Joint estimation of depth, reflectance and illumination for depth refinement

Kichang Kim Akihiko Torii Masatoshi Okutomi
Tokyo Institute of Technology
{kichang.k@ok., torii@, mxo@}ctrl.titech.ac.jp

Abstract

In this paper we propose a method for joint estimation of depth, reflectance and illumination from a single RGB-D image for depth refinement. This is achieved by a simple optimization based approach with smoothness constraints on depth, reflectance and illumination. We introduce an adaptively weighted local similarity constraint for reflectance, a normalized spherical-harmonic model for illumination, and an edge-aware local smoothness constraint for depth. This allows us to generate high quality depth without additional processes such as pre-training of stochastic models or image segmentation. Experimental results demonstrate that our method estimates high quality depth in comparison with ground-truth data not only for laboratory conditions but also for complex real-world scenes.

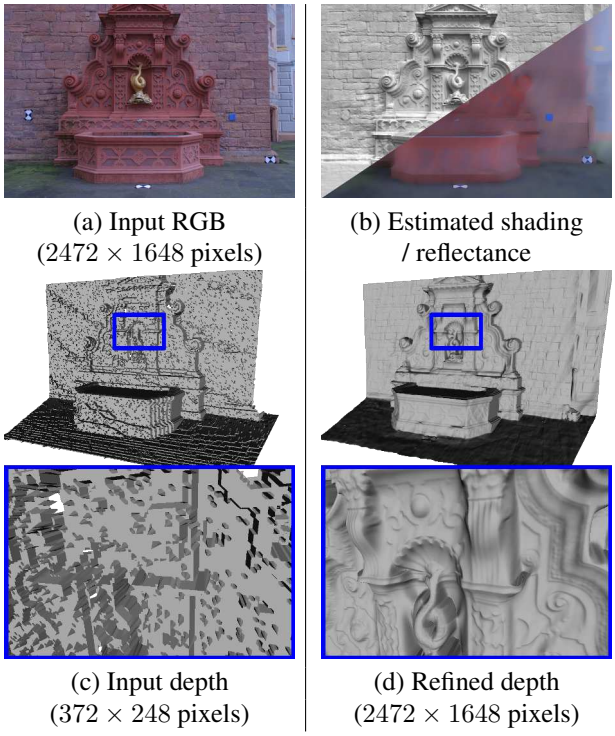
1. Introduction

3D modelling has become attractive due to recent advances in Structure from Motion [19, 29], Multiple View Stereo [7, 16], as well as sensor devices [1, 35]. Commercial RGB-D sensors like Kinect has made it easier for non-experts [15] to explore computer vision applications [28].

Since RGB-D sensors can acquire both 2D images and relatively rough depths, it also attracts researchers tackling the problem of estimating shape, reflectance and illumination which is known to be an extremely ill-posed problem [12, 21]. The estimation becomes more tractable by assuming that one or two of the parameters are given. This has been well studied in the context of shape from shading [13, 14, 21] and extended in photometric stereo [11, 31].

Using the depth image obtained from an RGB-D sensor as an initial rough estimate, recent works [2, 33] jointly or sequentially estimate reflectance, lighting and depth in high quality. Such methods work well in scenes with relatively simple objects. However, for general scenes with complex textures and object-shapes, high quality joint estimation remains a major challenge.

In this work, we aim to fill the quality gap between a high quality 2D image and a low quality depth image (figure 1). The main contribution is the design of a cost function with constraints for each component: depth, reflectance and illumination. We propose an adaptively weighted local similarity constraint for reflectance, a normalized spherical-harmonic model for illumination, and an edge-aware local smoothness constraint for depth as described in section 3. The proposed approach is able to handle scenes including objects with complex textures which are problematic for image-segmentation based approaches. Furthermore, our



(a) Input RGB (2472 × 1648 pixels)
(b) Estimated shading / reflectance
(c) Input depth (372 × 248 pixels)
(d) Refined depth (2472 × 1648 pixels)

Figure 1: **Joint estimation of depth, reflectance and illumination from a single RGB-D image.** Given a pair of high-quality RGB image (a) and a noisy low-quality depth (rendered as 3D shape (c)), we decompose the RGB image into reflectance, illumination (b) and depth using the low-quality depth as the initial guide. This enables us to generate a high quality shape (d).

method requires no pre/post processing for depth refinement. We demonstrate that such a simple formulation is already efficient at improving depth quality of real-world data in section 4.

We next describe fundamental works and important algorithms closely related to our work.

2. Related works

Shape from X. One of the earliest work for geometry estimation is Shape From Shading (SFS) introduced by Horn [13] that estimates surface normals (as geometry) from a single image under known-illumination and known-reflectance using a simple rendering model: $I = S(N)R$ where I is intensity, S is shading as a function of a surface normal N and R is reflectance. Forsyth [6] extended this simple rendering model to a more flexible model: spatially varying shading model. This allows them to handle variations in illumination resulting from inter-reflections, complex light sources, etc. Johnson and Adelson [18] showed that the restriction on reflectance can be relaxed by approximating natural illumination as a low order quadratic function.

