High Quality Shape from a RGB-D Camera using Photometric Stereo

Songyou Peng

Supervised by:

Dr. Yvain Quéau Prof. Daniel Cremers

Computer Vision Group

Department of Computer Science

Technical University of Munich

A Thesis Submitted for the Degree of
MSc Erasmus Mundus in Vision and Robotics (VIBOT)

· 2017 ·

Abstract

Low-cost RGB-D cameras are playing an increasingly important role in many computer vision tasks, however, their captured depth maps not only do not contain fine details of the objects but also have noisy or missing information. This dissertation proposes two novel methods which can refine the rough depth images based on the theory of photometric stereo.

The first method called RGB ratio model can resolve the nonlinearity problem in most previous methods and promise a closed-form solution. Modern depth refinement approaches usually could not separate the shape from the complicated albedo, which leads to visible artefacts on the refined depth. We propose another robust multi-light method which offers the advantage of recovering the real shape from the imperfect depth without any regularization, and it outperforms the state-of-the-art methods. Moreover, we combine our approach with image super-resolution such that the high-quality and high-resolution depth can be acquired. Quantitative and qualitative experiments have demonstrated the robustness and effectiveness of the suggested methods.

So you have to trust that the dots will somehow connect in your future. You have to trust in something — your gut, destiny, life, karma, whatever. This approach has never let me down, and it has made all the difference in my life.

— Steve Jobs

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would love to thank my supervisor Prof. Dr. Daniel Cremers for offering me this master thesis opportunity in the Computer Vision Group. It is a great honour to be a member of your group. This thesis would not have been possible without the continuous support from my advisor Dr. Yvain Quéau, who always believed me and gave my so many constructive suggestions. Thank you for your patience, effort and trust.

I really appreciate the enormous help offered by one my best friends Tianming Qiu. My stay in Munich would have been much harder without you. My gratitude also goes to Nan Yang, Hidenobu Matsuki, Yuesong Shen and all other masters students sharing the office 02.09.053 for your warm company and motivation. Besides, I am really grateful to Cancen Jiang and David Kong for your careful proofreading and valuable advice.

I am much obliged to Prof. Peter Sturm, who supervised my summer internship in INRIA during the summer of 2016 and recommended me here for the thesis. Your rigorous research attitude and extraordinary personality have set a great example to me in all respects.

I am particularly thankful to my VIBOT colleagues who made these wonderful two years one of the most special and satisfactory memories in my life. All the fun times, all the places we have visited, and the projects, assignments and exams we have gone through together will remain dear to me.

Most important of all, my dearest thanks goes to my beloved family. You have been unconditionally supporting of my decisions and have given me everything. I love you so much.

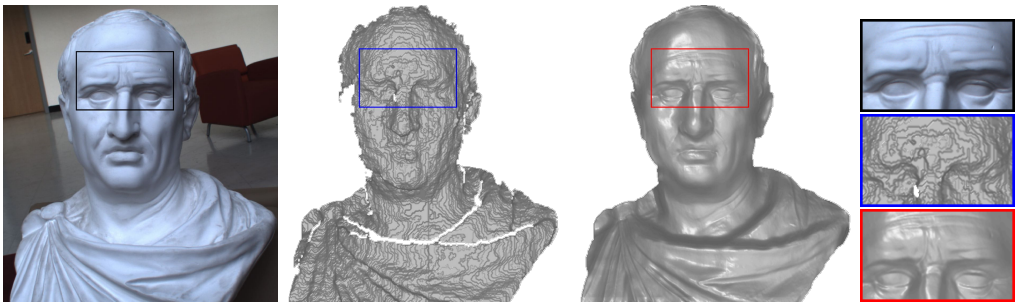
Songyou, well done, I am proud of you.

Chapter 1

Introduction

1.1 Motivation

With the advent of affordable RGB-D cameras, many research areas in modern computer vision, computer graphics and robotics have been boosted significantly, such as 3D modeling [1] and reconstruction [2], human motion capture [3] and visual SLAM [4,5], etc. Although the low-cost commercial RGB-D sensors provide the RGB image and its corresponding depth information, the quality of depth is often not satisfactory.



(a) RGB image (b) Input depth (c) Our refined depth (d) Details

Figure 1.1: Illustrations for the task of depth refinement. The depths are plotted as a 3D surface for the sake of a better visualization. The input RGB and depth images are from [6]. We can clearly see the improvement in the depth details using our RGBD-Fusion Like method.

As we can notice from Fig. 1.1(b), the depth map is very noisy and contains quantization effect with missing depth values. Moreover, we could not see much details of the statue from the rough depth. In contrast, Fig. 1.1(c) shows the depth with much higher quality after the

refinement using our approach which contributes to the preservation of geometric details. To highlight, Fig. 1.1(d) illustrates the significant improvement in the regions of the forehead where the details of wrinkles and eyes are recovered from the raw depth.

As a concrete example, it has been shown in [7] that the quality of 3D reconstruction may suffer from the noisy and imperfect depth measurements, and the estimation of camera trajectory will also drift severely because of the accumulated errors from the rough depths. There are some methods, e.g. KinectFusion [8], which tried to recover these details by fusing all the depth data from multiple views, nevertheless, the recovered details are still very limited. Provided we can enhance the quality of input rough depth in Fig. 1.1(b) to the refined depth in Fig. 1.1(c), all the tasks which rely on RGB-D sensors can be further improved.

1.2 Problem Statement

In this dissertation, we delve into the research of the refinement for a single depth map. To refine the depth information from consumer depth sensors, the intrinsic details of an object should be explored with the aid of RGB image(s), such as the the positions of the scene illuminations, their corresponding influences on the object's shading, and also the reflectance (albedo) of the object's material. Shape from shading (SFS) and photometric stereo (PS) are the fundamental approaches which can study the intrinsic properties of images, including layout patterns and geometric details. Therefore, they have been successfully integrated into many state-of-the-art depth enhancement methods.

These shape refinement approaches based on SFS or PS are usually formulated with the Lambertian reflectance model. Refining the depth from this model is an inverse problem, in which nonlinearity exists. Some methods [9, 10] directly applied the nonlinear optimization algorithm like Levenberg Marquardt or the alternating direction method of multipliers (ADMM) to solve the problem which led to runtime issue. Other methods [11] froze the nonlinear part with the outcome from the last iteration to cancel out the nonlinearity. However, such fixed-point scheme sometimes causes the optimization to diverge. As a consequence, we propose a method called the RGB ratio model. With the red, green and blue LED lights set up in various directions, we are able to build a lighting model for each channel of the color image acquired from RGB-D sensors. The proposed setup and model can resolve the nonlinearity and provide a closed-form solution. It will be shown in chapter 4 that the RGB ratio model generally achieves similar accuracy to the state-of-the-art methods and has better performance in some cases.

Furthermore, it should be mentioned that many shading-based depth refinement methods require making the assumption that the albedo is either uniform or constant [6, 12–15], and others impose some piecewise smoothness constraints on the albedo to make this inverse problem

well-posed [9–11, 16]. These methods work well on some uncomplicated objects, for example, the statue in Fig. 1.1, but have the difficulty in separating the changes in the albedo from the shape when the albedo becomes complicated. This makes sense because numerous real-world objects have rather elaborate patterns and colors. So the imposed piecewise smoothness is not commonly realistic. Consequently, we present another novel approach called the robust multi-light method which can handle the case of very complicated albedos. The idea is to acquire several images from various illuminations with fixed camera view, and use all the information to jointly refine light, albedo and depth iteratively. To simulate the scenario of multiple lighting conditions, we sway a white LED light source (could be the phone flashlight) and consistently capture images. Compared to other methods, one main advantage of the proposed method is that no regularization term needs to be imposed, which saves much unpredictable time for the tedious process of parameter tuning.

Last but not least, since the depth maps acquired from consumer RGB-D sensors usually do not have the satisfactory resolution, we used the higher-resolution RGB images as cues and have managed to integrate super-resolution with our depth refinement method. In this way, the final refined depth can attain the same resolution as the large RGB images with all the fine details. We believe this is the first shading-based depth super-resolution approach.

The main contributions of this dissertation are:

1. We propose a new RGB ratio model to resolve the nonlinearity and achieve similar accuracy to the state-of-the-art methods.

2. We introduce a robust multi-light method which outperforms other depth refinement approaches both quantitatively and qualitatively. Moreover, no regularization is imposed.

3. We combine the image super-resolution with our method and present the high-quality and high-resolution depth.

1.3 Outline

The outline of the thesis is organized as follows. In Chapter 2, we introduce the RGB-D cameras, related knowledge about shape from shading and photometric stereo as well as the state-of-the-art depth refinement approaches. Chapter 3 describes the methods developed in this thesis, including a modified version of RGBD-Fusion and two proposed methods. Extensive quantitative and qualitative evaluations of the performance of our methods are then presented in Chapter 4. Finally, Chapter 5 summarizes the developed approaches and provides the possible extensions for the future research.

Chapter 2

Background

2.1 RGB-D Cameras

2.1.1 General

RGB-D cameras have been widely used in many modern computer vision areas, for example 3D reconstruction [1, 2], visual odometry and mapping on quadrocopter [17] and visual SLAM algorithms [4, 5]. A RGB-D camera returns a color image which is usually in RGB color space, and a depth map, every pixel of which reflects the real-world distance between the camera and the corresponding position of the pixel. Depending on the technologies used to measure the depth information, the RGB-D camera can be divided to passive and active [18].

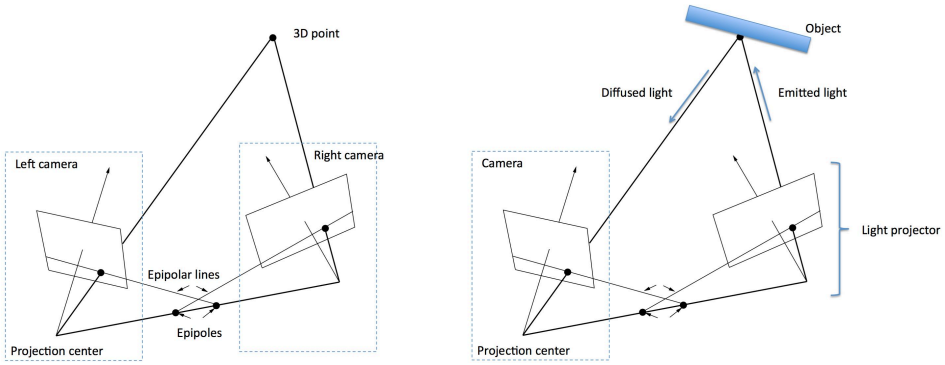
The so-called passive RGBD-camera usually contains two RGB cameras with a known translation between them. After taking one picture for each, the features in the two pictures are matched and then triangulation is applied to obtain the depth. An illustration is shown in Fig. 2.1(a).

Active technologies usually emit light into the environment so that it has the capability of acquiring depth images in a totally dark indoor scenario. They can be furthered categorized as the time of flight (ToF) or structured light approaches.

A ToF camera calculates the depth in each pixel by measuring the delay between the emission and the reflected time. ToF cameras typically emit either pulsed light or modulated light. The Microsoft Kinect 2.0 and IFM Efector are two examples of the ToF camera.

The RGB-D cameras with the structured light use a projector for a known pattern. Since the transformation between the camera and the projector is pre-given, a camera observes the projected pattern and then triangulates to calculate the depth (Fig. 2.1(b)). The ASUS Xtion Pro Live, Intel RealSense R200 and Ensenso are several well-known cameras using structured

light. It should be noted that many active stereo cameras project infrared (IR) light. Due to the fact that the sun is a source of infrared light, the use of these cameras is limited to the indoor environment.



(a) Passive stereo (b) Active stereo

Figure 2.1: Illustrations for the principle of passive and active stereo. Image courtesy of [19].

2.1.2 ASUS Xtion PRO LIVE

The ASUS Xtion Pro Live camera has two cameras and an IR projector as shown in Fig. 2.2(a). According to the official website [20], it provides the RGB image with the maximum resolution of 1280×1024, while the depth can be alternated either VGA resolution (640×480 with 30fps) or QVGA (320×240 with 60fps). Its depth is reported to range from 0.8 to 3.5m, but we found in the experiments that the blind area was less and merely around 0.5m. Xtion Pro Live has been used throughout our experiments and we chose different configurations depending on the applications. When the depth super-resolution is required, the RGB images have the resolution of 1280×1024, otherwise, we keep the same RGB resolution as the depth (640×480).

2.2 Shape from Shading & Photometric Stereo

The well-known shape from shading (SFS) problem was first introduced by Horn [21] in 1970 and then a large amount of literature flooded in to develop the field. The idea of SFS is, knowing the light source position, one can estimate the shape or the surface of an object from one single grayscale image. This inverse problem is highly ill-posed, as illustrated in Fig. 2.3.

From the mathematical perspective, the luminance can be separated as follows:

$$I = \rho S \tag{2.1}$$

(a) Camera structure [20]

(b) RGB image (c) Depth image

Figure 2.2: Upper row: the structure of ASUS Xtion Pro Live. Lower row: the RGB and depth images of an indoor scene acquired by this RGB-D camera. The unit in image (c) is millimetre.

where I is an intensity image, ρ is the reflectance (albedo) of the surface, and the S is the shading image. An example of such an image decomposition is shown in Fig. 2.4.

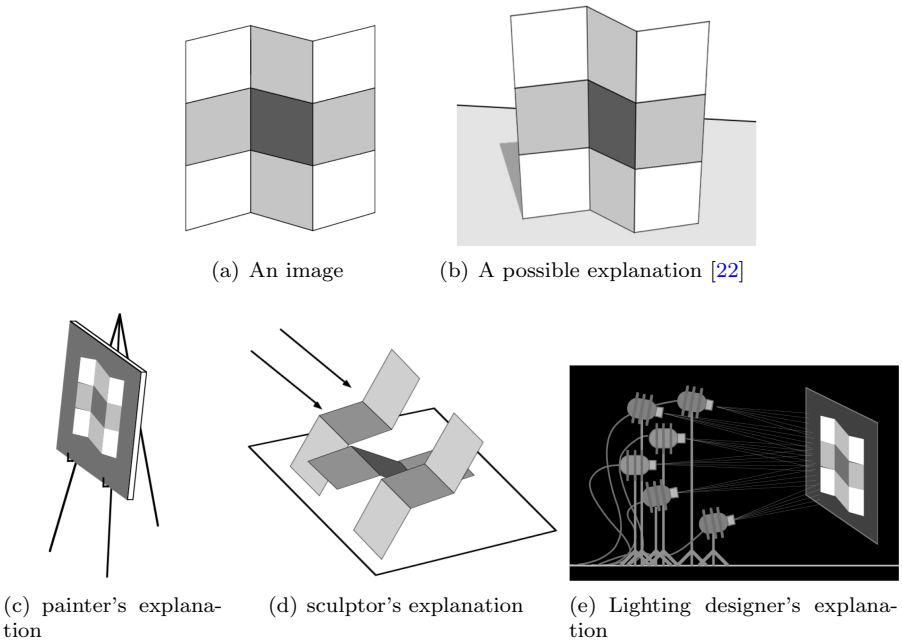
SFS approches assume the observed object follows the Lambert's cosine law [25], based on which the Eq. 2.1 can be reformulated to the Lambertian reflectance model:

$$I = \rho \, \mathbf{l}^\top \mathbf{n} \tag{2.2}$$

where we can notice that the shading S is the inner product of the light direction and the surface normal. Thus, the task of SFS is to retrieve the shape (surface normal) from the shading based on the Lambertian reflectance model. Moreover, many state-of-the-art shape or depth refinement methods used an extension of Lambertian model called spherical harmonics (SH) [26, 27] which can represent the illumination more realistically. It has been shown that the first-order SH model (Eq. 2.3) can account for 87.5% of real world light so we applied it throughout the whole thesis:

$$I = \rho \, (\mathbf{l}^\top \mathbf{n} + \varphi) \tag{2.3}$$

where φ can be understood as the ambient light parameter.



(a) An image (b) A possible explanation [22]

(c) painter's explanation (d) sculptor's explanation (e) Lighting designer's explanation

Figure 2.3: Various explanations for a twice-bent surface. Illustrations for the ambiguity suffered by SFS. Images courtesy of [23].

Figure 2.4: The decomposition of a color image of a toy panther. ρ is the reflectance (albedo) and S is the shading. Images are from MIT intrinsic images dataset [24].

