal{C}_{BP}(\mathbb{H}) = w_{D_{rgb}} \mathcal{C}_{D_{rgb}}(\mathbb{H}) + w_{S_{rgb}} \mathcal{C}_{S_{rgb}}(\mathbb{H}) + w_{D_{dg}} \mathcal{C}_{D_{dg}}(\mathbb{H}) + w_{S_{dg}} \mathcal{C}_{S_{dg}}(\mathbb{H}) \quad (3)$$

where  $\mathcal{C}_{D_{rgb}}(\mathbb{H})$ ,  $\mathcal{C}_{S_{rgb}}(\mathbb{H})$ ,  $\mathcal{C}_{D_{dg}}(\mathbb{H})$  and  $\mathcal{C}_{S_{dg}}(\mathbb{H})$  are respectively the RGB data cost, RGB smoothness cost, depth gradient data cost and depth gradient smoothness cost. We set  $w_{D_{rgb}} = 1.0$ ,  $w_{D_{dg}} = 1.0$ ,  $w_{S_{rgb}} = 0.1$  and  $w_{S_{dg}} = 0.1$  in our experiments. Each cost term is detailed as follows.

Denote the set of hole patches as  $\mathbb{H} = \{\mathbf{H}_l\}$  and the set of known patches as  $\mathbb{K} = \{\mathbf{K}_m\}$ .  $\mathbf{H}_l$  is itself a set containing all pixels that lie within the patch, with each pixel indexed by local patch coordinate  $\mathbf{p}$ . For notational convenience, we also define  $H_l(\mathbf{p})$  as the pixel location in image coordinates, such that  $I(H_l(\mathbf{p}))$  is the RGB intensity of the pixel, and likewise for  $I(K_m(\mathbf{p}))$ . Also,  $H_l^{-1}$  returns the corresponding known patch's index, such that  $\mathbf{K}_{H_l^{-1}}$  is the patch that repairs  $\mathbf{H}_l$ .

**RGB Data Cost:** Let  $Z_{D_{rgb}}$  denote the number of pixels covered by hole patches. The RGB data cost is defined so that the selected known patch closely matches the hole patch in the RGB image:

$$\mathcal{C}_{D_{rgb}}(\mathbb{H}) = \frac{1}{3Z_{D_{rgb}}} \sum_l \sum_{\mathbf{p} \in \mathbf{H}_l} \|I(H_l(\mathbf{p})) - I(K_{H_l^{-1}}(\mathbf{p}))\|^2. \quad (4)$$

**Depth Gradient Data Cost:** Let  $Z_{D_{dg}}$  be the number of non-hole pixels covered by hole patches, and  $D'$  as the depth gradient image. Since these pixels have depth values, their depth gradients can be calculated. The depth gradient data cost favors solutions in which the computed depth gradients closely match the original depth gradients for the non-hole pixels:

$$\mathcal{C}_{D_{dg}}(\mathbb{H}) = \frac{1}{2Z_{D_{dg}}} \sum_l \sum_{\mathbf{p} \in \mathbf{H}_l} \alpha(H_l(\mathbf{p})) \|D'(H_l(\mathbf{p})) - D'(K_{H_l^{-1}}(\mathbf{p}))\|^2 \quad (5)$$

where

$$\alpha(H_l(\mathbf{p})) = \begin{cases} 1 & H_l(\mathbf{p}) \text{ has a depth gradient} \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

**RGB Smoothness Cost:** Suppose  $\{\mathbf{H}_{l_1}, \mathbf{H}_{l_2}\}$  is a pair of overlapping hole patches, and  $\mathbf{K}_{H_{l_1}^{-1}}$  and  $\mathbf{K}_{H_{l_2}^{-1}}$  respectively denote their repairing known patches. Suppose also that pixel  $\mathbf{p}_a$  of  $\mathbf{K}_{H_{l_1}^{-1}}$  coincides with pixel  $\mathbf{p}_b$  of  $\mathbf{K}_{H_{l_2}^{-1}}$ , when  $\mathbf{K}_{H_{l_1}^{-1}}$  and  $\mathbf{K}_{H_{l_2}^{-1}}$  are pasted onto  $\mathbf{H}_{l_1}$  and  $\mathbf{H}_{l_2}$ . With  $Z_{ov}$  being the number of pixels in the overlapping regions of hole patches, we penalize solutions where the overlapping RGB values are inconsistent:

$$\mathcal{C}_{S_{rgb}}(\mathbb{H}) = \frac{1}{3Z_{ov}} \sum_{\{\mathbf{H}_{l_1}, \mathbf{H}_{l_2}\}} \sum_{\{\mathbf{p}_a, \mathbf{p}_b\}} \|I(K_{H_{l_1}^{-1}}(\mathbf{p}_a)) - I(K_{H_{l_2}^{-1}}(\mathbf{p}_b))\|^2. \quad (7)$$