Photometric Stereo (PS) can relax the restrictions in reflectance and illumination on SFS by using multiple images captured under varying illuminations. For a fixed viewpoint, Woodham [31] demonstrated that shape can be estimated under calibrated illuminations. Basri and Jacobs [3] then extended this to the uncalibrated case by representing illuminations as a low-order spherical harmonics model. Hernandez *et al*. [11] relaxed the fixed view condition to the non-fixed view condition by using visual hull information.

In another scope, given feature correspondences, Structure From Motion (SFM) estimates camera poses and (sparse) depths of feature points, which can be densified by Multiple View Stereo (MVS) [7]. SFM and MVS have been studied as a different context to SFS and photometric stereo but, recently, are often used as initial input for performing PS [27] and intrinsic image decomposition as described below.

Intrinsic image decomposition. Our work is also related to the intrinsic image decomposition which decomposes an image into reflectance and shading. As an earlier work, Horn [12] used gradient of magnitude to derive reflectance and illumination from a single image. Zhao *et al*. [36] showed that non-local texture similarity can be used for reducing decomposition ambiguity in intrinsic image decomposition. Bell *et al*. [4] proposed a novel algorithm based on the conditional random fields formulation. They also introduced large scale public data set for decomposition evaluation.

Intrinsic image decomposition, given shapes can be used for relaxing decomposition ill-posedness. Yu and Malik [34] used an inverse rendering approach for estimating reflectance with shape. Haber *et al*. [9] estimated a complex BRDF component and illumination from multiple images and a given shape. Laffont *et al*. [20] estimates relative diffuse reflectance using surface normals from MVS. They calculate relative reflectance by using the ratio of pixel values that have the same normals.

After commercial RGB-D sensors were introduced, several intrinsic image decomposition approaches using an RGB image and depths were introduced. Lee *et al*. [22], Chen and Koltun [5] and Jeon *et al*. [17] used a non-local smoothness term on reflectance. They assumed that if pixels have close normals, the pixels have similar shadings.

Shape refinement. To obtain high quality shape, there are various methods that use the data fusion approach. Nehab *et al*. [24] showed that precise shape can be obtained by refining position using surface normals. Boxin *et al*. [27] and Park *et al*. [25] showed that high quality shape can be obtained by a combination of photometric stereo and MVS. Barron and Malik [2] presented a novel statistical approach to decompose an RGB-D image into reflectance, illumination and refined depth. They used pre-trained GMM to resolve decomposition ambiguities on each component. They also represented shape and illumination as a combination of multiple segments to handle object occlusions and local variations of illumination.

Yu *et al*. [33] and Han *et al*. [10] introduced a shading based approach for shape refinement from an RGB-D image. Yu *et al*. [33] estimated scene illumination using initial depth and refine depth using estimated illumination. For reflectance, they used mean-shift clustering to segment the RGB image into small areas that have a uniform albedo. The relative albedo is then calculated between each segment in the same manner as [20]. Han *et al*. [10] proposed a similar depth refinement method for a uniform albedo object. They then estimate local lighting parameters for illumination which cannot be represented with a single spherical-harmonic illumination model. However they still require explicit image segmentation for handling multi-albedo objects. Wu *et al*. [32] proposed a real-time depth refinement algorithm from an RGB-D image with the GPU-accelerated Newton method.

Our method. Our method is closely related to [2] which jointly estimates depth, reflectance and illumination by formulating it as an energy minimization framework. However, we focuses on estimating high-quality shapes as the motivation as done in [10, 33]. Our method can be applied to general scenes while requiring neither pre-training nor explicit segmentation to deal with multiple reflectances. In the energy function, we propose a new smoothness constraint for reflectance where their weights are adaptively de-

fined from chromaticity and intensity changes. This clearly differs from the low intensity suppression idea used in [17]. We also use a soft constraint for the image fidelity term in contrast to the hard constraint used in [2].

The details of our method are described in next section.

3. Joint estimation of depth, reflectance and illumination

Given a high-quality 2D image and a corresponding (noisy) low-quality depth, we aim to estimate the refined high-quality depth by joint estimation of depth, reflectance and illumination. To achieve this, we assume that the input 2D image can be decoupled to diffuse reflectance (albedo) R, illumination L and depth D in the similar manner to the intrinsic image decomposition [2]:

$$\underset{R,D,L}{\arg\min}\, P_I(R,D,L) + P_R(R) + P_D(D) \tag{1}$$

P_I, P_R, and P_D are the costs defined using R, D, and L. We seek the best combination that separates each component while minimizing the total cost defined in equation 1. We next describe each cost in detail.

3.1. Data fidelity term

As with other works [2, 10], we enforce that the input RGB image $\hat{I} \in \mathbb{R}^3$ and the rendered image $I(R,D,L) \in \mathbb{R}^3$ (obtained using the estimated reflectance R, illumination L and depth D) have the same appearance. Then, the data fidelity term is:

$$P_I(R,D,L) = w_{df} \sum_i^n \|I_i(R,D,L) - \hat{I}_i\|^2 \tag{2}$$

where i indicates the index of pixels and w_{df} is the weighting parameter.