The goal of SFS is to estimate the shape or the surface of an object, which is represented by the surface normal $\mathbf{n}$. Two camera projection models are usually used for modelling the surface normal: orthographic and perspective projection. To derive $\mathbf{n}$ using orthographic projection, we consider projection along the z-axis [28]. Hence, A 3D point $P = (x, y, z)$ is mapped to the image point $p = (x, y)$. It should be noted that the depth, z, depends on the x and y coordinates. And we know the surface normal is orthogonal to the tangent plane in (x, y),

which can be written as:

$$\mathbf{n}(P) \propto \partial_x P \times \partial_y P = \begin{pmatrix} 1 \\ 0 \\ z_x \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ z_y \end{pmatrix} = - \begin{pmatrix} z_x \\ z_y \\ -1 \end{pmatrix} \tag{2.4}$$

After normalizing and choosing the outward direction, we acquire the unit-length surface normal with orthographic projection:

$$\mathbf{n}_{ortho} = \frac{1}{\sqrt{|\nabla z| + 1}} \begin{pmatrix} \nabla z \\ -1 \end{pmatrix} \tag{2.5}$$

For the more realistic perspective model, the 3D point P now becomes $\begin{pmatrix} (x - x_0)/f \\ (y - y_0)/f \\ z \end{pmatrix}$ and the corresponding normal is:

$$\mathbf{n}_{perspect} = \frac{1}{d} \begin{pmatrix} f\tilde{z}_x \\ f\tilde{z}_y \\ -1 - (x - x_0)\tilde{z}_x - (y - y_0)\tilde{z}_y \end{pmatrix} \tag{2.6}$$

where $\tilde{z} = \log z$, f the focal length, (x_0, y_0) the coordinates of principle points, and the normalizer $d = \sqrt{(f\tilde{z}_x)^2 + (f\tilde{z}_y)^2 + (-1 - (x - x_0)\tilde{z}_x - (y - y_0)\tilde{z}_y)^2}$.

From the definition of the surface normal, we can notice the SFS is an ambiguous problem. Even when the lighting and the albedo are known in Eq. 2.2, the inverse problem is ill-posed because the normal has 2 degrees of freedom. As we can see from Fig. 2.3, the solution of SFS is ambiguous. Horn and Brooks [29] proposed the so-called integrability constraint $z_{xy} = z_{yx}$, which was the first constraint imposed on the surface normal to make the SFS problem well-posed. Frankot and Chellapa [30] projected a non-integrable surface to the subspace spanning the valid smooth surface.

Provided we have several images from the same view but with different illuminations, Eq. 2.1 can be modelled as:

$$\mathbf{I} = \mathbf{BL} \tag{2.7}$$

Assuming there are $n \geqslant 3$ images from various illumination conditions, with m pixels in each image, $\mathbf{I} \in \mathbb{R}^{m \times n}$ is the stack of all the intensity images and each column of $\mathbf{I}$ represents a vectorized image from a lighting condition. $\mathbf{B} \in \mathbb{R}^{m \times 3}$ corresponds to $\rho \cdot \mathbf{n}^\top$ and $\mathbf{L} \in \mathbb{R}^{3 \times n}$ represents n various lightings. If the illuminations are known, we call this problem *calibrated* photometric stereo (PS), which was first introduced by Woodham [31] in 1980. The problem

is over-constrained so the surface normals can be estimated using a simple least squares. Some regions may sometimes suffer from the shadows for a certain illumination, so Forsythe and Ponce [32] formed a diagonal matrix to eliminate all those shadow points. Another interesting example of calibrated PS was proposed by Hernández *et al.* [33]. They controlled red, green, and blue lights from three directions and acquired the shape from only one color image, every channel of which was treated as a separate intensity image. This was one of the inspirations of our first proposed method RGB ratio model, which will be detailed in chapter 3.

However, the different lightings are not always controlled or given, then we call this kind of problem *uncalibrated* photometric stereo. Hayakawa [34] found that the surface normals and the albedo could be recovered with a 3×3 linear transformation using singular value decomposition. Yuille and Snow [35] further reduced the ambiguity to a generalized bas-relief (GBR) ambiguity by adding the integrability constraint. Now Eq. 2.7 can be represented as

$$\mathbf{I} = \mathbf{B}\mathbf{A}^{-1}\mathbf{A}\mathbf{L} \tag{2.8}$$

where $\mathbf{A}$ is the GBR matrix:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \mu & \nu & \lambda \end{pmatrix} \tag{2.9}$$

There has been a large amount of research work trying to solve the GBR ambiguity in uncalibrated PS. Alldrin *et al.* [36] used the prior emphasizing that the albedo distribution should have a low entropy, to resolve the ambiguity. Favaro and Papadhimitri [37] eliminated the ambiguity by exploiting the spatial maximum points of the inner product between the normals and the lights. They also found out the solution is unique when the normal is constructed under the perspective projection instead of the orthographic projection [38]. Quéau *et al.* [39] estimated the GBR parameters by imposing the total variation norm.

In terms of depth, the GBR ambiguity is equivalent to the equation [40]:

$$z(x, y) = \lambda z(x, y) + \mu x + \nu y \tag{2.10}$$

Based on the equation, we find out that the GBR ambiguity still exists for the task of depth estimation. However, since the rough depth will be given as the input from a RGB-D camera, the GBR ambiguity problem encountered by uncalibrated PS has resolved automatically. Therefore, we will discuss some state-of-the-art depth refinement approaches.

2.3 Depth and Shape Refinement

In recent years, there is a large amount of literature focusing on the depth or shape refinement based on either SFS (only use one single image) or PS (multiple images with different illuminations). They are collectively called shading-based methods. We will discuss these two streams respectively in the following.

2.3.1 SFS-based methods

Since SFS itself uses only one input image, it suffers from some intrinsic ambiguities that we have mentioned in the last section, even when the light and the albedo are specified, so there will be more than one possible solution. Now, although a rough depth map is given, the illumination and the albedo are unknown, so some regularization terms have to be imposed in order to acquire an exact solution of the inverse problem.

Han *et al.* [6] presented a framework which combines a global lighting model using the given color and depth with the help of SH model, with a local lighting model which varies spatially. The surface orientations should obey the integrability constraint on the smooth surface so they enforced the constraint by penalizing the curl of local neighbors. However, the albedo in their method is assumed to be uniform. To handle the multi-albedo objects, they had to apply another intrinsic image decomposition algorithm [41] and k-mean clustering to group the albedos into some areas with constant values inside. Such framework is not only unrealistic but also very time-consuming so not able to be adapted to the real-world applications.

Yu *et al.* [42] iteratively updated the SH lighting and the relative albedo, with which the method refined the rough shape. They performed mean shift to segment the input RGB image into small regions with uniform albedo, and then obtained the relative albedos among various segmented regions. To fill in the missing depth information, a constrained texture synthesis and patch-based repairing scheme were applied. In contrast, we efficiently apply a basic image inpainting approach as the pre-processing and the results are also satisfactory.

Wu *et al.* [9] extended their previous offline shading-based refinement work [12] to online shape refinement with highly parallel scheme and the help of the GPU. They first calculated the 2nd-order SH parameters with the assumption of uniform albedo, and then estimated the albedo by simply dividing RGB image with the shading term. We will show in chapter 3 that this process may lead to the severe albedo overfitting problem such that the albedo estimation is not correct. The shape is then refined in real-time by finding the surface that minimizes the difference between the shading and intensity image gradients. Thereafter, the coarse depth map was directly refined with smoothness and temporal constraint on the video by using a Gauss-Newton solver on the GPU.

Kim *et al.* [16] used a joint energy to estimate the depth, albedo and the light with smoothness regularization terms. An anisotropic Laplacian constraint on chromaticity was introduced for albedo and a local smoothness and bas-relief ambiguity similar to [43] constraints are imposed for depth. Based on our testing implementation, it has turned out that the Laplacian on chromaticity of the image cannot provide satisfactory albedos for the small indoor environment. What's more, it is a very tedious process to tune all the parameters for the constraints.

RGBD-Fusion method from Or-El *et al.* [11] can also deal with natural illumination conditions and make the depth recovery task in real time under GPU. They imposed the constraints not only on the albedo and depth estimation but also pixel-wise ambient lighting. Their method does not really converge because of their way of handling the nonlinearity. This inspired us to propose a new RGB ratio model to eliminate the nonlinearity.

Or-El's following work [10] can deal with specular objects with the help of IR camera and a more complicated reflectance model than spherical harmonics. In contrast, our proposed multi-light method still uses a Lambertian diffused reflectance model but can handle the specularity.

From what we have discussed, SFS methods are often limited to the uniform or constant albedo. For the sake of handling multi-albedo cases, some SFS methods [6,42] adapted segmentation methods to divide the input image into some constant albedo part, but the real-world objects are usually with complex multi-albedo and small regions, which makes the segmentation not accurate or incorrect. Some other methods [9–11, 16] try to add some piecewise smooth constraints on the albedo but never really acquired satisfactory outcome. Therefore, SFS-based methods using only one single image have the difficulty in separating the albedo from the surface normal, which may lead to the wrong depth estimation.

2.3.2 PS-based methods

Another category of shading-based depth refinement is PS-based methods. With the help of multiple images acquired from various illuminations, these approaches can resolve the ambiguities tolerated by SFS methods and have a better performance in the separation between the albedo and surface normal.

Haque *et al.* [14] proposed a method to reconstruct the shape and refine the depth using an IR camera without the need of RGB camera. However, similar to many other multi-view photometric reconstruction approaches [13,15], they assumed the albedo is restricted to uniform and thus, it is not suitable to use it for multi-albedo objects.

Their follow-up work from Chatterjee and Govindu [44] decomposed the input images under different illuminations with a standard photometric stereo manner. They used an iterative reweighted method to approximate the Rank 3 radiometric brightness matrix, then factorize it into the corresponding lighting, albedo and surface normal. They can cope with the multi-

albedo objects but still have to use the IR images instead of RGB images. In this case, at least one extra infrared light source is always required, while in our case only a cheap LED light or even just the flashlight on a phone is enough. Moreover, since the IR camera in ASUS Xtion Pro Live is limited to the resolution 640×480, while the RGB camera can reach 1280×1024, their approach can not perform depth super-resolution task like our multi-light method.

Wu *et al.* [45] used second-order spherical harmonics to model general illumination and use the shading constraint to help improve the object reconstruction. This method is extended to [12] whose results are furthered improved by integrating a weak temporal prior on lighting, albedo and the shape.

In this thesis, we propose two shape refinement methods based on the uncalibrated photometric stereo. Similar to [33], we firstly used red, green and blue LEDs for the active illuminations so we can treat every channel of the obtained color image as an intensity image with light from a different direction. Another proposed method needs only one white LED. With the RGB-D camera's angle of view fixed, we manually move the LED lights while the images are being taken. Moreover, the shading-based method has never been applied for the depth map super-resolution before. We successfully adapt our methods to depth super-resolution and achieved very pleasing outcomes.

Chapter 3

Methodology

Many computer vision applications such as 3D object reconstruction or visual SLAM require the depth information from RGBD cameras. However, the results of these applications are often limited by the bad quality of the depth acquisition from the consumer RGB-D camera. It would be very gratifying if we could improve the defective depth maps, such that all the depth-related tasks are likely to be further developed.

In this chapter, we first introduce some pre-processing techniques to fill the missing areas and reduce the noise and quantization effects on the input depth image. Then, we describe in detail one of the state-of-the-art depth refinement method from Or-El *et al.* [11] which we have chosen to implement as a starting point. A method based on an RGB ratio model is then followed and introduced to eliminate the nonlinearity in most of the modern depth enhancement methods. Finally, another proposed technique which does not require any regularization terms is presented. This method has exhibited the ability to deal with the objects with complicated albedos and extend to super-resolution for depth images.

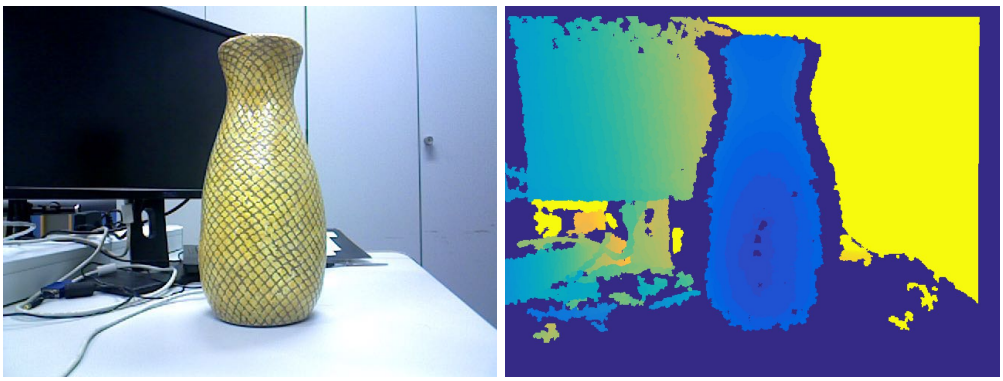
3.1 Pre-Processing

The first step for most of the image processing tasks is to pre-process the initial input image. Due to the hardware limitation of inexpensive RGB-D sensors, there usually exist holes with missing values on the depth images. Also, the depth data is often noisy so we need to do denoising and acquire a relatively smooth surface.

In this part, we will describe respectively the depth inpainting and denoising algorithm used in our pre-processing.

3.1.1 Depth inpainting

Image inpainting itself is a very mutual area and has been widely applied as a useful tool for many modern computer vision applications, e.g, restore the damaged parts of ancient paintings, and remove unwanted texts or objects in a photography [46]. Since the idea of image inpainting is to automatically replace the lost or undesired parts of an image with the neighbouring information by interpolating, it is reasonable to apply it to fill in the missing areas (Fig. 3.1).



(a) RGB image (b) Input depth image

Figure 3.1: The input RGB and depth image of a vase. The depth map is visualized using color from blue (near) to yellow (far), where the missing areas and quantization effects can be clearly noticed. The image (b) corresponds to Fig 3.2(a).

It should be noted that, the depth inpainting is applied to the input rough depth map so there is no need to use any powerful and advanced algorithms. The only request is to fill the missing areas with inexpensive computational time.

The idea of a classic inpainting algorithm [46] can be described as follows. Assuming we have an image I, U is the update information and Ω is the area with missing information. The resulting inpainted image can be treated as the addition between I and U within Ω. To build the update map U, there are two principles that [46] follow. One is the inpainted values inside Ω should be as smooth as possible. The other is that the lines reaching the edge of Ω should be continued and cross the missing area, while the values in Ω should be propagated from the nearest neighbours of Ω along the lines.

Again, due to the fact that our input depth images already have poor quality, the lines arriving at the boundary $\delta\Omega$ may be incorrect or produced by the noises. Thus, it is reasonable that our initial depth inpainting problem merely focuses on the smooth propagation from the neighbours and fill in the holes. Here, in each pixel (x, y) inside Ω, U can be modelled as a

discrete four-neighbour Laplacian operator:

$$U(x,y) = \Delta I_\Omega = 4I(x,y) - I(x+1,y) - I(x-1,y) - I(x,y+1) - I(x,y-1) \tag{3.1}$$

Now the inpainting problem can be represented as a minimization problem:

$$\min \iint_\Omega |U(x,y)|^2 dxdy \tag{3.2}$$

This energy function can be reformulated to a typical linear equation in matrix form:

$$\mathbf{A}\mathbf{x_{in}} = \mathbf{b} \tag{3.3}$$

Assuming m is the number of pixel inside Ω and nb is the sum of m and the number of neighbouring pixel around the boundary $\delta\Omega$, $\mathbf{A} \in \mathbb{R}^{nb \times m}$ is a Laplacian matrix and $\mathbf{b} \in \mathbb{R}^{nb}$ is a vector containing all the known boundary depth values as well as the 0 inside Ω. Solving the linear equation with simple least squares, we can acquire the inpainted values $\mathbf{x_{in}}$ within Ω. With this simple but efficient inpainting algorithm, we can fill the holes on the rough depth image as shown in Fig. 3.2.

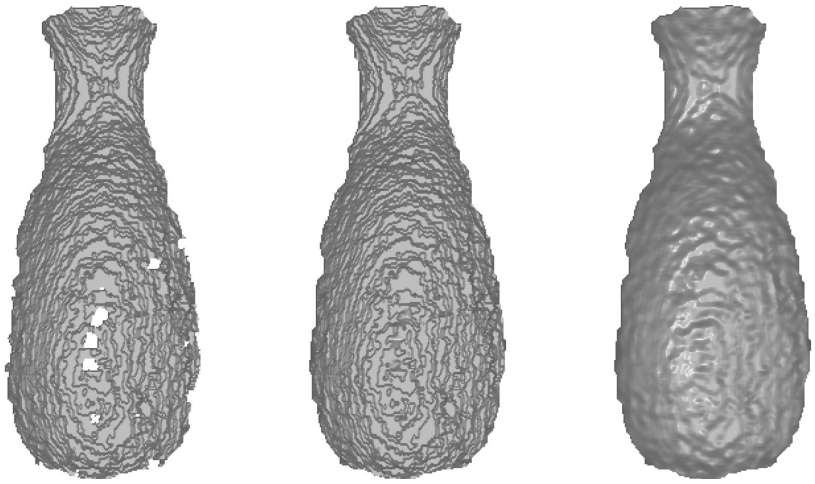
3.1.2 Depth denoising

The depth images acquired from low-cost RGB-D cameras usually not only contain various noises, but also quantized effect. As a standard pre-processing method, the image denoising technique is applied to our inpainted input depth map. Similar to the previous depth refinement methods [6,10,11,14,42,47], bilateral fitering [48] is used as our depth pre-processing smoother.