**Depth Gradient Smoothness Cost:** Similar to the RGB smoothness cost, we have a corresponding cost for the depth gradient image:

$$\mathcal{C}_{S_{dg}}(\mathbb{H}) = \frac{1}{2Z_{ov}} \sum_{\{\mathbf{H}_{l_1}, \mathbf{H}_{l_2}\}} \sum_{\{\mathbf{p}_a, \mathbf{p}_b\}} \|D'(K_{H_{l_1}^{-1}}(\mathbf{p}_a)) - D'(K_{H_{l_2}^{-1}}(\mathbf{p}_b))\|^2. \quad (8)$$

After belief propagation is performed to minimize  $\mathcal{C}_{BP}(\mathbb{H})$ , depth gradients of pixels within hole patches are replaced by depth gradients from the assigned known patches. With the transferred depth gradients and the known depth values along the hole boundary as boundary conditions, poisson integration [15] is used to fill in the depth values of the hole. Figure 7 illustrates this process.

### 3.3.2 Surface Normal Refinement

The estimated relative albedos, lighting and shape prior serve as useful inputs for normal refinement over the whole

scene. Suppose there are in total  $Z_{total}$  pixels. The surface normal refinement is formulated as a non-linear optimization using the total energy function:

$$\mathcal{E}(N) = w_{sfs}\mathcal{E}_{sfs}(N) + w_{prior}\mathcal{E}_{prior}(N) + w_{smooth}\mathcal{E}_{smooth}(N) + w_{norm}\mathcal{E}_{norm}(N). \quad (9)$$

$\mathcal{E}_{sfs}(N)$  is the shape-from-shading cost represented using 2nd order spherical harmonics. It constrains the normal according to the shading observed in the RGB image:

$$\mathcal{E}_{sfs}(N) = \frac{1}{Z_{total}} \sum_i \sum_{k=\{1,2,3\}} (I_{i,k} - p_{i,k}\hat{n}_i^T M_k \hat{n}_i)^2 \quad (10)$$

To resolve bas-relief ambiguity,  $\mathcal{E}_{prior}(N)$  constrains the normals to be similar to the prior normals computed from the repaired Kinect depth map (see Figure 5). Denote the prior normal as  $n'_i$ :

$$\mathcal{E}_{prior}(N) = \frac{1}{Z_{total}} \sum_i \|n_i - n'_i\|^2. \quad (11)$$

$\mathcal{E}_{smooth}(N)$  is a smoothness term with respect to 1st-order neighbors. For the set of 1st-order neighbors,  $\{i_1, i_2\}$ , we have:

$$\mathcal{E}_{smooth}(N) = \frac{1}{Z_{total}} \sum_{\{i_1, i_2\}} \|n_{i_1} - n_{i_2}\|^2. \quad (12)$$

Finally,  $\mathcal{E}_{norm}(N)$  is the norm regularization which constrains the normals to be of unit length:

$$\mathcal{E}_{norm}(N) = \frac{1}{Z_{total}} \sum_i (n_i^T n_i - 1)^2. \quad (13)$$

The total energy function  $\mathcal{E}(N)$  is a weighted sum of the four energy terms, with the weights fixed to  $w_{sfs} = 1.0$ ,  $w_{prior} = 0.1$ ,  $w_{smooth} = 0.05$  and  $w_{norm} = 0.05$ . The total energy function, which is non-linear in terms of normals  $n_i$ , is optimized by the trust-region-reflective algorithm. We initialize the normals to  $[0, 0, 1]^T$ , facing the camera.

## 4. Experimental Results

### 4.1. Lighting Estimation

In Figure 9, we investigate our approach’s ability to estimate environment light in an indoor scene, by comparing it to ground truth obtained using a mirrored sphere convolved with 2nd-order spherical harmonics. It can be observed that using more clusters and normals, which is made possible by the relative albedo estimation, leads to more accurate and robust light estimation. As the normals throughout the MST are used, the major light directions and intensity resemble that obtained from the mirrored sphere. Figure 10 also shows iterative refinement of light estimation throughout the alternating optimization process. We note that inconsistency in the environment light across the scene due to non-distant light sources will contribute to error.