Assuming Lambertian reflectance and spherical-harmonic lighting model [26], the rendering function (engine) I can be formulated as multiplication of the reflectance $R \in \mathbb{R}^3$ for three color channels and the shading S,

$$I_i(R,D,L) = S(N_i(D), L)R_i \tag{3}$$

where $N_i(D) \in \mathbb{R}^3$ is the surface normal at a pixel i computed from the depth D, and $L \in \mathbb{R}^9$ is the light basis.

Since the shading S is computed using a surface normal and illumination, we compute the surface normal as follows. For robust yet simple computation of the surface normal N_i, we form four triangle patches using depths of the neighboring pixels $t(i)$, compute the normals $\hat{N}_j$ from them, and compute the mean normal $\tilde{N}_i$ as

$$\tilde{N}_i(D) = \frac{\sum_{j \in t(i)} \hat{N}_j(D)}{4}. \tag{4}$$

Finally we obtain $N_i(D)$ as the L2-normalized vector of $\tilde{N}_i$.

Notice that we use the depths directly in the optimization process because we aim to produce high quality depths after the optimization. This differs from using and optimizing the surface normals [10, 33] which require a post-processing step that uses the estimated normals for depth refinement.

Using the surface normal $N(D) = (N_x, N_y, N_z)^\top \in \mathbb{R}^3$, the normalized lighting basis $L = \hat{L} / \|\hat{L}\|$, and the spherical harmonics lighting model [26], the shading can be represented as

$$\begin{aligned} S(N,L) = L_0 + L_1 N_y + L_2 N_z + L_3 N_x \\ + L_4 N_x N_y + L_5 N_y N_z \\ + L_6(N_z^2 - 1/3) \\ + L_7 N_x N_z + L_8(N_x^2 - N_y^2) \end{aligned} \tag{5}$$

It is interesting to note that the image rendering function 3 has an ambiguity in separating reflectance R and illumination L [21].

For the purpose of the depth refinement, estimating relative reflectances and illuminations is sufficient. We therefore fix the norm of illumination to a length of 1. Then we estimate the distribution of illumination and the relative reflectances among pixels. Additionally, we assume white-illumination, *i.e*. color variations on RGB images arise from reflectance variation only. This implies that RGB channels of illumination follow the same distribution and therefore the parameters for illumination can be reduced from 27 to 9. This assumption significantly reduces the computational cost since equation 5 has to be applied on all the pixels in the image.

3.2. Constraint term on reflectance

We assume that neighboring pixels have locally similar reflectances to include the smoothness constraint $P_R(R)$ on reflectance in the cost function 1. This assumption holds when the refletances have a uniform albedo. However, if the scene includes objects with complicated textures, this assumption contradicts at texture borders and hence the refinement process will be deteriorated. To solve this, we design the local similarity term that is adaptively weighted based on local changes in reflectance:

$$P_R(R) = w_r \sum_i^n \sum_{j \in p(i)} \|w_{ch}(i,j) w_{it}(i,j)(R_i - R_j)\|^2 \tag{6}$$

where $p(i)$ denotes a 3×3 window (8-adjacent pixels).

In detail, we first design the weight w_{ch} that adaptively controls the cost depending on changes in chromaticity,

$$w_{ch}(i,j) = \exp(-k_{ch}(1 - C(i,j))) \tag{7}$$

$$C(i,j) = \frac{I_i \cdot I_j}{\|I_i\| \|I_j\|} \tag{8}$$

where k_{ch} is a parameter for controlling the tolerance of chromaticity changes. In our rendering function 3, the direction of RGB vector $I_i/\|I_i\|$ is identical to the direction of reflectance R because the shading S is just a coefficient multiplied to the vector in this function. Therefore, the chromaticity changes computed as angles between RGB vectors reflect that of reflectances. We compute the angles between normalized RGB vectors of neighboring pixels using the same method as [17].

We next design the weight w_{it} that can adaptively control the cost depending on the scale changes in chromaticity. For detecting the scale changes in chromaticities.

$$w_{it}(i,j) = \exp(-k_{it}\|I_i - I_j\|^2) \tag{9}$$

where k_{it} is a parameter for controlling the tolerance of intensity changes. Since our shading function uses a smoothly varying illumination model, the sharp changes in reflectances (and/or depths) will appear as the sharp changes in image intensities [12]. We therefore detect these sharp changes using the distances between RGB vectors in neighboring pixels.

Note that we setup the controlling parameters such that k_{ch} is relatively small for high sensitivity to chromaticity changes whereas k_{it} is big to avoid over detection.

3.3. Constraint terms on depth

For the depth D, we apply two types of constraint terms: local surface smoothness and bas-relief.

$$P_D(D) = w_s P_{smooth}(D) + P_{bas}(D) \tag{10}$$

$P_{smooth}(D)$ is a term for constraining smoothness of local surface (depth). This term is based on the assumption that pixel intensities become similar if the reflectances and surface normals are similar in the Lambertian surface model. P_{bas} is a term for handling bas-relief ambiguity. This term for constraining bas-relief ambiguity is required because, when only the reflectances are given, the surface normal cannot be determined since there are multiple combinations of shapes and illuminations that generates the same image. We next describe these two terms in detail.