The advantage of bilateral filter is that it reduces the noise while preserving the edge in the input image. More than a regular Gaussian smooth filter, which uses only the difference of the image values (depth in our case) between the center pixel the neighbours, the bilateral filter also utilizes the space difference as a reference to build up the weighting function. The filtered pixel value can be modelled as a weighted sum of neighbouring pixels:

$$\hat{I}(\mathbf{p}) = \frac{1}{W} \sum_{\mathbf{q} \in \mathcal{N}} I(\mathbf{q}) e^{-\left(\frac{\|I(\mathbf{p}) - I(\mathbf{q})\|^2}{2\sigma_r^2} + \frac{\|\mathbf{p}-\mathbf{q}\|^2}{2\sigma_d^2}\right)} \tag{3.4}$$

where $\hat{I}(\mathbf{p})$ is the filtered value at pixel $\mathbf{p} = (x,y)$, $\mathcal{N}$ represents the neighbouring pixels with $\mathbf{p}$ in the center, and W is the sum of the all the neighbouring weights centering in $\mathbf{p}$, σ_r and σ_d denote the parameters controlling the spatial and depth weighting respectively. The smoothed result on our input depth image is shown in Fig. 3.2.



(a) Input depth (b) After inpainting (c) After smoothing

Figure 3.2: Illustrations for the pre-processing on the depth image of the vase.

After the pre-processing procedure, we have an initial smooth and inpainted depth image. It will be used as the input of all the depth refinement methods detailed in the following sections.

3.2 RGBD-Fusion Like Method

RGBD-Fusion is a state-of-the-art depth recovery method proposed by Or-El *et al.* [11] in 2015. This novel method is adequate for natural scene illumination and able to enhance the depth map much faster than other methods. It is reasonable to gain a comprehensive understanding in the field of depth refinement by implementing this method.

It is worth mentioning that we did not just follow the paper step by step without injecting any our own ideas. For example, instead of estimating the pixel-wise ambient light with a separate energy function, we jointly calculated all four first-order spherical harmonics parameters (3 for point-source light direction and 1 for ambient light) with the fast least squares. In this case, we reduced the number of tuning parameters from 8 to 5 while the results have the only negligible difference. And throughout the whole estimation process of light, albedo and depth, we only used the information within the given mask which also speeded up the algorithm. This is the reason we call our first method "RGB-Fusion Like" method.

The natural uncalibrated illumination condition means the light is no longer a point light source, thus a Lambertian model is not sufficient. Basri and Jocobs [26] have found that low order spherical harmonics (SH) model can well set out the irradiance of the diffused objects

under the natural scene. More specifically, the first-order SH model can capture 87.5% of natural lighting, whose form is extended from the Lambertian reflectance model:

$$I(x,y) = \rho(x,y)(\mathbf{l}^\top \mathbf{n}(x,y) + \varphi) \tag{3.5}$$

where $I : \mathcal{M} \to \mathbb{R}^C$ denotes the intensity values. $\rho : \mathcal{M} \to \mathbb{R}^C$ is the albedo, $\mathbf{l}^\top = \begin{pmatrix} l_x & l_y & l_z \end{pmatrix}$ describes the light direction and φ represents the ambient light. $\mathbf{n} : \mathcal{M} \to \mathbb{S}^2 \subset \mathbb{R}^3$ is the unit length surface normal, which is dependent on the depth z. We define that $C = 1$ represents the grayscale image while $C = 3$ for color image. $(x, y) \in \mathcal{M}$ represents the pixel coordinate inside the given mask $\mathcal{M}$ of an object. Eq. 3.5 can be rewritten as:

$$I(x,y) = \rho(x,y)\ \mathbf{s}^\top \tilde{\mathbf{n}}(x,y) \tag{3.6}$$

where

$$\mathbf{s} = \begin{pmatrix} \mathbf{l} \\ \varphi \end{pmatrix} \quad \tilde{\mathbf{n}}(x,y) = \begin{pmatrix} \mathbf{n}(x,y) \\ 1 \end{pmatrix} \tag{3.7}$$

$\mathbf{s} \in \mathbb{R}^4$ is the first-order SH parameter. It should be mentioned that the 1st-order SH model is used as the fundamental model throughout the whole methodology part.

After introducing the preliminary knowledge, the overall energy function for the RGBD-Fusion Like method which can jointly estimate lights, albedo and depth is described below:

$$\begin{aligned} E(\rho, z, \mathbf{s}) = \sum_{(x,y)\in\mathcal{M}} \sum |I(x,y) - \rho(x,y)\mathbf{s}^\top \tilde{\mathbf{n}}(x,y)|^2 + \lambda_\rho \sum_{(x,y)\in\mathcal{M}} \sum_{k\in\mathcal{N}(x,y)} |\omega_k(x,y)(\rho(x,y) - \rho_k)|^2 \\ + \lambda_z \sum_{(x,y)\in\mathcal{M}} \sum |z(x,y) - z_0(x,y)|^2 + \lambda_l \sum_{(x,y)\in\mathcal{M}} \sum |\Delta z(x,y)|^2 \end{aligned} \tag{3.8}$$

For the sake of simplicity, we will use $\|\cdot\|_2^2 = \sum_{(x,y)\in\mathcal{M}} \sum (\cdot)^2$ to reshape the equation, and then I, z and ρ are vectorized to $\mathbb{R}^m$ within the mask, while $\tilde{\mathbf{n}} \in \mathbb{R}^{m\times 4}$ with each row as $\begin{bmatrix} \mathbf{n}(x,y)^\top & 1 \end{bmatrix}$. m is the number of pixel inside the mask $\mathcal{M}$. We also directly replace $\mathcal{N}$ with $\mathcal{N}(x,y)$ to represent all the neighbourhood of the pixel inside the mask. Finally, Eq. 3.8 can be reformulated as:

$$E(\rho, z, \mathbf{s}) = \|I - \rho \cdot \tilde{\mathbf{n}}(z)\mathbf{s}\|_2^2 + \lambda_\rho \| \sum_{k\in\mathcal{N}} \omega_k(\rho - \rho_k)\|_2^2 + \lambda_z \|z - z_0\|_2^2 + \lambda_l \|\Delta z\|_2^2 \tag{3.9}$$

The mark $\cdot$ represents element-wise multiplication here. The function successively consists of an SFS term, an albedo anisotropic Laplacian term, a depth data fidelity term and a depth isotropic Laplacian term. Now we will go through the details step by step.

3.2.1 Light estimation

Here the unit length surface normal $\mathbf{n}$ is formulated with orthographic projection, i.e.

$$\mathbf{n}(x,y) = \frac{1}{\sqrt{|\nabla z(x,y)|^2 + 1}} \begin{pmatrix} \nabla z(x,y) \\ -1 \end{pmatrix} \tag{3.10}$$

$\nabla z(x,y) = \begin{bmatrix} z_x(x,y) & z_y(x,y) \end{bmatrix}^\top$ represents the gradient of depth image $z(x,y)$ in x and y directions. Since we have the input depth from pre-processing, the initial normal $\mathbf{n}_0$ is known.

To compute the spherical harmonics parameters, we assume the albedo ρ equals to 1 for each pixel. Since there are known intensity values and surface normal in each pixel within the mask, we will have an overdetermined least square problem from the energy in Eq. 3.9:

$$\min_{\mathbf{s}} \|\tilde{\mathbf{n}}\mathbf{s} - I\|_2^2 \tag{3.11}$$

This process only applies once at the beginning of the process since the least squares is not sensitive to the details on the surface, thus the estimation from the smooth surface is enough [11].

3.2.2 Albedo estimation

As mentioned in chapter 2, many depth recovery techniques based on SFS or PS assume constant or uniform albedo. Such assumption does not fit in with the real-world objects so they perform poorly on the shape estimation for multi-albedo cases. In order to acquire a decent shape outcome, an effective multi-albedo estimation process is a matter of importance.

We know from Eq. 3.5 that, assuming we have the knowledge of input intensity and estimated shading, the albedo image can be directly obtained from $I/(\tilde{\mathbf{n}}\mathbf{s})$. However, due to the fact that both input image I and the surface normal $\mathbf{n}$ are noisy, the albedo acquired like this is prone to the overfitting, which makes the acquired albedo contain all the undesired spatial layout details, as illustrated in Fig. 3.3. To resolve the overfitting problem, we should impose some restrictions on the estimation of albedo. Land's Retinex theory [49] addressed that the reflectance of an object is normally flat with sparse strong variations. Indeed, plenty of real-world objects has piecewise smooth appearance as shown in Fig. 3.4, which means most parts of a layout are dominated by certain colors. Therefore, a prior that emphasizes the piecewise smoothness on the albedo should be defined. Such assumption of piecewise smoothness has already been widely used in many photometric stereo methods [39, 50].

With this assumption, the albedo of an object can be roughly divided into several pieces with different intensities, which can be treated as the image segmentation problem to some extent. Thus, we should refer to some classic variational segmentation methods and adapt the

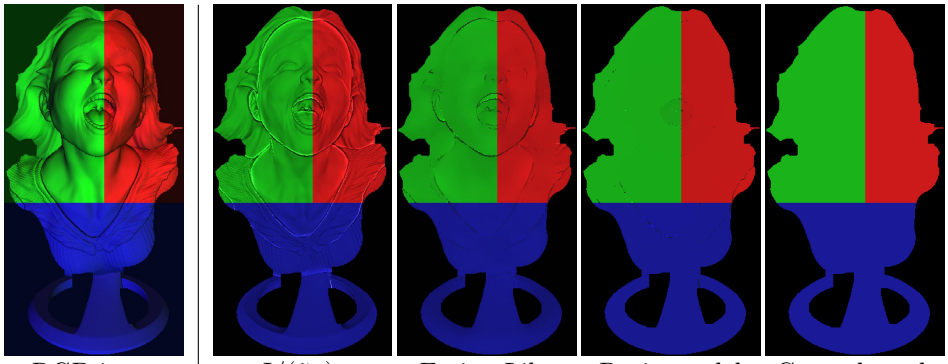


Figure 3.3: Comparison for the estimated albedo under the synthetic data. Note that the light parameter is given. From left to right: the albedo obtained without regularization, albedo estimated by RGBD-Fusion like method, albedo from the proposed ratio model and the ground truth albedo.

edge preserving smoothness term to our problem. Similar to the idea in [51], an anisotropic Laplacian term is imposed to estimate the albedo. Now, the SH parameters $\mathbf{s}$ and the surface normal $\mathbf{n}$ are freezed and the overall regularized minimization problem in Eq. 3.9 is:

$$\min_{\rho} \|\rho \cdot \tilde{\mathbf{n}}\mathbf{s} - I\|_2^2 + \lambda_\rho \| \sum_{k \in \mathcal{N}} \omega_k (\rho - \rho_k)\|_2^2 \tag{3.12}$$

where k indicates the neighbouring index of a certain pixel, which 4-connected set is chosen for $\mathcal{N}$ in our case. The weight ω_k is defined as below, and it is dependent on two parameters σ_I and σ_z which accounts for the discontinuity in both intensity and depth.

$$\omega_k = \exp\left(- \frac{\|I - I_k\|_2^2}{2\sigma_I^2} - \frac{\|z - z_k\|_2^2}{2\sigma_z^2} \right) \tag{3.13}$$

3.2.3 Depth enhancement

After acquiring the first-order spherical lighting parameters $\mathbf{s}$ and the albedo ρ, we can refine our depth with the help of Eq. 3.5 and Eq. 3.10. Now our minimization problem with respect to the depth z in Eq. 3.9 can be written as below. The data fidelity term is applied to resolve the SFS ambiguities and enables our refined surface close to the input. The Laplacian smoothness term makes sure that there is no strong discontinuity in the output.

$$\min_{z} \|\rho \cdot \tilde{\mathbf{n}}(z)\mathbf{s} - I\|_2^2 + \lambda_z \|z - z_0\|_2^2 + \lambda_l \|\Delta z\|_2^2 \tag{3.14}$$

where z_0 is the input depth and Δ represents the Laplacian operator. Because the normal defined in Eq. 3.10 contains a denominator related to the depth gradient, the SFS term in the energy is nonlinear. Many optimization methods can be applied to solve the nonlinear problem, e.g. Levenberg-Marquardt algorithm or ADMM, but they are not suitable for our application due to expensive computational time. Here a "fixed point" method which is similar to iteratively reweighted least square (IRLS) is introduced to deal with problem efficiently.

The idea of the fixed-point approach is in each iteration, the normalizer in the surface normal can be treated as a weighting term and determined by the depth from the last iteration. With the help of this trick, the normalizer is known and Eq. 3.14 is linear again. We can solve the linear system using any fast linear optimization method. In each iteration t, this process can be represented element-wise as follows:

$$\begin{aligned} \mathbf{n}(z^{(t)}, z^{(t-1)}) &= w(z^{(t-1)}) \begin{pmatrix} \nabla z^{(t)} \\ -1 \end{pmatrix} \\ w(z^{(t-1)}) &= \frac{1}{\sqrt{1 + |\nabla z^{(t-1)}|^2}} \end{aligned} \tag{3.15}$$

And now the depth refinement problem in Eq. 3.14 is reformulated as below in each iteration:

$$\min_{z^{(t)}} \|\rho \cdot \tilde{\mathbf{n}}(z^{(t)}, z^{(t-1)})\mathbf{s} - I\|_2^2 + \lambda_z \|z^{(t)} - z_0\|_2^2 + \lambda_l \|\Delta z^{(t)}\|_2^2 \tag{3.16}$$

As long as the energy decreases in each iteration, the process is repeated.

To sum up this section, it should be noted that the SFS term was used as a core in all light, albedo and depth estimation in the overall energy. The whole process of the RGBD Fusion-Like method has been described in Alg. 1 and some real-world results are shown in Fig. 3.4.

Algorithm 1 RGBD-Fusion Like Depth Refinement

Input: Initial depth image z_0, RGB image I

1: Estimate the SH parameter, $\mathbf{s} = \arg\min_{\mathbf{s}} E(\rho = 1, z_0)$ {Eq. 3.11}
2: Estimate the albedo, $\rho = \arg\min_{\rho} E(z_0, \mathbf{s})$ {Eq. 3.12}
3: t = 1, $z^{(t-1)} = z_0$
4: **while** $E(\rho, z^{(t)}, \mathbf{s}) - E(\rho, z^{(t-1)}, \mathbf{s}) < 0$ **do**
5: $\quad z^{(t)} = \arg\min_{z} E(\rho, z, \mathbf{s})$ {Eq. 3.16}
6: $\quad t := t + 1$
7: **end while**

Output: Refined depth image $z^{(t)}$

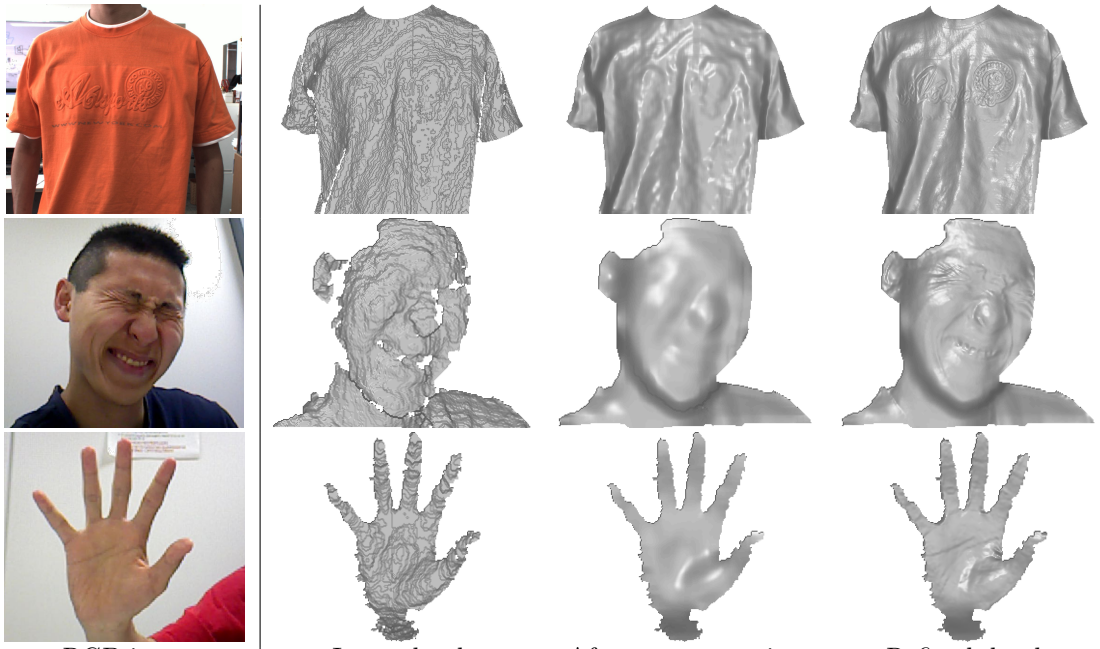


Figure 3.4: Illustrations for our implementation of RGBD-Fusion Like method. Top row is a T-shirt from [6]. Middle and bottom row are the author's face and palm.

3.2.4 Limitations

Although our RGBD-Fusion Like method works moderately well in some real cases, it is not difficult to find the limitations and improve correspondingly.

- the surface normal modelled by the orthographic projection is merely an ideal case which is not really in line with the real world camera model. And the intrinsic parameters such as the focal length and the coordinate of the principle point are either usually given as a preliminary knowledge, or obtained from camera calibration without much effort. Hence, it is reasonable to formulate the surface normal with the perspective projection.