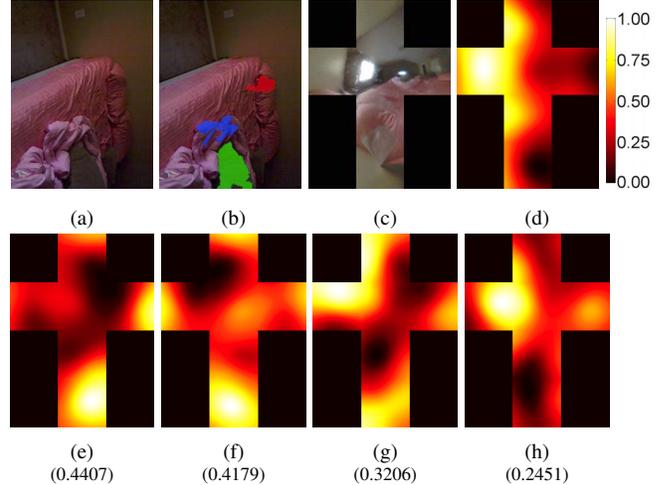


Figure 9. Light estimation experiment. (a) Input scene. (b) Clusters colored for illustration. (c) Ground truth environment map. (d) Ground truth 2nd-order SH. (e) Estimation by red cluster in (b). (f) Estimation by green cluster. (g) Estimation by blue cluster. (h) Estimation by all regions in the MST. Bracketed numbers show RSME.

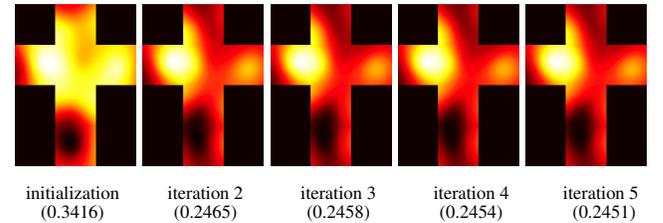


Figure 10. Iterative refinement of light estimation throughout AO. Bracketed numbers show RMSE, which is converging.

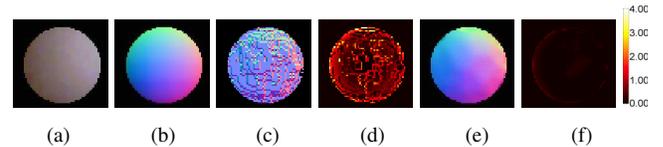


Figure 11. Normal estimation of a Lambertian ball in the scene, *jeans*. (a) Input image. (b) Ground-truth normal map. (c) Raw normal map. (d) Squared error map of raw normals (RMSE=0.5178), (e) Our estimated normal map, (f) Squared error map of our estimated normals (RMSE=0.1401).

### 4.2. Ground Truth Comparison

Next we validate our approach by conducting an analytical experiment in which we estimate normals of a Lambertian ball in an indoor scene (named *jeans* in the supplement). Figure 11 shows the results of our approach in refining the raw normals computed directly from the depth map. The RMSE is improved from 0.5718 to 0.1401<sup>2</sup>. The more apparent error along the sphere boundary is due to the greater noise in Kinect RGB images near object boundaries.

<sup>2</sup>While the RMSE of relative light intensity is in the range  $[0, 1]$ , the RMSE of normals is in the range  $[0, 2]$ , as the squared error of normals is in range  $[0, 4]$ . For example, normals  $[0, 0, 1]^T$  and  $[0, 0, -1]^T$  result in a maximum squared error of 4.