Local surface smoothness. We construct the smoothness term using surface normal continuities based on pixel intensities as

$$P_{smooth}(D) = \sum_{i}^{n} \sum_{j,k \in l(i)} \{w_{sc}(i,j,k)\|N_j(D) - N_k(D)\|^2\}^2 \tag{11}$$

where $l(i)$ indicates four lines (horizontal, vertical, two diagonals) passing through a position i. We compute the weight w_{sc} using the differences in neighboring pixel intensities to suppress the errors induced by object boundaries and self-occlusions.

$$w_{sc}(i,j,k) = \exp(-k_{sc} \cdot \min(\|I_i - I_j\|, \|I_i - I_k\|)). \tag{12}$$

We compute this weight for horizontal, vertical and diagonal directions separately.

Bas-relief ambiguity. We construct the bas-relief term P_{bas} in a similar but simpler way to [2]

$$P_{bas}(D) = w_{rd} \sum_{i \in B} \{F(D_i - \hat{D}_i)\}^2 + w_{bd} \sum_{i \in B^{\complement}} \sum_{j \in n(i)} (D_i - D_j)^2 \tag{13}$$

where B is a label which indicates the pixel having an initial depth measurement $\hat{D}_i$ and $B^{\complement}$ is the pixel with no value. This label can be easily obtained from RGB-D images or SfM+MVS.

The first term handles the pixels having initial depths. We penalize if the distance between the original and estimated depths is more than a threshold τ as

$$F(x) = \begin{cases} 0, & \text{if} |x| \leq \tau \\ |x| - \tau, & \text{otherwise} \end{cases} \tag{14}$$

The second term handles pixels that have no initial depth values. The edge-preserving smoothing in equation 11 may generate areas that are not constrained by any other smoothness terms. We simply constrain that their normals are parallel and are facing the optical axis of the camera as commonly used in MVS even if it is inaccurate.

3.4. Implementation & Optimization

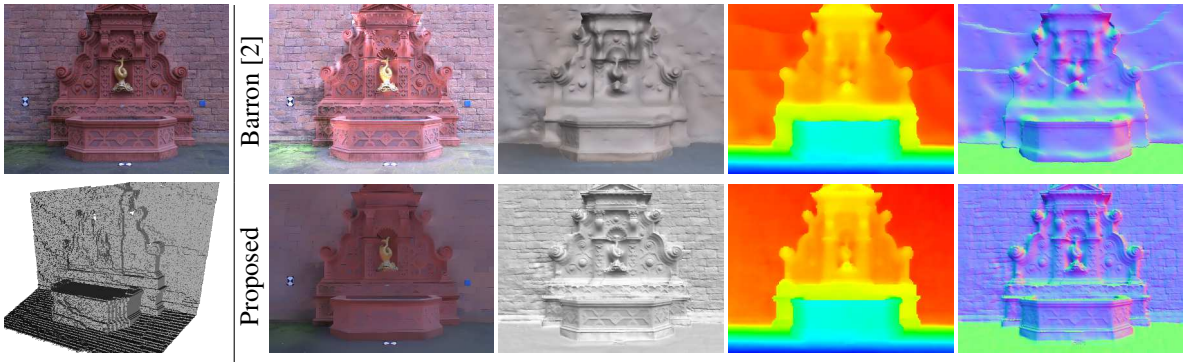
We can estimate geometry, reflectance and illumination by minimizing the cost function (equation 1) which can be solved by a non-linear least squares method, *e.g.* the trust-region method. Before optimization, we upsample low-resolution depth to be the same size as the RGB image using nearest neighbor algorithm. We also add boundary condition $\{R_i \mid 0 \leq R_i \leq 1\}$ for reflectance because reflectance cannot be negative and cannot amplify incident light. We use Ceres Solver [8] to minimize the cost in our experiment.

In our cost function, the total number of variables (R, D and L) is $3n \times n + 9$ for the number of image pixels n. As this is very large, a single step optimization is feasible but not efficient. Thus, we minimize it by iterating a two-step optimization. More specifically, we iterate the following two steps:

(a) optimize R and L for fixed D.

(b) optimize D using the updated R and L.

To further improve stability and efficiency, we use a



(a) Input (b) Reflectance (c) Shading (d) Depth (e) Normal

Figure 2: **Results on the Fountain dataset [30]**. (a) Input RGB image (top) and raw depth (bottom). (b-e) Reflectance, shading, refined depth, and normal computed using [2](top) and our proposed method (bottom). Note that normal (e) is calculated from the depth for visualization.

(a) Barron[2] (b) Proposed (c) Ground-truth

Figure 3: Rendered depths of figure 2 (d) for the Fountain dataset [30].