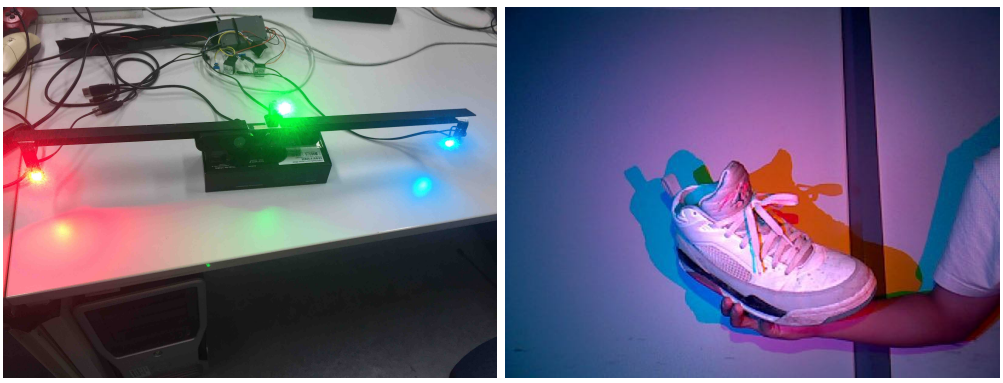
- In our RGBD-Fusion Like method, only the intensity is applied because the pixel values in RGB channels are more or less the same under the natural scene illumination. When we estimated the SH lighting parameters and the albedo in 3 channels separately, the results are quite similar to each other. So using all three channels rather than just the grayscale image will neither provide much extra information nor improve the depth enhancement. Instead, it will just decelerate the whole algorithm. We ought to find a way to take better advantages of all three channels in a photometric stereo manner.

- The most important inspiration for us to propose the new RGB ratio model in the next section is, the RGBD-Fusion Like method was not always convergent in terms of depth enhancement part because of the fixed-point optimization method. In the 4th line of the Alg. 1, we force the iteration to stop when the energy for the depth refinement starts increasing. This is due to the reason that the fixed-point method actually tries to solve the non-linearity in a tricky way, which is not totally mathematically correct. Therefore, we thought of the idea of RGB ratio model, which can eliminate the denominator inside the normal and promise a real linear problem.

3.3 Proposed Method I: RGB Ratio Model

According to the limitations in the last section, we thought of the idea of RGB ratio model. First of all, we replace the orthographic projection model with the perspective one to represent the surface normal. And then, to fully use the information of the RGB three channels while eliminating the nonlinearity in the objective function in the depth refinement, we use the ratio model between every pair of the RGB image.

It should be noted that we need to add active red, green and blue LED lights for the sake of emphasizing the difference among RGB channels. The green LED is installed in the middle with the red and blue ones on the two sides of ASUS Xtion Pro Live camera (both are around 30 cm to the green LED). The hardware setup of our system and a color image taken with such setup are illustrated in Fig. 3.5.



(a) LED lights setup (b) A color image with our lighting setup

Figure 3.5: Illustrations for the color LEDs setup and the acquired RGB image.

Now to derive our new ratio model, we treat each channel of the color image I as a single intensity image, denoted by I_R, I_G, I_B. Therefore, 3 equations can be obtained from Eq. 3.5.

$$\begin{aligned} I_R &= \rho_R(\mathbf{l}_R^\top \mathbf{n} + \varphi_R) \\ I_G &= \rho_G(\mathbf{l}_G^\top \mathbf{n} + \varphi_G) \\ I_B &= \rho_B(\mathbf{l}_B^\top \mathbf{n} + \varphi_B) \end{aligned} \tag{3.17}$$

Using R and G channel as an example, we acquire the ratio model between R and G:

$$\frac{I_R - \rho_R \varphi_R}{I_G - \rho_G \varphi_G} = \frac{\rho_R \mathbf{l}_R^\top \mathbf{n}}{\rho_G \mathbf{l}_G^\top \mathbf{n}} \tag{3.18}$$

Similarly, we can acquire another two ratio models which are between green and blue, and blue and red channels respectively. It can be noticed from Eq 3.18 that, the non-linearity problem mentioned before has been solved because the denominator in the surface normal $\mathbf{n}$ is cancelled out. Also, our normal is derived from perspective camera model and can be represented as a function of $\log z$. For the sake of simplicity we directly represent $z = \log z$ and redefine $\mathbf{n}$ without the normalizer in the following part.

$$\mathbf{n}(x,y) = \begin{pmatrix} f z_x(x,y) \\ f z_y(x,y) \\ -1 - \tilde{x} z_x(x,y) - \tilde{y} z_y(x,y) \end{pmatrix} \tag{3.19}$$

where f is the focal length, $(\tilde{x}, \tilde{y}) = (x - x_0, y - y_0)$, with (x_0, y_0) the coordinates of principle points and (x, y) the coordinates of a pixel inside the given mask $\mathcal{M}$, $[z_x, z_y]$ is the gradient of depth z.

The overall energy for the proposed RGB ratio model method is:

$$\begin{aligned} E(\mathcal{P}, z, \mathbf{s}) = \ & \|\mathcal{R}(\mathcal{P}, z)\|_2^2 + \lambda_z ||z - z_0||^2 + \lambda_\rho^1 \|\omega \nabla \mathcal{P}\|^2 + \sum_c \|\rho_c \, \mathbf{s}_c^\top \tilde{\mathbf{n}} - I_c\|_2^2, \\ & c \in \{R, G, B\} \end{aligned} \tag{3.20}$$

where $\mathcal{P} \in \mathbb{R}^{3m}$ is the stack of RGB albedo, and $\mathcal{R}$ denotes our ratio model which will be defined separately in the following part. m is the number of pixels inside $\mathcal{M}$. This energy for RGB ratio model is successively composed of a proposed ratio shading term, a depth fidelity term, an albedo smoothness term, an albedo fidelity term and a SH estimation term. Now we will explain the proposed algorithm based on the new ratio model.

3.3.1 Algorithm details

Similar to the RGBD-Fusion Like method, the algorithm is separated into 3 parts: light estimation, albedo estimation and depth enhancement. However, our new method requires an initial estimation of the color albedo as the input of our iterative method, and hence, we calculate the initial SH parameters $\mathbf{s}^0$ with Eq. 3.5 and the color albedo $\mathcal{P}^0$ with Eq. 3.12 using the old model. Noted that the initial estimation is performed with respect to all RGB three channels.

Albedo refinement: with the acquired initial SH parameters and albedo, we can start the iterative refinement process. In order to refine the color albedo with our ratio model, in each iteration, we need to reshape the ratio model described in Eq. 3.18 as follows:

$$\begin{aligned} I_G\mathbf{l}_R^\top\mathbf{n}\rho_R - I_R\mathbf{l}_G^\top\mathbf{n}\rho_G &= \rho_R\rho_G(\varphi_G\mathbf{l}_R^\top\mathbf{n} - \varphi_R\mathbf{l}_G^\top\mathbf{n}) \\ I_B\mathbf{l}_G^\top\mathbf{n}\rho_G - I_G\mathbf{l}_B^\top\mathbf{n}\rho_B &= \rho_G\rho_B(\varphi_B\mathbf{l}_G^\top\mathbf{n} - \varphi_G\mathbf{l}_B^\top\mathbf{n}) \\ I_R\mathbf{l}_B^\top\mathbf{n}\rho_B - I_B\mathbf{l}_R^\top\mathbf{n}\rho_R &= \rho_B\rho_R(\varphi_R\mathbf{l}_B^\top\mathbf{n} - \varphi_B\mathbf{l}_R^\top\mathbf{n}) \end{aligned} \tag{3.21}$$

For each pixel, we can reformulate the Eq. 3.21 to a matrix form:

$$\begin{pmatrix} I_G\mathbf{l}_R^\top\mathbf{n} & -I_R\mathbf{l}_G^\top\mathbf{n} & 0 \\ 0 & I_B\mathbf{l}_G^\top\mathbf{n} & -I_G\mathbf{l}_B^\top\mathbf{n} \\ -I_B\mathbf{l}_R^\top\mathbf{n} & 0 & I_R\mathbf{l}_B^\top\mathbf{n} \end{pmatrix} \begin{pmatrix} \rho_R \\ \rho_G \\ \rho_B \end{pmatrix} = \begin{pmatrix} \rho_R\rho_G(\varphi_G\mathbf{l}_R^\top\mathbf{n} - \varphi_R\mathbf{l}_G^\top\mathbf{n}) \\ \rho_G\rho_B(\varphi_B\mathbf{l}_G^\top\mathbf{n} - \varphi_G\mathbf{l}_B^\top\mathbf{n}) \\ \rho_B\rho_R(\varphi_R\mathbf{l}_B^\top\mathbf{n} - \varphi_B\mathbf{l}_R^\top\mathbf{n}) \end{pmatrix} \tag{3.22}$$

This small linear system can be easily generalized to a big sparse linear system denoted by $\mathbf{A}_\rho\mathcal{P} = \mathbf{b}_\rho$, where $\mathbf{A}_\rho \in \mathbb{R}^{3m\times 3m}, \mathbf{b}_\rho \in \mathbb{R}^{3m}$. $\mathcal{P} \in \mathbb{R}^{3m}$ represents the stack of RGB albedos.

To acquire the RGB albedos, some regularization terms are required similar to Eq. 3.12. Now in each iteration, we fix the normal and the SH parameters, the minimization problem of color albedo in Eq. 3.20 in each iteration now becomes:

$$\mathcal{P}^{(t)} = \arg\min_{\mathcal{P}} \|\mathbf{A}_\rho^{(t-1)}\mathcal{P} - \mathbf{b}_\rho^{(t-1)}\|^2 + \lambda_\rho^1\|\omega\nabla\mathcal{P}\|^2 + \lambda_\rho^2\|\mathcal{P} - \mathcal{P}^{(t-1)}\|^2 \tag{3.23}$$

where the weight $\omega = \text{diag}(\begin{pmatrix} \omega_R \\ \omega_G \\ \omega_B \end{pmatrix}) \in \mathbb{R}^{3m\times 3m}$, which can be denoted as:

$$\omega_c = \exp(-\frac{\sigma_c||\nabla I_c||^2}{\max||\nabla I_c||^2}), \quad c \in \{R, G, B\} \tag{3.24}$$

σ_c is a tuning parameter for each channel c. We can notice from Fig. 3.6 the importance of imposing the weight ω. Without the weight, the isotropic smoothness regularization will not take care of the boundary of the albedo, which leads to the bad depth enhancement.

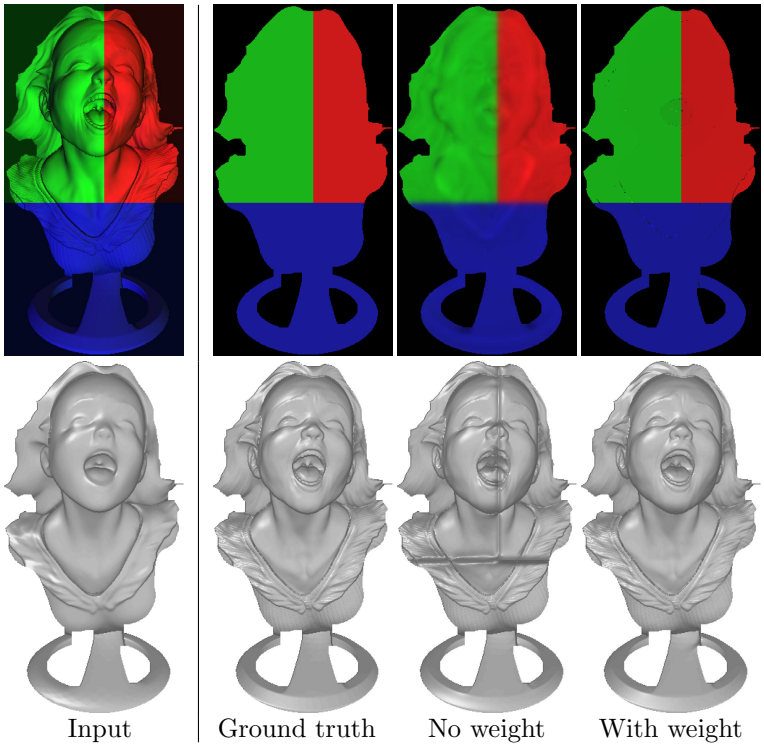


Figure 3.6: Illustrations for the importance of the weight ω inside regularization term in Eq. 3.23 when estimating the albedo. Top: first one is the input color image and the other three are the albedos. Bottom: 3D surface from depth. Note that the light parameter is given.

There are three interesting aspects of the albedo estimation which are worth having a few more words:

1. If we do not use a data fidelity term $\|\mathcal{P} - \mathcal{P}^{(t-1)}\|^2$, the albedo will get increasingly dark after several iterations. This is due to the fact that there also exist the RGB albedos in $\mathbf{b}_\rho$, so $\rho = 0$ will become the solution of our ratio model term. Therefore, adding the data term can not only avoid such problem but help refine the albedo iteratively.

2. One observation about Eq. 3.22 is that, if the SH parameters are the same among the three channels, the right side of the equal sign becomes 0. This is one of the reasons why we really need to set up 3 LED lights with a distance to each other, which will provide us with enough difference on the light directions.

3. Instead of using an anisotropic Laplacian regularization like in RGBD Fusion Like method, the smoothness term in Eq. 3.23 only takes the use of the gradient of ρ with a weight only

depending on the RGB image's gradient. It takes fewer efforts to build such a smoothness term than the anisotropic term but turns out the acquired albedo is still satisfactory.

Depth refinement: After acquiring the color albedo in time step t, we are going to refine the depth with the help of the ratio model. First we reshape Eq. 3.18 with the surface normal $\mathbf{n}$ as the argument:

$$\begin{aligned}
\rho_G(I_R - \rho_R\varphi_R)\mathbf{l}_G^\top\mathbf{n} - \rho_R(I_G - \rho_G\varphi_G)\mathbf{l}_R^\top\mathbf{n} = 0 \\
\rho_B(I_G - \rho_G\varphi_G)\mathbf{l}_B^\top\mathbf{n} - \rho_G(I_B - \rho_B\varphi_B)\mathbf{l}_G^\top\mathbf{n} = 0 \\
\rho_R(I_B - \rho_B\varphi_B)\mathbf{l}_R^\top\mathbf{n} - \rho_B(I_R - \rho_R\varphi_R)\mathbf{l}_B^\top\mathbf{n} = 0
\end{aligned} \tag{3.25}$$

since the normal $\mathbf{n}$ now is a function of z, Eq. 3.25 can be actually simplified as below (the derivation details can be found in Appendix B):

$$\Psi z = 0 \tag{3.26}$$

When the estimated color albedo and light are fixed, the depth refinement problem in Eq. 3.20 is:

$$z^{(t)} = \arg\min_z ||\Psi z||^2 + \lambda_z||z - z_0||^2 \tag{3.27}$$

Light estimation Since estimating light with the proposed ratio model is an ill-posed problem, we decided to use Eq. 3.11 to calculate the SH parameters for each channel when the albedo and the surface normal are frozen. The minimization problem from Eq. 3.20 can then be written as:

$$\mathbf{s}^{(t)} = \underset{\mathbf{s}=(\mathbf{s}_R,\mathbf{s}_G,\mathbf{s}_B)}{\arg\min} \sum_c \|\rho_c\, \mathbf{s}_c^\top \tilde{\mathbf{n}} - I_c\|_2^2,\ c \in \{R, G, B\} \tag{3.28}$$

Alg. 2 outlines the process of RGB ratio model method and Fig. 3.7 illustrates the effectiveness of this approach.

3.3.2 Limitations

The RGB ratio model method can estimate the albedo and the depth better than our RGBD-Fusion Like method in some cases owing to that the non-linearity optimization problem in the RGBD-Fusion has been solved. Nevertheless, there still exist some defects:

- Three LED lights have to be set up far away from each other. As already mentioned about Eq. 3.21, the albedo refinement may fail if the lights are set too close. This can lead to some inconvenience, such as the requirement of relatively larger space to set up the system than the RGBD-Fusion like method.