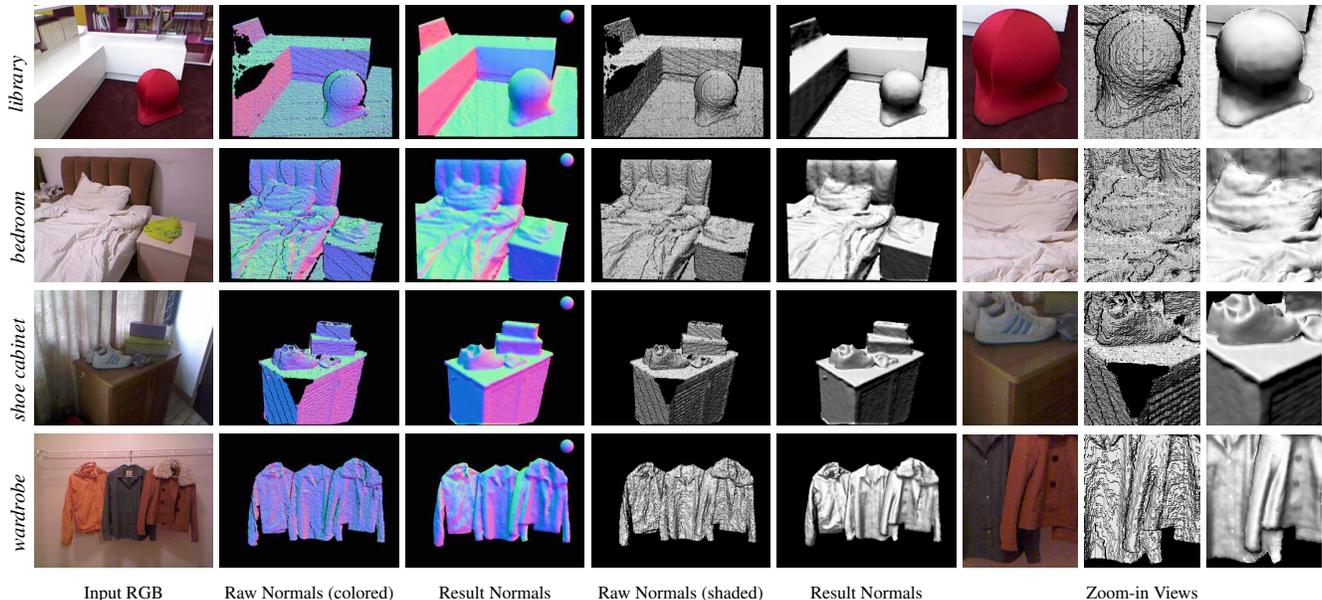


Figure 12. Kinect scenes repaired by our approach.

### 4.3. Repairing Kinect Scenes

We tested our approach on four indoor scenes captured by Kinect, namely, *library*, *bedroom*, *shoe cabinet* and *wardrobe*. These are common indoor scenes with shading detail that our approach can make use of to refine the reconstructed surface. Figure 12 shows the results. In *library*, the *structural holes* on the books and shelf are repaired by the propagated patches, and the round surface of the stool is well reconstructed by shading despite the presence of noise and holes in the input depth and normal map. In *bedroom*, details of the pillow are faithfully reconstructed, e.g., the crease at the top-right corner. In *shoe cabinet*, structural propagation enables the proper repair of the hole at the corner, which provides a correct shape prior compared to smoothing (see also Figure 8). To this, shading adds further details, e.g., the marks on the shoe. Finally, in *wardrobe*, shape-from-shading significantly improves the surface where very fine details such as the folded collar and button regions can be clearly seen. Please refer to the supplementary materials for three additional results.

### 4.4. Comparison with Other Methods

To demonstrate the possible improvements obtainable with noisy Kinect depth data in our method, we compare our depth-assisted approach with a state-of-the-art shape-from-shading algorithm [1], which operates with only an RGB image using generic albedo and illumination priors. As shown in Figure 13, our depth-assisted method achieves significantly better surface normal reconstructions. We believe that the priors used in [1] may be more appropriate for single objects than for full scenes that are captured

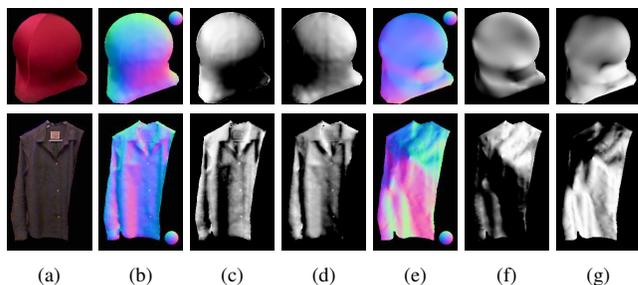


Figure 13. Comparison to SfS technique of [1]. (a) Input RGB image. (b-d) Our recovered normals and two normal maps  $N$  shaded as  $N \cdot L$  with  $L = (-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})^T$  and  $L = (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})^T$ . (e-g) Recovered normals and shaded images of [1] using generic albedo and illumination priors.

by a Kinect. In this comparison, we used the code provided in [1] with the default parameters. Our approach uses only the regions with the highest-confidence relative albedos (from the MST) for lighting estimation, rather than the entire image. Our supplement contains additional results.

Figure 14 compares our albedo normalization result with the state-of-the-art intrinsic image separation technique of [9], which also makes use of Kinect depth data. The result of [9] was provided to us by the authors. Their work assumes the input to be a nearly flawless depth map obtained from video streams of a moving Kinect, and does not operate as well with a noisy depth map available from a single Kinect image. In contrast, our technique performs more effective albedo normalization because the relative albedos are obtained with the help of estimated lighting. This results in more refined shading details, e.g., on the bed.