Data #	0	1	2	3	4	5	6	7	8	9	average
Input	0.678	0.615	0.500	0.251	0.202	0.215	0.215	0.231	0.300	0.348	0.362
Barron [2]	**0.143**	0.136	0.129	0.101	0.104	0.101	0.105	0.231	**0.114**	0.133	0.129
Our	0.167	**0.114**	**0.104**	**0.087**	**0.081**	**0.083**	**0.085**	**0.100**	0.117	**0.119**	**0.108**

Table 1: Depth RMSE values for the Fountain dataset [30].

standard coarse-to-fine strategy starting from $1/8$ down-sampled RGB and depth, then iteratively increase resolution until it is the same as the original resolution.

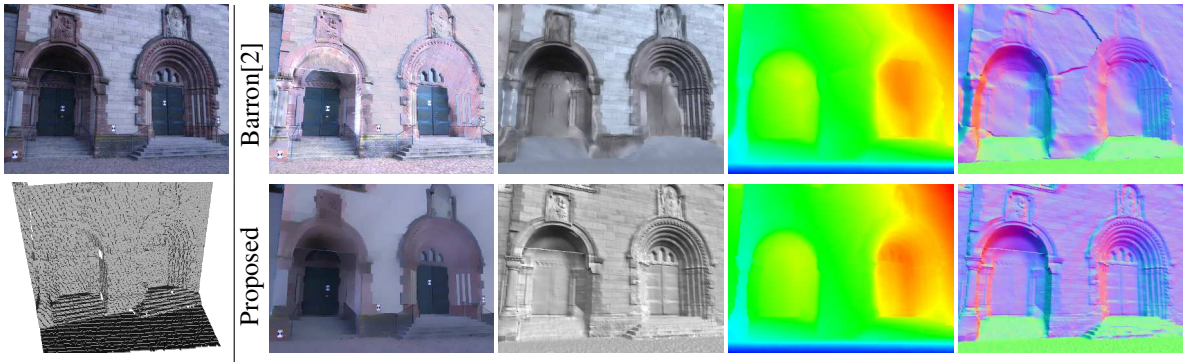
4. Experimental Results

In this section we describe the experiments performed on common datasets and outline the quantitative and qualitative results of our method compared to the state-of-the-art methods [2, 10]. All of experiment were performed on Windows PC with Intel i7 CPU and 32GB memory.

Parameter setup. Throughout all experiments, we used the same values for the parameters $k_{it} = 100$, $k_{ch} = 1000$, $k_{sc} = 100$, $w_{df} = 1$, $w_r = 30$, $w_{bd} = 30$ and $w_{rd} = 100$. Only the depth smoothness weight w_s and the depth error threshold τ are changed for each dataset.

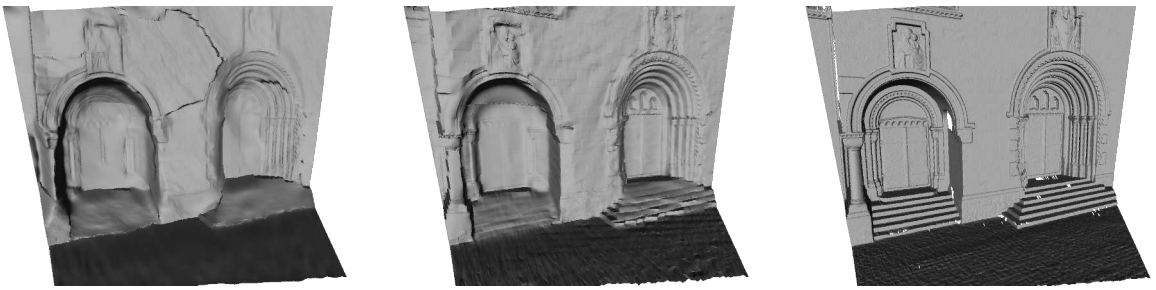
Strecha dataset. We performed experiment on the Strecha dataset [30] which is a standard dataset for 3D reconstruction benchmarks for qualitative and quantitative evaluation of our method. It contains high-quality RGB images, their camera parameters and (laser scanned) ground-truth mesh data.

We first generate ground-truth depth images from the mesh data (Fountain-P11 and Herz-Jesu-P8). We next prepare the input low-quality depth images by down-sampling to 160×120 pixels, adding noise, and quantizing the depth values to acquire a Kinect-like effect similar to [2]. We also down-sampled RGB images to 640×480 pixels. Finally, we



(a) Input (b) Reflectance (c) Shading (d) Depth (e) Normal

Figure 4: Results on the Herz-Jesu dataset[30]. See the caption of figure 2 for details.



(a) Barron[2] (b) Proposed (c) Ground-truth

Figure 5: Rendered depths of figure 4 (d) for the Herz-Jesu dataset [30].