Algorithm 2 RGB Ratio Model method

Input: Initial depth image z_0, RGB image I, mask, focal length, principle point

1: $\mathbf{s}^{(0)} = \arg\min_{\mathbf{s}} E(\mathcal{P}=1, z_0)$ {Eq. 3.28}
2: Estimate initial color albedo and build $\mathcal{P}^{(0)}$ {Eq. 3.12}
3: t = 1, $z^{(0)} = z_0$
4: **while** $\frac{\|E(\mathcal{P}^{(t)}, z^{(t)}, \mathbf{s}^{(t)}) - E(\mathcal{P}^{(t-1)}, z^{(t-1)}, \mathbf{s}^{(t-1)})\|}{E(\mathcal{P}^{(t-1)}, z^{(t-1)}, \mathbf{s}^{(t-1)})} > \epsilon$ **do**
5: $\mathcal{P}^{(t)} = \arg\min_{\mathcal{P}} E(\mathcal{P}^{(t-1)}, z^{(t-1)}, \mathbf{s}^{(t-1)})$ {Eq. 3.23}
6: $z^{(t)} = \arg\min_{z} E(\mathcal{P}^{(t)}, \mathbf{s}^{(t-1)})$ {Eq. 3.27}
7: $\mathbf{s}^{(t)} = \arg\min_{\mathbf{s}} E(\mathcal{P}^{(t)}, z^{(t)})$ {Eq. 3.28}
8: $t := t + 1$
9: **end while**

Output: Refined depth image $z^{(t)}$ and stacked color albedo $\mathcal{P}^{(t)}$

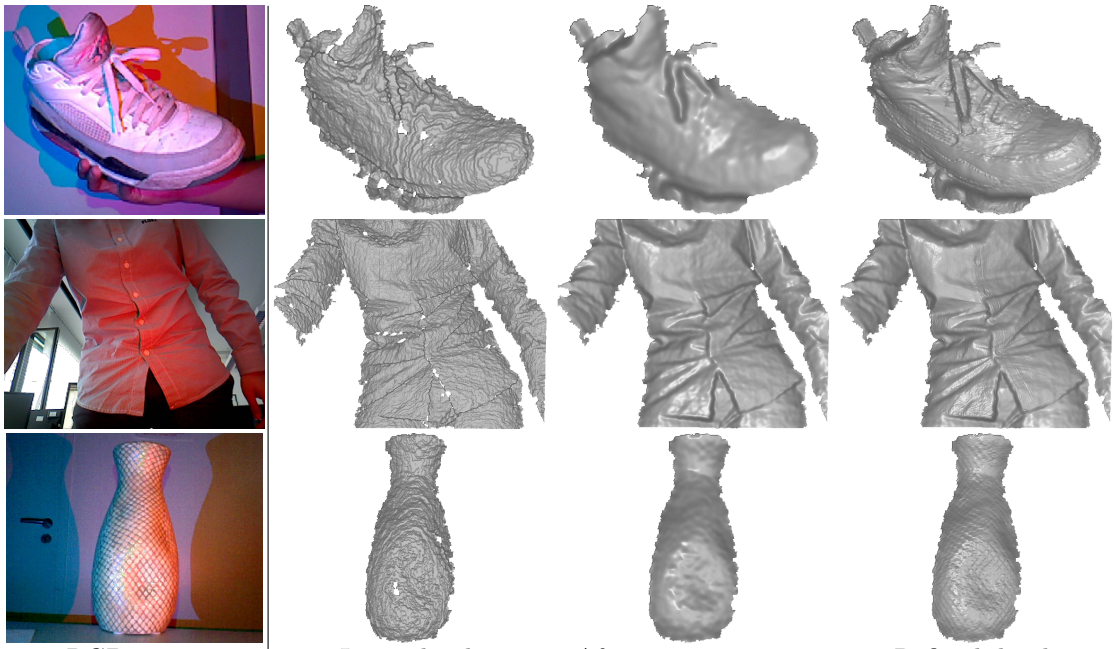


Figure 3.7: Illustrations for the depth refinement of our proposed RGB ratio model. Many fine geometric details have been refined on the depths, which shows the effectiveness of this method. It should be mentioned that the middle row is under the natural scene illumination and our method still works well.

- The active red, green and blue LEDs are likely to bring extra specularity. If we want to refine the depth of a specular object, the specular reflection will be from not only the natural scene illumination, but also RGB lights from 3 directions, which will make the refined results even worse.

- Auto white balance (AWB) has a big impact on the refined results. This is due to the fact that the success of our model highly relies on the difference among 3 channels in a color image. And AWB will mix up the information among the channels so it is very necessary to turn it off. This impedes the generalization of our model because AWB has been set as a default in many modern inexpensive cameras.

3.4 Proposed Method II: Robust Multi-Light Model

3.4.1 Inspiration

We can notice that the albedo estimation of both our RGBD-Like method and RGB ratio model is highly dependent on the regularization terms which emphasize the piecewise smoothness. This is a standard approach for almost all the state-of-the-art depth refinement method to estimate the albedo. They work fine when the albedo itself is very simple with big patches of patterns and only several dominant colors. However, there are more real-world objects containing complicated layout colors and patterns which all these methods with such a process of albedo estimation will fail. Were the albedo estimation not working, the outcome of final depth refinement has no chance to be correct. What is more, The parameters for the regularization terms are often needed to be different for various objects, so the parameters tuning for the regularization is tedious and time-consuming.

As a consequence, it is reasonable to propose a new method which is able to eliminate all the regularization terms and estimate the complicated albedo. The necessity of using regularizations for calculating albedo have been mentioned in section 3.2.2, which in short is about avoiding the overfitting problem and eliminating ambiguities if only the shading term is applied. Provided we have several color images for a still object with lights coming from various directions, all the images can be jointly used to estimate the albedo of that object. Hence, unlike the depth refinement using only one image, the ambiguity in the shading term has been resolved so there is no need to impose any regularization. Fig. 3.9 has shown the robustness of our proposed method for the albedo estimation.

In order to simulate the scenario that an object is illuminated from various directions, we simply sway a white LED in different directions and take several images (even the flashlight on any phone is sufficient). An example of a vase with different lightings is shown in Fig. 3.8.

Figure 3.8: Illustrations for the obtained color images of a vase from various light directions with a white LED light. Even the phone flashlight is sufficient for giving various lightings.

3.4.2 Algorithm details

Since we do not control the lighting with red, green and blue LEDs but use one white light instead, the ratio model in Eq. 3.18 is not applicable here. We decide to use the standard 1st-order SH model described in Eq. 3.5 again to construct the input color images.

Similar to those two methods mentioned before, the proposed algorithm consists of three parts: light estimation, albedo estimation and depth enhancement. We need to iteratively update all of them but unlike the RGB ratio model method, we do not need to have an initial estimated light and albedo beforehand. Instead, the albedo can be assumed to be 1 everywhere and the illuminations can be set to frontal directions at the beginning. Then we start refining everything iteratively in the loop, as shown in Alg. 3.

First of all, the shading model we use for the proposed method is still based on SH model in Eq. 3.5, so the corresponding photometric stereo minimization problem now becomes:

$$\sum_i \sum_c \sum_{(x,y)\in\mathcal{M}} |\rho_c(x,y)\mathbf{s}_{i,c}^\top \tilde{\mathbf{n}}(x,y) - I_{i,c}(x,y)|^2 \tag{3.29}$$

$c \in \{R,G,B\},\ i \in \{1,\cdots,n\}$, where n stands for the total number of varying light directions. A simplified version of this shading energy is:

$$\sum_i \sum_c \|\rho_c \cdot \tilde{\mathbf{n}}\mathbf{s}_{i,c} - I_{i,c}\|_2^2 \tag{3.30}$$

Now the overall energy for our proposed robust multi-light method is characterized as:

$$E(\rho, z, \mathbf{s}) = \sum_i \sum_c \|\rho_c \cdot \tilde{\mathbf{n}}(z)\mathbf{s}_{i,c} - I_{i,c}\|_2^2 + \lambda_z \|z - z_0\|_2^2 \tag{3.31}$$

The mark $\cdot$ denotes element-wise multiplication. As noticed, the new overall energy is extremely simple with only one shading term and one depth fidelity term, without any regularization terms for the albedo or depth estimation. And we have found out that $\lambda_z = 100$ works well for all

cases, which means our system can be easily used by anybody without problems.

Light estimation In each iteration, we first freeze the albedo and the surface normal and then refine the SH parameters for all input images from the overall energy in Eq. 3.31. To estimate the light with the simple least squares, we reshape the shading term to a linear problem with the SH light as the argument.

Let $\cdot$ denote element-wise multiplication between a vector and each column of a matrix, we have matrices $\mathbf{A}_{\mathbf{s}_{i,c}} = \rho_c \cdot \tilde{\mathbf{n}}(z)$ for every channel in all input color images. Therefore, the illuminations can be estimated with the shading term in the energy as below:

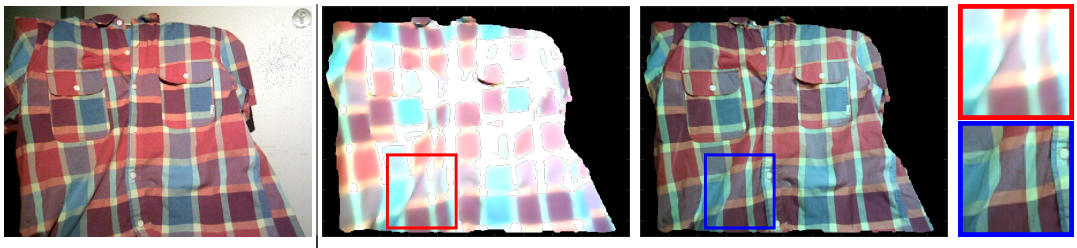
$$\min_{\mathbf{s}_{i,c}} \|\mathbf{A}_{\mathbf{s}_{i,c}}\mathbf{s}_{i,c} - I_{i,c}\|_2^2 \tag{3.32}$$

Albedo estimation Similar to the light estimation, we need to reshape the SFS term in the overall energy in order to solve it with least squares. The energy for the albedo estimation from the overall energy looks like this:

$$\min_{\rho_c} \|\mathbf{A}_{\rho_c}\rho_c - \mathbf{I}_c\|_2^2 \tag{3.33}$$

$\mathbf{A}_{\rho_c} \in \mathbb{R}^{mn\times m}$ is the stack of n diagonal matrices $\text{diag}(\tilde{\mathbf{n}} \cdot \mathbf{s}_{i,c})$, where $\text{diag} : \mathbb{R}^m \to \mathbb{R}^{m\times m}, i \in \{1, \cdots, n\}$, and $\mathbf{I}_c \in \mathbb{R}^{mn}$ denotes the stack of all the vectorized I_i in channel c, which can be represented as:

$$\mathbf{A}_{\rho_c} = \begin{pmatrix} \text{diag}(\tilde{\mathbf{n}} \cdot \mathbf{s}_{1,c}) \\ \vdots \\ \text{diag}(\tilde{\mathbf{n}} \cdot \mathbf{s}_{n,c}) \end{pmatrix} \quad \mathbf{I}_c = \begin{pmatrix} I_{1,c} \\ \vdots \\ I_{n,c} \end{pmatrix} \tag{3.34}$$



RGB image | RGBD-Fusion [11] | Proposed method

Figure 3.9: Comparison for the albedo estimation between our proposed robust multi-light method and RGBD-Fusion method. RGBD-Fusion was under our implementation since their source code did not provide the albedo estimation. The details show the robustness of our method and the ability of eliminating the shadows on the object.

Depth enhancement After having the estimated light and the color albedo, we can continue refining the depth. Again, we need to rearrange the energy function with the depth z as the argument. First, let us start with the simplest case and consider one pixel in one of the input images. If we expand Eq. 3.5 with the the perspective projection normal in Eq. 3.19, we have:

$$I(x,y) = \rho(x,y) \cdot \begin{pmatrix} l^1 & l^2 & l^3 \end{pmatrix} \begin{pmatrix} fz_x(x,y) \\ fz_y(x,y) \\ -1-(x-x_0)z_x(x,y)-(y-y_0)z_y(x,y) \end{pmatrix} / d(x,y) + \rho(x,y) \cdot \varphi \tag{3.35}$$

where $d = \sqrt{(fz_x(x,y))^2 + (fz_y(x,y))^2 + (-1-(x-x_0)z_x(x,y)-(y-y_0)z_y(x,y))^2}$ is a normalizer. After rearranging, we have:

$$\frac{l^1 f - l^3(x-x_0)}{d(x,y)} z_x(x,y) + \frac{l^2 f - l^3(y-y_0)}{d(x,y)} z_y(x,y) = I(x,y) + \frac{l^3}{d(x,y)} - \rho(x,y)\varphi \tag{3.36}$$

we extend this equation to all the pixels in the mask $\mathcal{M}$ and acquire:

$$\frac{l^1 f - l^3 \tilde{x}}{d} \cdot z_x + \frac{l^2 f - l^3 \tilde{y}}{d} \cdot z_y = I + \frac{l^3}{d} - \varphi \cdot \rho \tag{3.37}$$

Provided we denote the gradient matrices in x and y directions as D_x and D_y, Eq. 3.37 becomes:

$$\left(\mathrm{diag}(\frac{l^1 f - l^3 \tilde{x}}{d}) D_x + \mathrm{diag}(\frac{l^2 f - l^3 \tilde{y}}{d}) D_y \right) z = I + \frac{l^3}{d} - \varphi \cdot \rho \tag{3.38}$$

This is the linear equation for one image. Now we define the $\mathbf{A}_z$ and $\mathbf{b}_z$ for our system:

$$\begin{aligned} \mathbf{A}_{z_c} &= \begin{pmatrix} \mathrm{diag}(\frac{l^1_{1,c} f - l^3_{1,c}\tilde{x}}{d_1}) D_x + \mathrm{diag}(\frac{l^2_{1,c} f - l^3_{1,c}\tilde{y}}{d_1}) D_y \\ \vdots \\ \mathrm{diag}(\frac{l^1_{n,c} f - l^3_{n,c}\tilde{x}}{d_n}) D_x + \mathrm{diag}(\frac{l^2_{n,c} f - l^3_{n,c}\tilde{y}}{d_n}) D_y \end{pmatrix}_{mn \times m} \\ \mathbf{b}_{z_c} &= \begin{pmatrix} I_{1,c} + \frac{l^3_{1,c}}{d_1} - \varphi_{1,c} \cdot \rho_c \\ \vdots \\ I_{n,c} + \frac{l^3_{n,c}}{d_n} - \varphi_{n,c} \cdot \rho_c \end{pmatrix}_{mn \times 1} \end{aligned} \tag{3.39}$$

It is worth mentioning that in each iteration, we freeze the denominator d with the z from last

iteration to solve the non-linearity. Finally, we stack $\mathbf{A}_{z_c}$ and $\mathbf{b}_{z_c}$ for each channel $c \in \{R, G, B\}$:

$$\mathbf{A}_z = \begin{pmatrix} \mathbf{A}_{z_R} \\ \mathbf{A}_{z_G} \\ \mathbf{A}_{z_B} \end{pmatrix}, \quad \mathbf{b}_z = \begin{pmatrix} \mathbf{b}_{z_R} \\ \mathbf{b}_{z_G} \\ \mathbf{b}_{z_B} \end{pmatrix} \tag{3.40}$$

After all the derivations, we finally model our energy for the depth enhancement as:

$$\min_z \|\mathbf{A}_z z - \mathbf{b}_z\|_2^2 + \lambda_z \|z - z_0\|_2^2 \tag{3.41}$$

The conjugate gradient (CG) method has been applied to optimize all three sub energies. The structure of the proposed algorithm is described in Alg. 3 and one refined depth example is shown in Fig. 3.10. More examples can be found in chapter 4.

Algorithm 3 Robust Multi-Light Model Method

Input: Initial depth image z_0, RGB image I, mask, focal length, principle point

1: t = 1, $z^{(t-1)} = z_0$, $\rho_R^{(0)}, \rho_G^{(0)}, \rho_B^{(0)} = 1$

2: **while** $\frac{\|E(\rho^{(t)}, z^{(t)}, \mathbf{s}^{(t)}) - E(\rho^{(t-1)}, z^{(t-1)}, \mathbf{s}^{(t)})\|}{E(\rho^{(t-1)}, z^{(t-1)}), \mathbf{s}^{(t-1)}} > \epsilon$ **do**

3: $\quad \mathbf{s}^{(t)} = \arg\min_{\mathbf{s}} E(\rho^{(t-1)}, z^{(t-1)})$ {Eq. 3.32}

4: $\quad \rho^{(t)} = \arg\min_{\rho} E(z^{(t-1)}, \mathbf{s}^{(t)})$ {Eq. 3.33}

5: $\quad z^{(t)} = \arg\min_{z} E(\rho^{(t)}, z^{(t-1)}, \mathbf{s}^{(t)})$ {Eq. 3.43}

6: $\quad t := t + 1$

7: **end while**

Output: Refined depth image $z^{(t)}$ and albedo $\rho^{(t)}$

3.4.3 When super-resolution meets depth refinement

We can notice that there exist over-smoothing problems in many 3D scanning applications [1], which is due to the fact that the depth map given by the consumer RGB-D camera usually has a really low resolution. The scanning or reconstruction results will be then improved if the depth resolution is increased. For most of the well-known affordable RGB-D cameras, the depth resolution is far smaller than the RGB image resolution. For instance, ASUS Xtion Pro Live can acquire 1280×1024 RGB images and 640×480 depth images. Microsoft Kinect 2.0 owns 1920×1080 RGB resolution but only 512×424 depth one, and Intel RealSense R200 has a 1920×1080 RGB camera while its depth reoslution is 640×480. It would be very useful if we can not only refine the depth map in its original scale but close to the RGB resolution.