Figure 14. Left: albedo normalization result of Figure 9 by [9]. Right: our result.

## 5. Discussion

High-quality normals are vital prerequisites for different practical applications. Figure 15 shows a point cloud significantly refined with our resultant normals using the method of [10]. In addition, the resultant normals enable realistic re-lighting and high-quality 3D surface reconstruction. We kindly refer readers to our supplementary video for various demonstrations and comparisons.

**Limitations:** Like other patch-based image completion methods, the effectiveness of our patch-based hole repairing step is subject to the quality and compatibility of the surrounding known patches. While the RGB data is in general of higher quality than the depth data, its noise can still affect the quality of shape-from-shading. For scenes with local light sources, the environment light may differ significantly in different parts of the scene. This issue could potentially be addressed by solving for the environment light separately among local regions.

**Conclusion:** We presented a useful postprocessing method to improve the quality of surface normals obtained from Kinect. When used with the latest Kinect, which has higher resolution in RGB than in depth, the proposed method could also be utilized for the problem of depth map denoising and upsampling [20, 3, 13], since the geometry is solved at the RGB image resolution and its use of shading significantly reduces the effects of depth sensor noise. In future work, we plan to consider the lighting visibility of scene points based on the depth map, as this should improve the estimation of lighting, relative albedos, and shape-from-shading.

**Acknowledgements** This work was partially supported by Singapore University of Technology and Design (SUTD) StartUp Grant ISTD 2011 016, by SUTD-MIT International Design Centre (IDC) Research Grant IDSF12001100H, and by the National Research Foundation (NRF: 2012-0003359) of Korea funded by the Ministry of Education, Science and Technology. We thank Shuochen Su for his help on results.

## References

- [1] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV (4)*, pages 57–70, 2012.
- [2] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *IJCV*, 35(1):33–44, Nov. 1999.

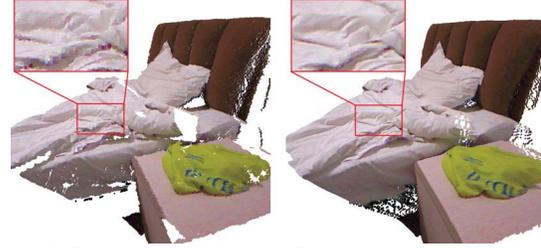


Figure 15. Left: raw point cloud. Right: point cloud refined with our resultant normals.

- [3] J. Dolson, J. Baek, C. Plagemann, and S. Thrun. Upsampling range data in dynamic environments. In *CVPR*, 2010.
- [4] J.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *CVIU*, 109(1):22–43, 2008.
- [5] B. V. Funt and G. D. Finlayson. Color Constant Color Indexing. *PAMI*, 17(5):522–529, May 1995.
- [6] B. K. P. Horn and M. J. Brooks. *Shape from shading*. MIT Press, Cambridge, MA, USA, 1989.
- [7] R. Huang and W. A. P. Smith. Shape-from-shading under complex natural illumination. In *ICIP*, pages 13–16, 2011.
- [8] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *CVPR*, pages 2553–2560, 2011.
- [9] K. J. Lee, Q. Zhao, X. Tong, M. Gong, S. Izadi, S. U. Lee, P. Tan, and S. Lin. Estimation of intrinsic image sequences from image+depth video. In *ECCV*, pages 327–340, 2012.
- [10] Z. Lu, Y.-W. Tai, M. Ben-Ezra, and M. S. Brown. A framework for ultra high resolution 3d imaging. In *CVPR*, 2010.
- [11] S. K. Nayar and R. M. Bolle. Reflectance based object recognition. *IJCV*, 17(3):219–240, 1996.
- [12] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *ECCV (1)*, pages 528–541, 2012.
- [13] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *ICCV*, 2011.
- [14] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
- [15] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, July 2003.
- [16] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, 2001.
- [17] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum. Image completion with structure propagation. *ACM Trans. Graph.*, 24(3), 2005.
- [18] R. Woodham. Photometric method for determining surface orientation from multiple images. *Opt. Eng.*, 19(1), 1980.
- [19] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *CVPR*, pages 969–976, 2011.
- [20] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *CVPR*, 2007.
- [21] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, D. Terzopoulos, and T. F. Chan. Outdoor photometric stereo. In *ICCP*, 2013.
- [22] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE PAMI*, 21(8):690–706, 1999.