Data #	0	1	2	3	4	5	6	7	average
Input	0.551	0.435	0.490	0.541	0.440	0.514	0.684	0.666	0.540
Barron [2]	0.206	0.189	0.154	0.138	0.127	0.162	0.198	0.184	0.170
Our	**0.172**	**0.154**	**0.127**	**0.117**	**0.106**	**0.120**	**0.141**	**0.147**	**0.137**

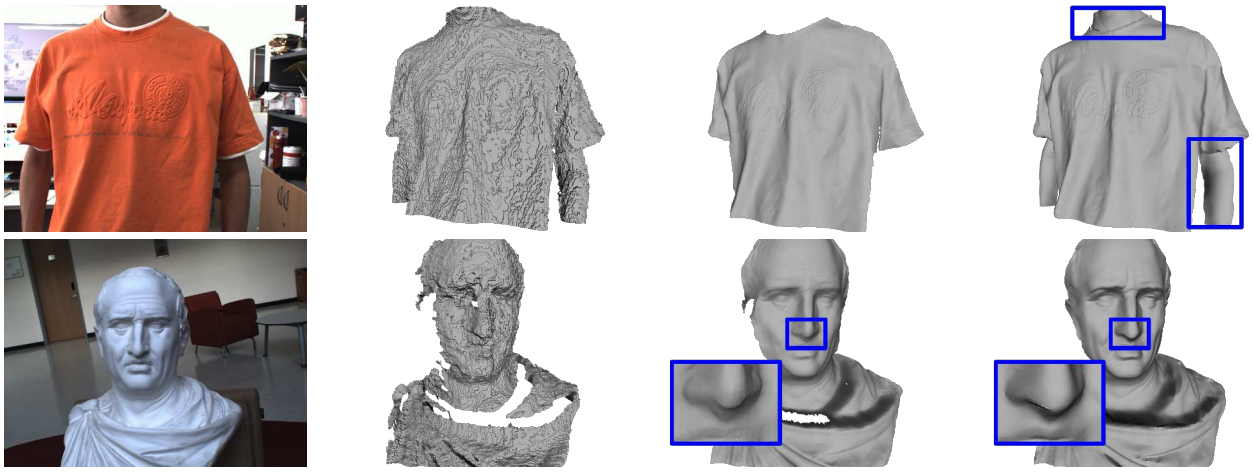
Table 2: Depth RMSE values for the Herz-Jesu dataset [30].

refine the depth using the RGB image with the parameters $w_s = 0.01$ and $\tau = 0.05$. Note that this real-world dataset includes multiple reflectances and complex lighting conditions. We evaluate the root mean squared error (RMSE) between our refined depth and the ground-truth depth.

Figures 2, 3, 4 and 5 show examples of input RGB image, rendered input depth and estimated results(Barron [2] (top) and our method (bottom)). The depth results of [2] have over-smoothing artifacts (figures 3 and 5 (a)) and cracks (figure 5 (a)). These cracks are due to the pre-segmentation failure. In contrast, our method can produce high-quality depths (figures 3 and 3 (b)). Additionally, our edge-preserved smoothness term works effectively on discontinuous surfaces (near the pillar of figure 5 (b)). Furthermore, the quantitative evaluations in tables 1 and 2 clearly show that our method produces more accurate depths when compared using the ground-truth depth.

RGB-D sensor dataset in [10]. In this experiment, we use the images of RGB-D sensors provided in [10]. We set the parameters $w_s = 0.05$ and $\tau = 0.01$.

Figure 6 captures a human wearing a T-shirt (top) and a Cicero sculpture (bottom). As our method represents shadows and a spatial variation of illumination as diffuse reflectance variation, it can be applied to more general object which has complicated reflectance under natural illumination without explicit segmentation as used in [10]. Notice that our method recovered not only the T-shirt but also arms and a neck on the body data (top in figure 6 (d)). Also, notice that the shape around the nose of sculpture is inaccurately recovered by [10] due to shadow whereas our method can deal with it by considering the shadow as variations in reflectance.



(a) Input RGB (b) Input depth (c) Han *et al*. [10] (d) Proposed

Figure 6: Results on the RGB-D dataset in [10]. (a) Input RGB images. (b) Input raw depths. (c) Depths refined by [10]. (d) Depths refined by our method.

NYU dataset. We conducted experiments on the NYU V2 dataset [23] commonly used in the intrinsic image decomposition evaluation. The dataset contains RGB-D images taken by Kinect V1 sensors. We tested the same 33 scenes used in [2]. We used $w_s = 0.04$ and $\tau = 0.02$ for all scenes in the NYU dataset.

Figure 7 shows the results obtained by [2] (top) and our method (bottom) on the NYU dataset. Scene numbers indicate the image indexes in the dataset. Notice that our method successfully recovered more details when compared to the over-smoothed results of [2]. The towel and tiles on the wall look more detailed in scene 195 of figure 7. Also wrinkles on the curtain and the bedclothes look more natural in scene 541 of figure 7.

However, our method has some limitations that causes it to induce noise on some refined shapes. Our data fidelity term cannot absorb errors induced by extreme illumination variations. Scene 190 of figure 7 presents the effect of spatially varying illumination. The wall of our result is unnaturally distorted near the table-light in contrast to [2]. Our constraint term on reflectance may not distinguish monochromatic variations. Scene 518 of figure 7 shows the artifact of monochromatic reflectance that appears as surface unevenness on the curtain, pillow and bedclothes.

5. Conclusion

We have proposed a method for the joint estimation of depth, reflectance and illumination for a depth refinement. The proposed method is able to produce high quality depth from a single RGB-D image with no additional pre/post processing steps. The technical novelty is in the design of a cost function with adaptively weighted smoothness terms for depth, reflectance and illumination. The experimental results on real-world datasets demonstrated that our method can refine the shapes accurately. The comparisons with state-of-the-art methods clearly showed the advantages qualitatively and quantitatively.