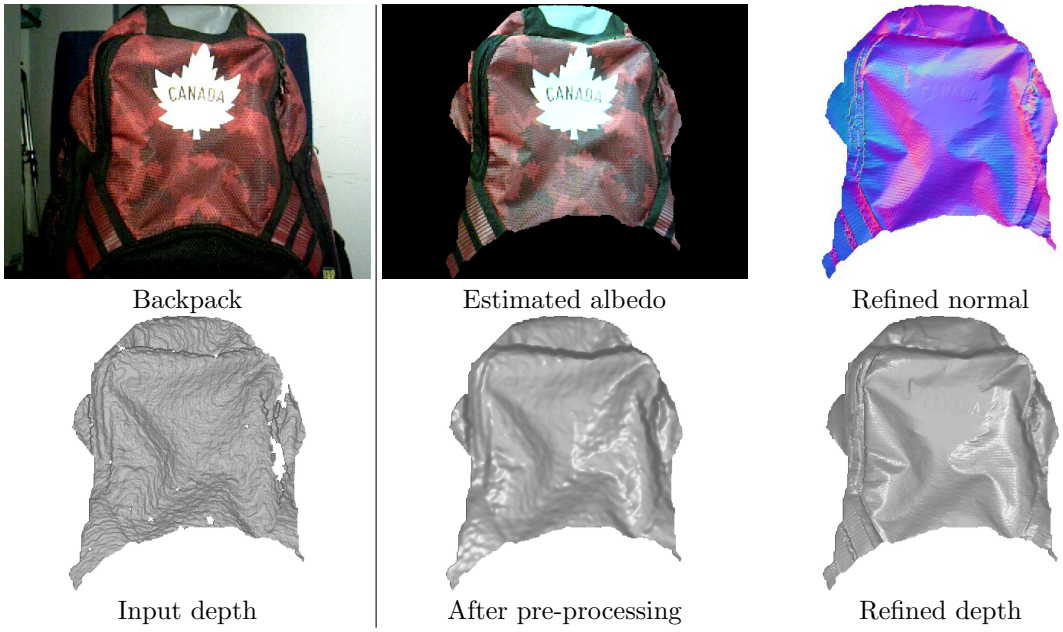


Figure 3.10: Illustrations for our proposed robust multi-light method. Here $n = 10$ images with various lighting conditions have been used, one of which is the top left RGB image. Fine details on the surface of the backpack are recovered without any artefacts suffered by other methods.

In this section, we will present our approach to the combination of the robust multi-light method and super-resolution, with the help of which we will provide decent high-quality and high-resolution depth maps.

The scale factor between the RGB and depth image is around 2 for ASUS XtionPro Live, so we can at least enlarge our map twice larger than the original size, which means the refined depth resolution will be 1280×960. And we assume that the input depth image has been registered well (done easily with OpenNI) such that the upsampled depth Z is aligned with the large RGB image after simple interpolation.

Provided the acquired small depth and the super-resolution depth are denoted as z and Z respectively, a standard single depth image super-resolution problem can be represented like in [52]:

$$z = KFZ \tag{3.42}$$

where K represents a downsampling operator [53] and F a blurring filter. For the sake of simplicity, we do not consider the blurring filter F but only the simple downsampling operator K. If z and Z are vectorized, K is a 4-connected isotropic downsampling matrix.

In the light and albedo estimation part for the depth super-resolution, the energy in Eq. 3.32

and Eq. 3.33 are directly used. When we build $\mathbf{A}_{\mathbf{s}_c}$ and $\mathbf{A}_{\rho_c}$, the surface normal $\mathbf{n}(z)$ is replaced with $\mathbf{N}(Z)$ which is the normal of the large depth. However, the depth enhancement part (Eq. 3.43) should be adapted to the super-resolution framework. As we know, calculating Z from the super-resolution equation in Eq. 3.42 is ill-posed so our shading term can be treated as the regularization term. Now, Z is again used to replace z during the construction of $\mathbf{A}_z$ and $\mathbf{b}_z$, which are now denoted by $\mathbf{A}_Z$ and $\mathbf{b}_Z$.

With the input small depth $z_0 \in \mathbb{R}^m$ and a downsampling kernel $K \in \mathbb{R}^{m \times M}$ where m and M represent the number of the pixel within the small mask and the big mask respectively, the super-resolution depth refinement energy now is changed to:

$$\min_Z \|\mathbf{A}_Z Z - \mathbf{b}_Z\|_2^2 + \lambda_z \|K \cdot Z - z_0\|_2^2 \tag{3.43}$$

After optimizing this energy within the framework of the robust multi-light method, we will acquire the super-resolution version refined depth, which has been illustrated in Fig. 3.11.

(a) One input image (960 × 1280) (b) Estimated albedo (960 × 1280) (c) Estimated normal(960 × 1280)

(d) Input depth image (480 × 640) (e) Smoothed depth image (480 × 640) (f) Refined super-resolution depth (960 × 1280)

Figure 3.11: Results of the super-resolution depth of a paper bag using our robust multi-light method. Input depth size is 480 × 640, and the refined depth's is 960 × 1280.

Chapter 4

Results and Evaluation

Parameter setup First of all, we need to specify the parameters we used throughout the whole evaluation part. The default parameters are used for the RGBD-Fusion method, which has 8 in total. It should be mentioned that, since both proposed methods don't have a smoothness term for the depth enhancement, λ_z^2 in RGBD-Fusion and λ_l in our RGBD-Fusion Like method are set to 0 for the sake of fairness of comparison. λ_z^2 and λ_l are set to the values in table 4.1 only during the quantitative evalution because we want to illustrate the importance of this depth smoothness term for the RGBD-Fusion method.

Table 4.1: Parameters of all the methods used throughout all the experiments.

Method	Total number	Parameters
RGBD-Fusion [11]	8	$\lambda_\rho = 0.1, \lambda_\beta^1 = 0.1, \lambda_\beta^2 = 0.1, \tau = 0.05, \sigma_c = \sqrt{0.05}, \sigma_d = \sqrt{50}, \lambda_z^1 = 0.004, \lambda_z^2 = 0.0075$
RGBD-Fusion Like method (Eq. 3.9)	5	$\lambda_\rho = 10, \sigma_I = \sqrt{0.05}, \sigma_z = \sqrt{50}, \lambda_z = 500, \lambda_l = 2$
Proposed I: RGB Ratio model (Eq. 3.20)	4	$\lambda_\rho^1 = 10^{15}, \lambda_\rho^2 = 10^{13}, \sigma_c = 100, \lambda_z = 100$
Proposed II: Robust Multi-Light (Eq. 3.31)	1	$\lambda_z = 100$

4.1 Quantitative Evaluation

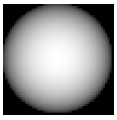
In the following, we will first explain how we generate the synthetic dataset. Then, we will quantitatively compare the refined depths from both proposed methods and the RGBD-Fusion [11] under various cases of the albedo. The runtime will be discussed at the end of this section.

4.1.1 Synthetic data generation

In order to quantitatively validate the performance of our proposed methods and our implementation of the RGBD-Fusion, we use the well-known "The Joyful Yell" dataset with 3 point light sources and ambient lights.

To simulate the natural scene illumination, we set the frontal directions for red, green and blue lightings so the first-order SH parameters (3D position of the light source + ambient light) modelled as:

$$\mathbf{s}_R = \mathbf{s}_G = \mathbf{s}_B = \begin{bmatrix} 0 & 0 & -1 & 0.2 \end{bmatrix}^\top$$

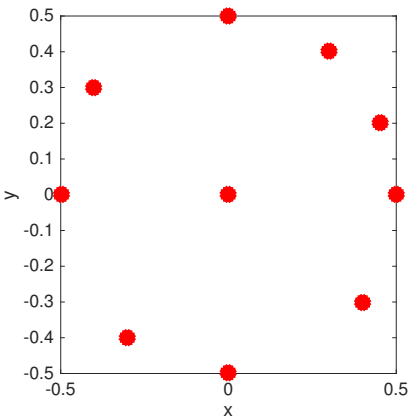


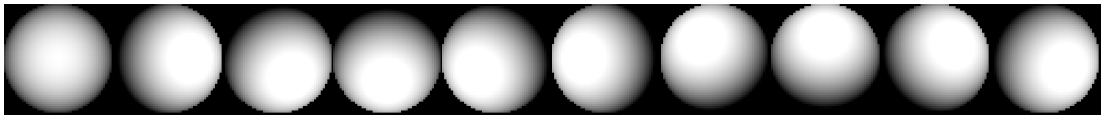
The ball on the right represents the light direction. Then, in order to reproduce the LED configuration for the proposed RGB ratio model, we define the 3 lighting directions as:

$$\mathbf{s}_R = \begin{bmatrix} 0 & 0 & -1 & 0.15 \end{bmatrix}^\top$$
$$\mathbf{s}_G = \begin{bmatrix} 0.3 & 0.2 & -1 & 0.25 \end{bmatrix}^\top$$
$$\mathbf{s}_B = \begin{bmatrix} -0.2 & 0.3 & -1 & 0.2 \end{bmatrix}^\top$$

Finally, we need to create a sequence of same images with various directional lights for our robust multi-light model. A "lighting" matrix L and the corresponding 2D positions (omit the third dimension because they are the same) of 10 point light sources can be illustrated as below, where red points represent the light positions.

$$L = \begin{pmatrix} 0.5 & 0 & -1 & 0.2 \\ 0.3 & 0.4 & -1 & 0.2 \\ 0 & 0.5 & -1 & 0.2 \\ -0.4 & 0.3 & -1 & 0.2 \\ -0.5 & 0 & -1 & 0.2 \\ -0.3 & -0.4 & -1 & 0.2 \\ 0 & -0.5 & -1 & 0.2 \\ 0.4 & -0.3 & -1 & 0.2 \\ 0 & 0 & -1 & 0.2 \\ 0.45 & 0.2 & -1 & 0.2 \end{pmatrix}^\top$$





After defining the lighting setup for various methods, three various albedo scenarios ranging from simple to complicated are listed below:

- Red, green and blue piecewise constant areas
- Colorful patterns with a few small details inside[1]
- Colorful patterns with complicated details[2]

With the albedos and all the pre-defined lights, we can use the Lambertian model mentioned in section 2 to create the synthetic color images like the first row in Fig. 4.2.

4.1.2 Results Accuracy

Metrics Two metrics have been defined to quantitatively evaluate the performance of depth refinement: root mean square error (RMSE) between ground truth and the estimated depth, and mean angular error (MAE) between the ground truth and estimated normal directions. We input the rough depth (Fig. 4.1(a)) and obtain the refined depth with various methods and then compare these two metrics with the ground truth, which is shown in Fig. 4.1(b).

Let z, z_g, N, N_g be the input and the ground truth depth as well as their corresponding normals respectively, m the total number of pixels inside the given mask $\mathcal{M}$ and i the index inside the mask, the loose definitions for RMSE and MAE are:

$$e_{RMSE} = \sqrt{\frac{\sum_{i}^{m} (z(i) - z_g(i))^2}{m}} \tag{4.1}$$

$$e_{MAE} = \frac{\sum_{i}^{m} \arccos(N(i) \cdot N_g(i))}{m} \tag{4.2}$$

The RMSE reflects the global quality of the refined depth (low frequency), while the MAE assesses the precision of the recovered depth (high frequency).

According to table 4.1, 4.2 and Fig. 4.2, there are some interesting observations:

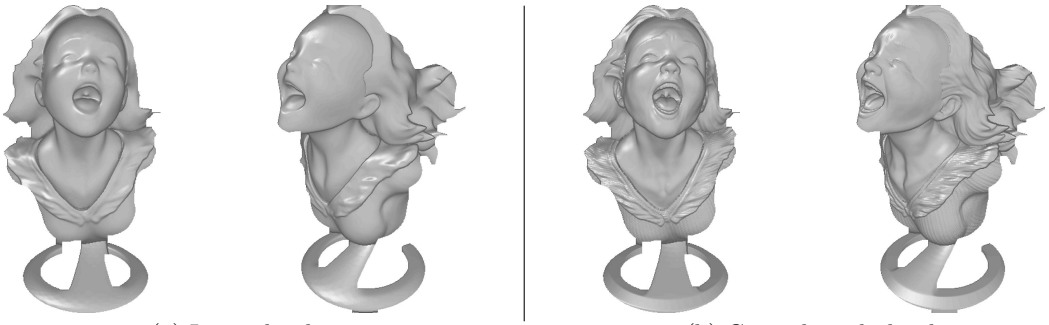
- Our RGBD-Fusion Like method uses fewer parameters (5 against 8) and less runtime (7s against 21s) than RGBD-Fusion [11] but achieves almost the same accuracy.
- Adding the Laplacian smoothness term in the depth enhancement energy of RGBD-Fusion method makes a huge improvement on the refined results. In contrast, both our proposed methods have no smoothness term but provide equal or better results.

[1] EBSD map. Image courtesy of https://mtex-toolbox.github.io/files/doc/EBSDSpatialPlots.html
[2] 1000 Visual Mashups. Image courtsesy of https://www.flickr.com/photos/qthomasbower/3470650293

Table 4.2: Quantitative evaluations among 4 methods. RMSE and MAE are in pixels and degrees respectively. "No smooth" means no laplacian smoothness term in depth enhancement.

	Simple RGB		Pattern		Complicated	
Method	RMSE	MAE	RMSE	MAE	RMSE	MAE
Input reference	3.3305	16.3096	3.3305	16.3096	3.3305	16.3096
RGBD-Fusion [11] (no smooth)	3.3418	18.9115	3.3872	27.0026	3.3411	25.6574
RGBD-Fusion [11]	3.1751	17.2197	3.1890	18.4722	3.1708	18.0850
Fusion-Like (no smooth)	3.3475	17.5911	3.3459	23.4808	3.3898	35.2610
Fusion-Like	2.8700	17.1776	2.8749	17.7302	2.8848	19.6452
RGB ratio model	**1.9437**	5.0574	2.9116	17.5238	3.1006	21.2286
Robust multi-light model	2.3125	**3.8708**	**1.5794**	**1.7368**	**1.8424**	**2.6815**

- Single depth image refinement methods (RGBD-Fusion and RGB ratio model) have a chance to acquire satisfactory results only when the albedo is simple with several big color patches. However, they will fail and give even worse results than the input depth in terms of RMSE and MAE when the albedos get complicated. Most of the small details on the albedo of "Pattern" and "Complicate Pattern" cannot be acquired by these methods, which yield the wrong depth estimation with artefacts. This is due to the fact that their albedo estimations highly rely on the regularization terms which prefer piecewise smoothness, but this does not meet the condition of most real-world objects.
- It can be effortlessly noticed that our robust multi-light method has a strong ability to handle the cases with extremely complicated albedo. Instead of using any regularization terms, our method use only one shading term (Eq. 3.33) to estimate the albedo with extra images illuminated from various light directions. Compared to the albedo estimated by other methods, the albedo from our multi-light method could recover most of the details.



(a) Input depth (b) Ground truth depth

Figure 4.1: The input depth and the ground truth depth from two views. We only use the cases of frontal direction for the quantitative evaluation, which are the first and the third images.

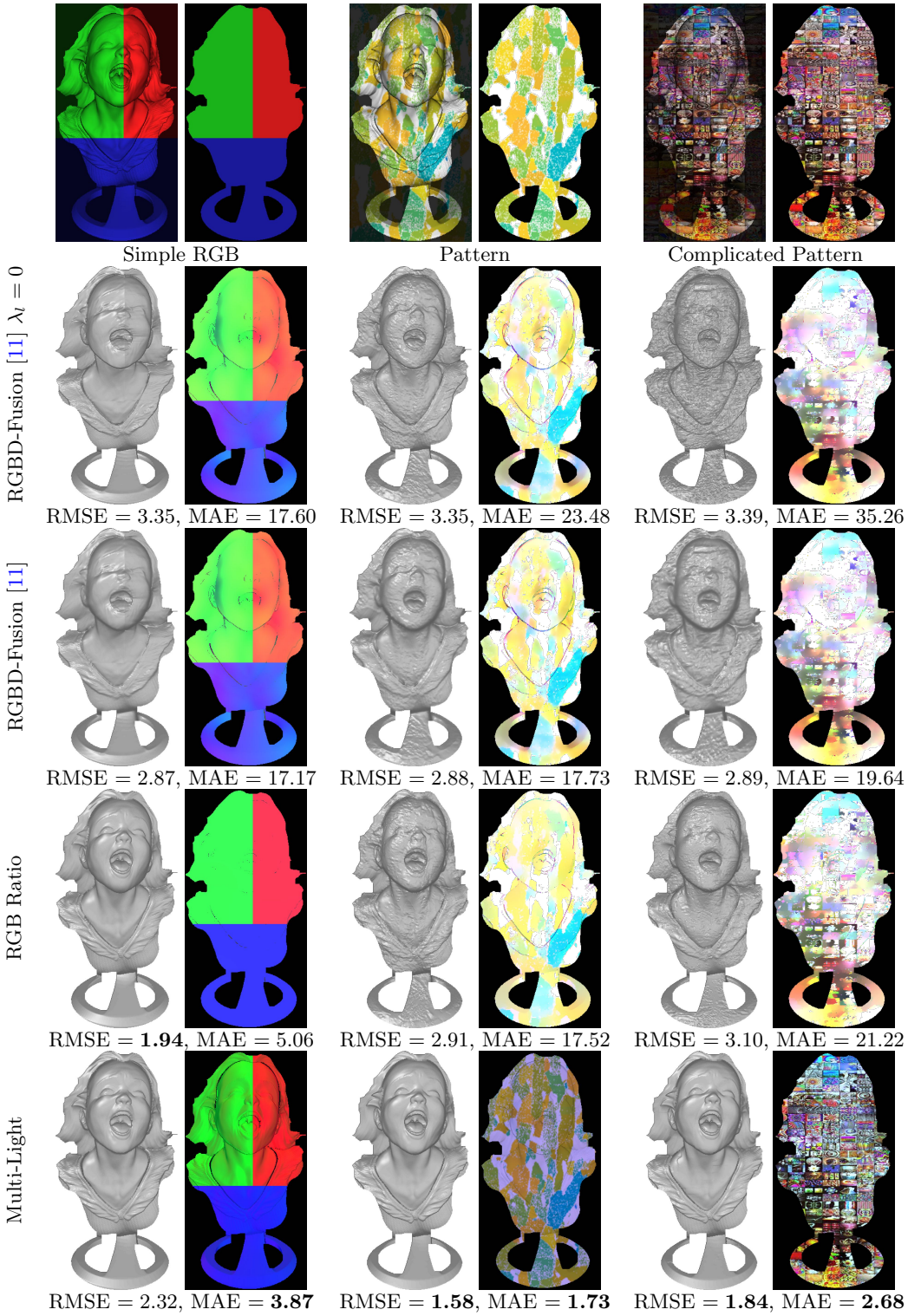


Figure 4.2: Evaluation of our two proposed methods RGB ratio model and robust multi-light method against the RGBD-Fusion [11], under three albedos scenarios from simple to complicated. The first row is the input color images and their ground truth albedos, while the rest are the estimated depths and albedos using the parameters defined in table 4.1. The errors for the rough input depth are RMSE of 3.33 and MAE of 16.30. Our proposed methods can deal with the complicated albedo and outperform RGBD-Fusion in all tests.