Our method has some limitations. It cannot handle the scenes which are not suitable to our rendering model: Non-Lambertian surfaces and complex illuminations (interreflection, refraction and spatially varying illumination). In addition to that the weighted local similarity used in our reflectance constraint cannot distinguish whether the borders arise from texture changes or object occlusions. Resolving these limitations using multiple RGB images is a potential future work.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 15H05313, 25240025.

References

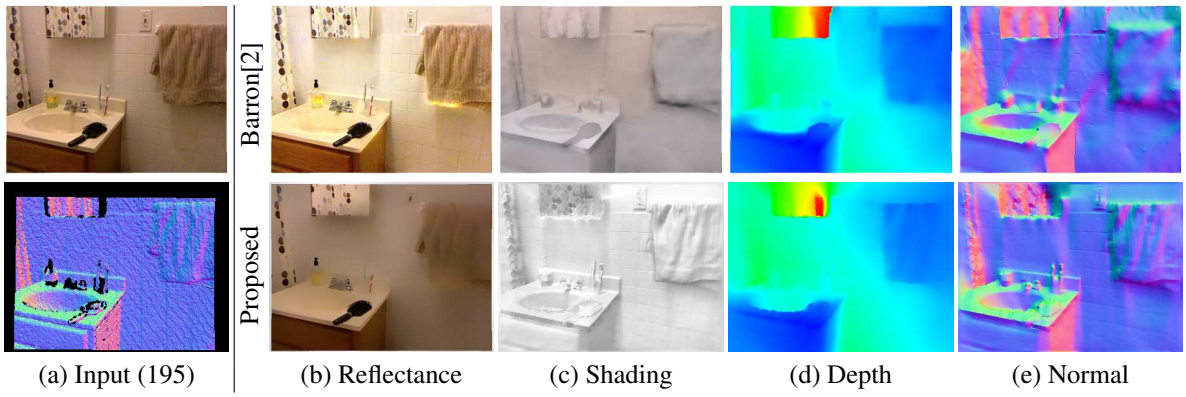
[1] ASUS. Xtion pro live. `http://www.asus.com/Commercial_3D_Sensor/Xtion_PRO_LIVE/`.

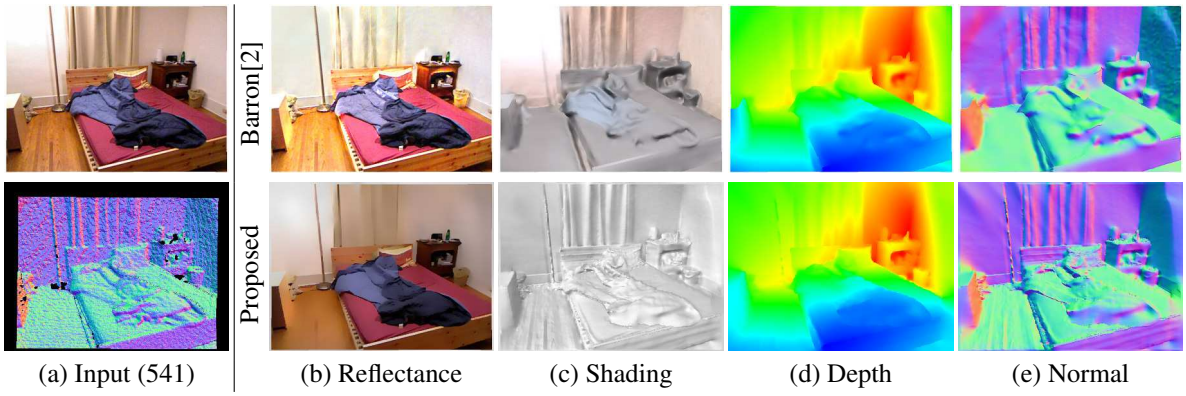
[2] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013.

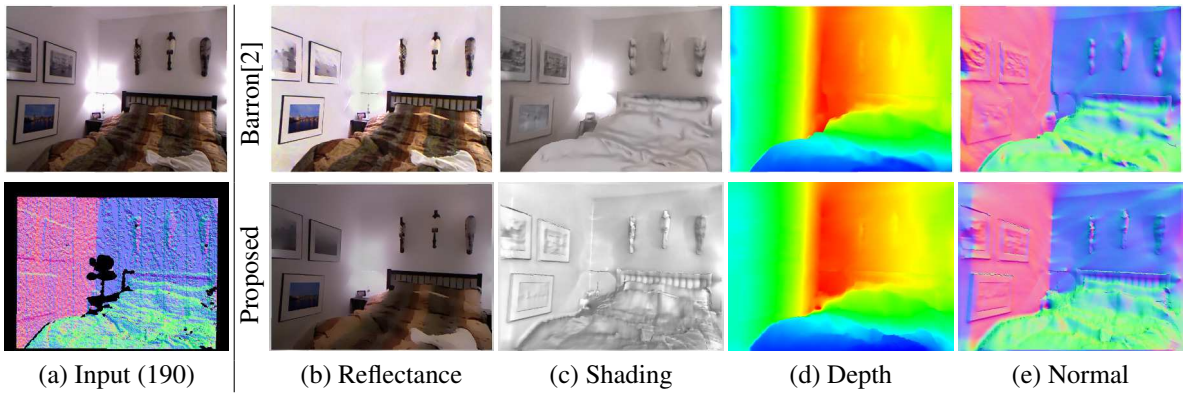
[3] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *CVPR*, 2001.

[4] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. In *SIGGRAPH*, 2014.

[5] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013.







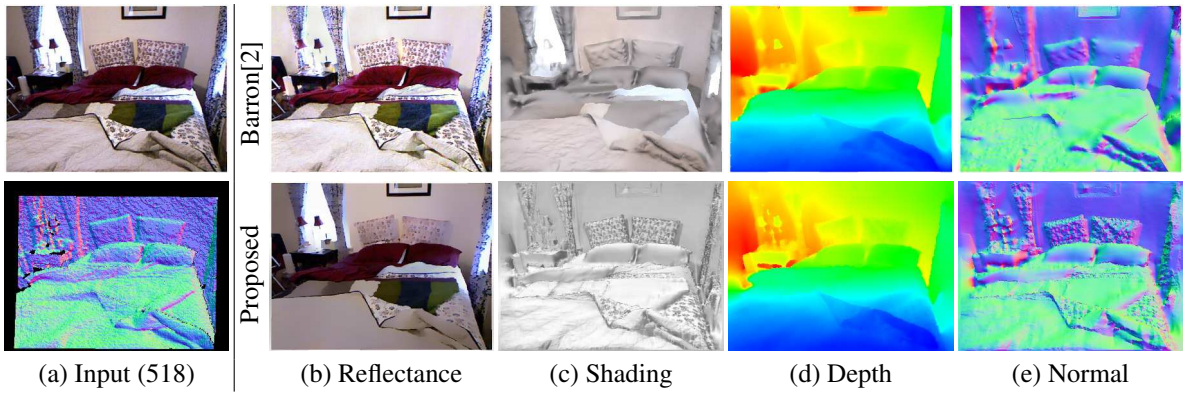


Figure 7: Results on the NYU dataset [23]. See the caption of figure 2 for details. The bottom of (a) shows input depth normal instead of the actual depth. The number in (a) is the scene index.

[6] D. A. Forsyth. Variable-source shading analysis. *IJCV*, 91(3):280–302, 2011.

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *CVPR*, 2007.

[8] Google. Ceres solver. `http://ceres-solver.org/`.

[9] T. Haber, C. Fuchs, P. Bekaer, H.-P. Seidel, M. Goesele, and H. Lensch. Relighting objects from image collections. In *CVPR*, 2009.

[10] Y. Han, J.-Y. Lee, and I. S. Kweon. High quality shape from a single rgb-d image under uncalibrated natural illumination. In *ICCV*, 2013.

[11] C. Hernandez, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *PAMI*, 30(3):548–554, March 2008.

[12] B. K. P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 3(4):277 – 299, 1974.

[13] B. K. P. Horn. Obtaining shape from shading information. In *Shape from Shading*, pages 123–171. MIT Press, Cambridge, MA, USA, 1989.

[14] K. Ikeuchi and B. K. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17(13):141 – 184, 1981.

[15] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 559–568, New York, NY, USA, 2011. ACM.

[16] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, 2011.

[17] J. Jeon, S. Cho, X. Tong, and S. Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, 2014.

[18] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *CVPR*, 2011.

[19] B. Klingner, D. Martin, and J. Roseborough. Street view motion-from-structure-from-motion. In *ICCV*, 2013.

[20] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. Coherent intrinsic images from photo collections. In *SIGGRAPH Asia*, 2012.

[21] E. H. Land, John, and J. Mccann. Lightness and retinex theory. *Journal of the Optical Society of America*, pages 1–11, 1971.

[22] 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*, 2012.

[23] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[24] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *SIGGRAPH*, 2005.

[25] J. Park, S. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon. Multiview photometric stereo using planar mesh parameterization. In *ICCV*, 2013.

[26] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, 2001.

[27] B. Shi, K. Inose, Y. Matsushita, P. Tan, S.-K. Yeung, and K. Ikeuchi. Photometric stereo using internet images. In *3DV*, 2014.

[28] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[29] N. Snavely, S. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008.

[30] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

[31] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):121–171, 1980.

[32] C. Wu, M. Zollhfer, M. Niener, M. Stamminger, S. Izadi, and C. Theobalt. Real-time shading-based refinement for consumer depth cameras. In *SIGGRAPH Asia*, 2014.

[33] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin. Shading-based shape refinement of rgb-d images. In *CVPR*, 2013.

[34] Y. Yu and J. Malik. Recovering photometric properties of architectural scenes from photographs. In *SIGGRAPH*, 1998.

[35] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, 2012.

[36] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *PAMI*, 34(7):1437–1444, July 2012.