4.1.3 Runtime

First of all, all the tests are performed in MATLAB R2015b under Mac OSX 10.10.5, Intel Core i5 2.7GHz, 2 cores, 8GB memory. The resolution of our synthetic image is 540×960.

We first compare the runtime for 4 methods, which is shown in table 4.3. It is noticeable that our implementation of RGBD-Fusion like method is much faster than the original approach while the accuracy from our implementation is quite similar to the original one. On the one hand, we only consider the pixel inside the mask while the official implementation[3] uses the whole image. On the other hand, instead of estimating the ambient light for each pixel with an extra energy, we directly treat the ambient light as one parameter inside the first-order spherical harmonics so the ambient light could be obtained along with lighting directions.

It should be mentioned that both our proposed approaches are indeed about twice slower than RGBD-Fusion method. However, the RGBD-Fusion method stops when the overall energy starts increasing, which happens merely in $1-3$ iterations based on our experiment. In contrast, the minimization for our methods are both convergent so the runtime highly relies on the threshold for the relative change of the energy values.

Table 4.3: The comparison of the runtime among RGBD-Fusion method, our implementation RGBD-Fusion Like method, proposed RGB ratio model and robust multi-light method in synthetic data.

Method	Runtime (s)
RGBD-Fusion [11]	21.64
RGBD-Fusion Like	7.75
Proposed I: RGB Ratio	49.33
Proposed II: Multi-Light	52.82

Now for the proposed robust multi-light method, we are interested in understanding how the number of different illuminations makes the difference in the runtime as well as the RMSE and MAE. To perform the experiment, we randomly pre-defined 40 illumination directions and constructed the corresponding synthetic color images as discussed in section 4.1.1.

It has been shown in Fig. 4.3 that when the number of images increases, the runtime for each iteration has linearly ascent, while two errors decrease and reach a platform. This is reasonable because the details on a certain part of the object can be retrieved when there exists light on it. Therefore, the more various lightings we have, the more details we can obtain. If we take the comprehensive consideration, $10 \sim 20$ would be the suitable number for illuminations.

[3]Source code of official RGBD-Fusion implementation can be found: https://cs.technion.ac.il/~royorel/rgbd_fusion.html

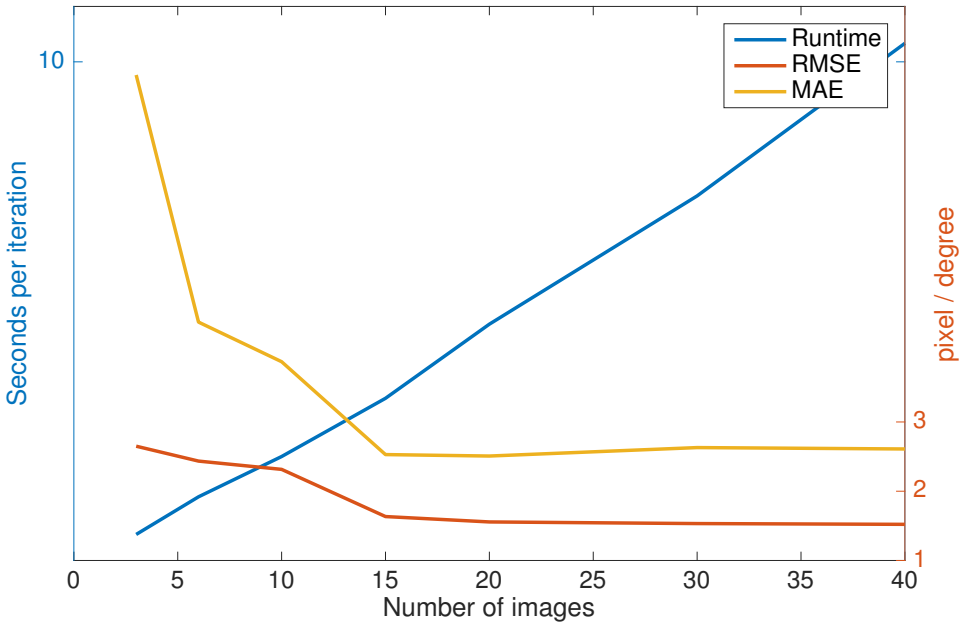


Figure 4.3: Illustrations for the runtime, RMSE and MAE in various number of illuminations for the proposed robust multi-light method. $10 \sim 20$ is the suitable number for different lightings.

4.2 Qualitative Evaluation

Except for the quantitative evaluation, we have found out that our robust multi-light method also outperforms the state-of-the-art methods qualitatively in many aspects. In this part, we will first show the robustness of depth estimation of the proposed multi-light method on the objects with complicated albedo. Then, we will demonstrate the performance on the specular (non-Lambertian) objects. In addition, as mentioned in chapter 2, uncalibrated photometric stereo suffers from the GBR ambiguity. So we will finally exemplify that our method also outperforms other photometric stereo algorithms because of the input depth cues.

4.2.1 Complicated albedo objects

It has already been illustrated in Fig. 4.2 that our robust multi-light method has the capability of acquiring the shape of synthetic objects with very complex albedo. In contrast, the state-of-the-art depth enhancement methods with one image work well for many simple color objects, but could not separate the complicated albedo from the real shape of the object, which leads to severe artefacts on the final estimated depth. Figure 4.4 has proved that this judgment

still holds correct for the real-world objects. To be fair, we set the parameter for Laplacian smoothness term in RGBD-Fusion to zero.

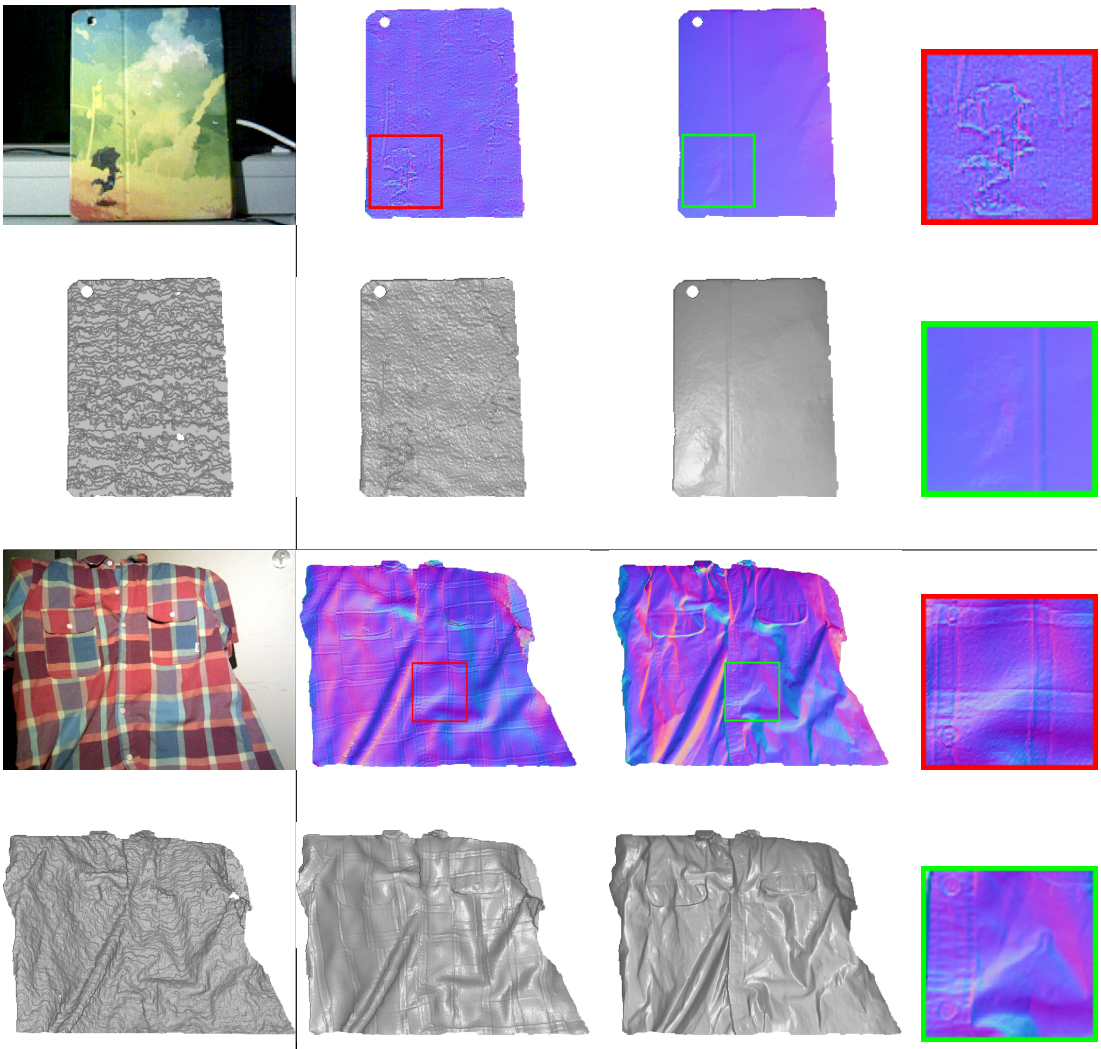


Figure 4.4: Comparisons between our multi-light model and RGBD-Fusion for two specular objects. On the first column, the RGB images of the iPad cover and the shirt are ones of the 10 various illuminations. The first and third rows correspond to the surface normals, while the second and fourth are the refined depths. Our method can correctly estimate the surface (normals) when structural patterns (but no depth variation) exist, while the depths from RGBD-Fusion contains visible artefacts.

As we can notice, the first example in Fig. 4.4 is a flat colorful iPad cover with an underlying narrow slot on it. The depth and the surface normal recovered from RGBD-Fusion method contain quantities of wrong details from the albedo at the same time that the slot almost disappears. On the contrary, our method has successfully refined the flat surface of the cover with the visible slot. Similarly, in the second example, there are no depth differences among the patterns on the colorful shirt, yet the RGBD-Fusion method could not manage to figure out this fact. Again, our method is able to get rid of nearly all the patterns and recover the real shape of the shirt.

4.2.2 Specular (non-Lambertian) objects

Again, we also compare our multi-light method with the state-of-the-art depth refinement approach [11]. Same as the last section, we turned off the Laplacian smoothness term in their method for the sake of fairness.

The Lambertian reflectance model is built up based on the Lambert's cosine law which assumes objects are diffused. It should not work with the non-Lambertian objects theoretically. Indeed, the shapes recovered from the RGBD-Fusion method have very obvious artefacts in the specular areas, as we can notice in Fig. 4.6. This is because the specular areas are "light polluted" and no color or shape details can be retrieved. However, it turns out that our multi-light approach can still work well and recover the real shapes with the presence of specularity, although the Lambertian model is also applied.

(a) One illumination (b) The other illumination (c) Estimated albedo with multi-light method

Figure 4.5: Demonstrations for the albedo remedy of specularity of a paper bag. The first two images are among the 10 various illumination conditions. The third image represents the albedo estimated by our multi-light method. We can clearly see the specular parts in the images appear different from other parts in the albedo, which is the result of the remedy of specularity.

To explain the reason why our method is working with the specular objects, we first assume that 10 input color images are given. We know the specularity in each image differs from the other 9 images owing to the fact that we sway the active LED light. This means, the specularity

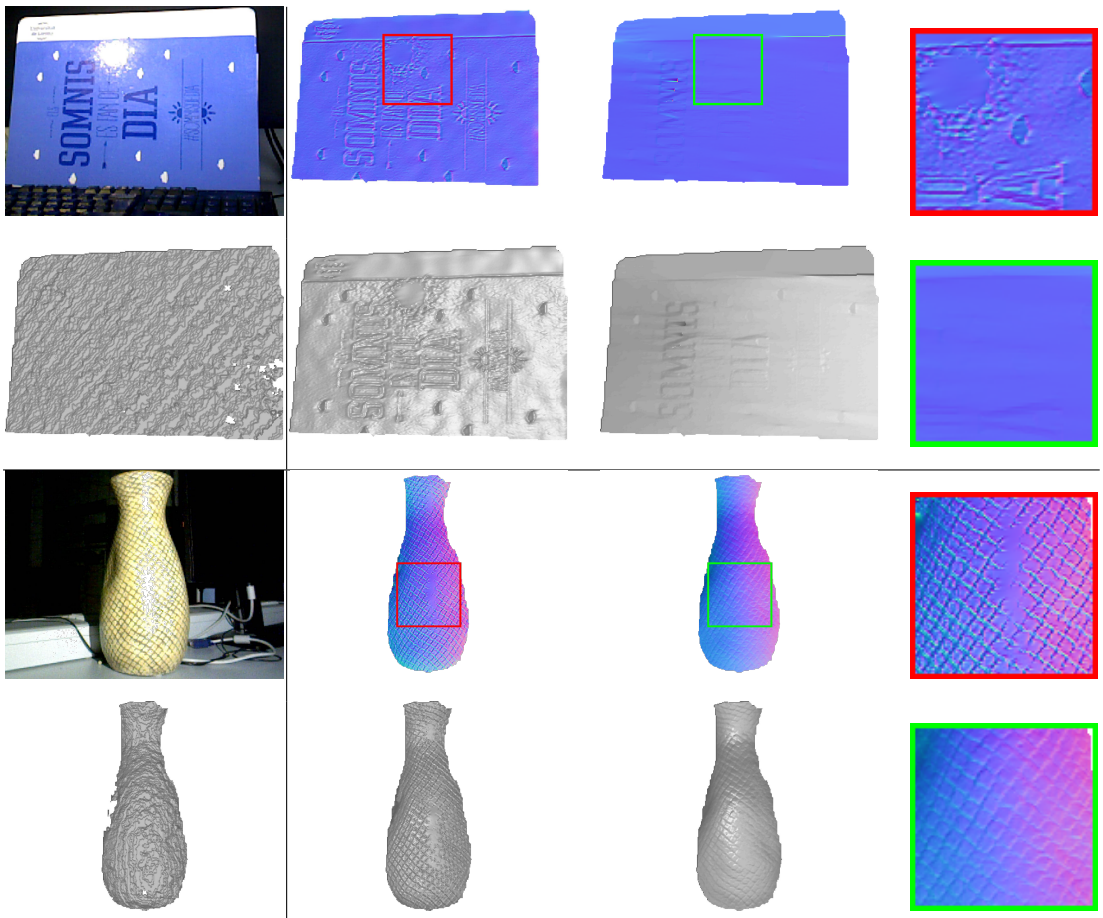


Figure 4.6: Comparisons between our multi-light method and RGBD-Fusion for two specular objects. The RGB images in the first column are among 10 various illuminations. The first and third rows correspond to the surface normals , while the second and fourth are the refined depths. We can notice the RGBD-Fusion method has strong artefacts on the refined depth in the specular part, while our method can still correctly acquire all the correct details under the specularity.

appearing in a certain image has a high probability not to be specular again in the other 9. Meanwhile, the albedo and depth refinement of our algorithm use all 10 images instead of just the specular one, hence, the rest 9 images under the least squares would remedy the specularity in that area. An example to illustrate the compensation effect on the albedo of the specularity is shown in Fig. 4.5.

4.2.3 Comparison with photometric stereo method

Also, we would like to show the advantages of our method over other uncalibrated photometric stereo methods to which no rough depth is given as the input. Here we compare our robust multi-light method with a state-of-the-art PS method [54] from Favaro and Papadhimitri [4]. Here we call their method LDR-PS.

Figure 4.7: Comparison of the recovered depth between our multi-light method and the uncalibrated photometric stereo method LDR-PS on the synthetic dataset. The refined depth from our method is almost the same as the ground truth, while LDR-PS suffers from the GBR ambiguity, especially on the neck and pedestal of the statue.

First, we make the comparison on the synthetic dataset of colorful patterns from the last section. It is noticeable in Fig. 4.7 that the depth enhanced with our method is almost the same as the ground truth. The result acquired from LDR-PS seems to have a similar appearance with the ground truth from the frontal direction (it looks darker because the estimated depth from LDR-PS method is not always correct). However, if we rotate the reconstructed map to the side, we can notice that the discontinuity between the head and body has been over-smoothed and the pedestal is distorted. This is the so-called generalized bas-relief ambiguity.

The same problem happens in the real data as well (Fig. 4.8). Both methods hold almost identical frontal view but the depth in the side view acquired from LDR-PS turns out to have a wrong interpretation.

[4]The LDR-PS code can be obtained from author's website http://www.cvg.unibe.ch/tpapadhimitri/

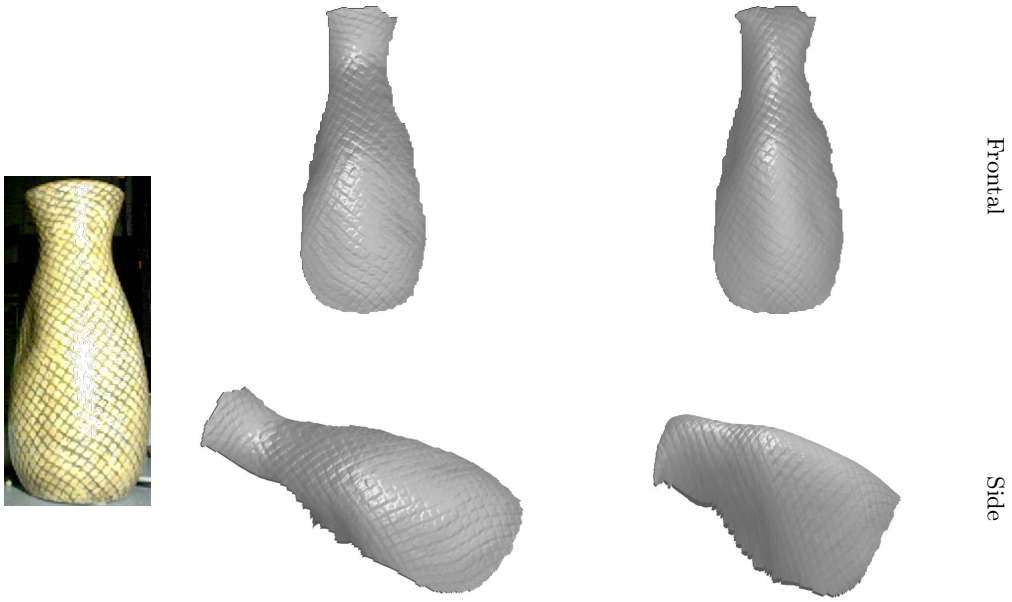


Figure 4.8: Comparison of the recovered depth between our multi-light method and the uncalibrated photometric stereo method LDR-PS on the real-world vase. Our method is able to recover the real shape, while the LDR-PS could not acquire the correct shape because of the GBR ambiguity.

Chapter 5

Conclusions and Future Work

In this thesis, we have developed two new depth refinement algorithms which significantly enhance the quality of the noisy and coarse depth images from consumer RGB-D cameras. The proposed RGB ratio model utilizes red, green and blue LEDs as the active lights and iteratively refines the depth map with the ratio Lambertian models for every pair of channels of the input RGB image. This method requires only one color image and resolves the nonlinearity in the inverse problem. Nevertheless, the method may be limited by the size of indoor environment because the active lights should be set as far away from each other as possible. Also, similar to other depth enhancement methods based on one single image, it can merely handle the objects with constant or simple albedo. The recovered depths may contain artefacts because this type of methods has difficulty in estimating the complicated albedo. Therefore, we present another robust multi-light method which is able to handle such issue.

The robust multi-light method uses multiple images, in which the object is illuminated from different directions, to jointly estimate the depth, albedo as well as lighting conditions. The robustness of this method is achieved by the capacity of recovering the complicated albedo and real shape without any regularization term. Unlike most of the previous methods which have a number of tuning parameters, we only have one fixed parameter which works well in almost all cases. Moreover, this method has been integrated with the image super-resolution scheme such that a high-quality and high-resolution refined depth map can be obtained. We believe this is the first depth image super-resolution approach based on photometric stereo. To conclude, our robust multi-light method shows the potential of high-resolution 3D reconstruction from an affordable RGB-D camera.

There are various possible directions for the future work on our depth or shape refinement research:

- It has been shown in [55] that the noise level of depth acquisition from low-cost depth sensors grows quadratically with respect to the increasing distance. Hence, the refined depth should be theoretically more accurate if every depth pixel is weighted according to the corresponding measurement noise.

- As we mentioned, the depths for complicated objects refined by single-image based methods contain artefacts, due to the fact that the designed constraints on the albedo are not practical so the estimated depth may be affected by the inaccurate albedo. For instance, many approaches use anisotropic Laplacian to impose the piecewise smoothness on the albedo, which is not commonly correct for the real-world objects. Hence, provided we can use a more realistic regularization term, the estimated albedo and the refined depth are supposed to be better. Recently some researchers have proposed a general framework to jointly deal with many classic computer vision tasks like deblurring, demosaicing and denoising which are normally modelled with $\|Ax - b\|^2 + R(x)$. Instead of using certain regularizations like TV, they separate the equation using the Primal-Dual or ADMM scheme and directly solve the proximal operator of the arbitrary regularizer R with a BM3D denoiser [56] or a deep denoising neural network [57]. It would be very interesting if we could use such scheme to estimate the albedo.

- The existing 3D object reconstruction/modelling methods are subject to low-resolution and bad-quality depths. It is promising to integrate them with our shading-based depth super-resolution method, which will potentially improve the reconstruction accuracy.

- As with other methods, one prerequisite of our methods is that the depth image needs to be registered to the RGB image. It is possible to integrate the depth image registration within the refinement framework implicitly such that we can directly acquire depth and color images from the RGB-D sensors without any other third-party software like OpenNI.

Appendix A

List of Notations

Table A.1: List of notations

Notation	Description
I	Input image
ρ	Albedo
z	Depth image
Z	Super-resolution depth image
$\mathbf{l}$	Light direction in $\mathbb{R}^3$
φ	Ambient light parameter
$\mathbf{s}$	First-order Spherical Harmonics coefficients in $\mathbb{R}^4$
$\mathbf{n}$	Surface normal
$\mathcal{M}$	Binary mask
m	Number of pixels inside mask $\mathcal{M}$
n	Number of different illuminations
$\mathbf{I}$	Stack of vectorized images
$\mathcal{P}$	Stack of vectorized RGB albedos
∇	Gradient operator
Δ	Laplacian operator
c	a certain channel, $c \in \{R, G, B\}$
(x, y)	pixel coordinate

Appendix B

Derivation of Matrix Ψ in RGB Ratio Model

In the depth refinement part of RGB ratio model, we can acquire the Eq. 3.25, which we re-write it here:

$$\begin{aligned}
\rho_G(I_R - \rho_R\varphi_R)\mathbf{l}_G^\top\mathbf{n} - \rho_R(I_G - \rho_G\varphi_G)\mathbf{l}_R^\top\mathbf{n} = 0 \\
\rho_B(I_G - \rho_G\varphi_G)\mathbf{l}_B^\top\mathbf{n} - \rho_G(I_B - \rho_B\varphi_B)\mathbf{l}_G^\top\mathbf{n} = 0 \\
\rho_R(I_B - \rho_B\varphi_B)\mathbf{l}_R^\top\mathbf{n} - \rho_B(I_R - \rho_R\varphi_R)\mathbf{l}_B^\top\mathbf{n} = 0
\end{aligned} \tag{B.1}$$

Here we will explain how to derive $\Psi z = 0$ from Eq. 3.25.

Since the surface normal now is represented using perspective projection with the denominator eliminated, the normal $\mathbf{n} \in \mathbb{R}^3$ in each pixel can be written as:

$$\mathbf{n} = \begin{pmatrix} fz_x \\ fz_y \\ -z - \tilde{x}z_x - \tilde{y}z_y \end{pmatrix} \tag{B.2}$$

z_x and z_y are the first-order derivative of depth $z \in \mathbb{R}^m$ in x and y directions, f is the focal length and $\tilde{x} = x - x_0,\ \tilde{y} = y - y_0$, with (x_0, y_0) the coordinates of principle points and (x, y) the coordinate of a pixel inside mask $\mathcal{M}$. Let us consider only the first equation (ratio model between the red and green channels) in Eq. B.1 and denote:

$$\begin{aligned}
p_{GR} &= \rho_G(I_R - \rho_R\varphi_R) \\
p_{RG} &= \rho_R(I_G - \rho_G\varphi_G)
\end{aligned} \tag{B.3}$$

After putting Eq. B.2 into Eq. B.1 and expanding the equation, we acquire:

$$\begin{aligned}&(p_{GR}\mathbf{l}_G^1 f - p_{RG}\mathbf{l}_R^1 f + p_{RG}\mathbf{l}_R^3\tilde{x} - p_{GR}\mathbf{l}_G^3\tilde{x})z_x\\ +&(p_{GR}\mathbf{l}_G^2 f - p_{RG}\mathbf{l}_R^2 f + p_{RG}\mathbf{l}_R^3\tilde{y} - p_{GR}\mathbf{l}_G^3\tilde{y})z_y\\ +&(p_{RG}\mathbf{l}_R^3 - p_{GR}\mathbf{l}_G^3)z = 0\end{aligned} \tag{B.4}$$

Then if we consider all the pixels inside the mask $\mathcal{M}$ and we have the gradient matrices $D_x \in \mathbb{R}^{m\times m}$ and $D_y \in \mathbb{R}^{m\times m}$ in x and y directions. Now assuming $z, \rho, I \in \mathbb{R}^m$, and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ represent all the coordinates inside the mask, we first have four diagonal matrices:

$$\begin{aligned}P_{GR} &= \mathrm{diag}\big(\rho_G \cdot (I_R - \rho_R\varphi_R)\big)\\ P_{RG} &= \mathrm{diag}\big(\rho_R \cdot (I_G - \rho_G\varphi_G)\big)\\ \tilde{X} &= \mathrm{diag}\big(\mathbf{x} - x_0\big)\\ \tilde{Y} &= \mathrm{diag}\big(\mathbf{y} - y_0\big)\end{aligned} \tag{B.5}$$

The mark $\cdot$ represents element-wise multiplications between two vectors. Eq. B.4 becomes:

$$\begin{aligned}&(P_{GR}\mathbf{l}_G^1 f D_x - P_{RG}\mathbf{l}_R^1 f D_x + P_{RG}\mathbf{l}_R^3\tilde{X}D_x - P_{GR}\mathbf{l}_G^3\tilde{X}D_x)z\\ +&(P_{GR}\mathbf{l}_G^2 f D_y - P_{RG}\mathbf{l}_R^2 f D_y + P_{RG}\mathbf{l}_R^3\tilde{Y}D_y - P_{GR}\mathbf{l}_G^3\tilde{Y}D_y)z\\ +&(P_{RG}\mathbf{l}_R^3 - P_{GR}\mathbf{l}_G^3)z = 0\end{aligned} \tag{B.6}$$

After re-arranging, we have the simplified Eq. B.6:

$$(P_{GB}L_G - P_{RG}L_R)z = 0 \tag{B.7}$$

Then, we can similarly acquire such equationm, and $P_{GB}, P_{BG}, P_{BR}, P_{RB}$ for other two ratio models in Eq. B.1, which can be denoted as belows. Finally, $\Psi \in \mathbb{R}^{3m\times m}$ is acquired.

$$\begin{aligned}&(P_{GR}L_G - P_{RG}L_R)z = 0\\ &(P_{BG}L_B - P_{GB}L_G)z = 0 \quad \Rightarrow \Psi z = 0\\ &(P_{RB}L_R - P_{BR}L_B)z = 0\end{aligned} \tag{B.8}$$

where

$$\begin{aligned}L_R &= (\mathbf{l}_R^1 f D_x + \mathbf{l}_R^2 f D_y - \mathbf{l}_R^3 - \mathbf{l}_R^3\tilde{X}D_x - \mathbf{l}_R^3\tilde{Y}D_y)\\ L_G &= (\mathbf{l}_G^1 f D_x + \mathbf{l}_G^2 f D_y - \mathbf{l}_G^3 - \mathbf{l}_G^3\tilde{X}D_x - \mathbf{l}_G^3\tilde{Y}D_y)\\ L_B &= (\mathbf{l}_B^1 f D_x + \mathbf{l}_B^2 f D_y - \mathbf{l}_B^3 - \mathbf{l}_B^3\tilde{X}D_x - \mathbf{l}_B^3\tilde{Y}D_y)\end{aligned} \tag{B.9}$$

Bibliography

[1] Jürgen Sturm, Erik Bylow, Fredrik Kahl, and Daniel Cremers. CopyMe3D: Scanning and printing persons in 3D. In *German Conference on Pattern Recognition*, pages 405–414. Springer, 2013.

[2] Robert Maier, Jürgen Sturm, and Daniel Cremers. Submap-based bundle adjustment for 3d reconstruction from rgb-d data. In *German Conference on Pattern Recognition*, pages 54–65. Springer, 2014.

[3] Licong Zhang, Jürgen Sturm, Daniel Cremers, and Dongheui Lee. Real-time human motion tracking using multiple depth cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2389–2395. IEEE, 2012.

[4] Nikolas Engelhard, Felix Endres, Jürgen Hess, Jürgen Sturm, and Wolfram Burgard. Real-time 3d visual slam with a hand-held rgb-d camera. In *Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden*, volume 180, pages 1–15, 2011.

[5] Christian Kerl, Jurgen Sturm, and Daniel Cremers. Dense visual slam for rgb-d cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2100–2106. IEEE, 2013.

[6] Yudeog Han, Joon-Young Lee, and In So Kweon. High quality shape from a single rgb-d image under uncalibrated natural illumination. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1617–1624, 2013.

[7] Robert Maier. Out-of-core bundle adjustment for 3d workpiece reconstruction. Master's thesis, Technische Universität München, Germany, September 2013.

[8] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgib-

bon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE international Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136. IEEE, 2011.

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

[10] Roy Or-El, Rom Hershkovitz, Aaron Wetzler, Guy Rosman, Alfred M Bruckstein, and Ron Kimmel. Real-time depth refinement for specular objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4378–4386, 2016.

[11] Roy Or-El, Guy Rosman, Aaron Wetzler, Ron Kimmel, and Alfred M Bruckstein. Rgbd-fusion: Real-time high precision depth recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5407–5416, 2015.

[12] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1108–1115. IEEE, 2011.

[13] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1161–1168, 2013.

[14] Mohammadul Haque, Avishek Chatterjee, Venu Madhav Govindu, et al. High quality photometric reconstruction using a depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2275–2282, 2014.

[15] Yvain Quéau, Jean Mélou, Jean-Denis Durou, and Daniel Cremers. Dense multi-view 3d-reconstruction without dense correspondences. *arXiv preprint arXiv:1704.00337*, 2017.

[16] Kichang Kim, Akihiko Torii, and Masatoshi Okutomi. Joint estimation of depth, reflectance and illumination for depth refinement. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[17] Albert S. Huang, Abraham Bachrach, Peter Henry, Michael Krainin, Daniel Maturana, Dieter Fox, and Nicholas Roy. Visual odometry and mapping for autonomous flight using an rgb-d camera. In *Int. Symposium on Robotics Research (ISRR)*, Flagstaff, Arizona, USA, Aug. 2011.

[18] Christian Kerl. Odometry from rgb-d cameras for autonomous quadrocopters. Master's thesis, Technical University Munich, Germany, Nov. 2012.

[19] Radu Horaud. A short tutorial on three-dimensional cameras, 2013. http://perception.inrialpes.fr/~Horaud/Courses/pdf/Horaud_3Dcameras_tutorial.pdf.

[20] ASUS. Xtion pro live. https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/. Accessed: 2017-05-15.

[21] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970.

[22] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015.

[23] Edward H Adelson and Alex P Pentland. The perception of shading and reflectance. *Perception as Bayesian inference*, pages 409–423, 1996.

[24] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2335–2342. IEEE, 2009.

[25] Klett, Eberhard Witwe, Detleffsen, Christoph Peter, et al. *IH Lambert... Photometria sive de mensura et gradibus luminis, colorum et umbrae*. sumptibus viduae Eberhardi Klett, 1760.

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

[27] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA A*, 18(10):2448–2459, 2001.

[28] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[29] Berthold KP Horn and Michael J Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986.

[30] Robert T. Frankot and Rama Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):439–451, 1988.

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

[32] David A Forsyth and Jean Ponce. A modern approach. *Computer vision: a modern approach*, pages 88–101, 2003.

[33] Carlos Hernández, George Vogiatzis, and Roberto Cipolla. Overcoming shadows in 3-source photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):419–426, 2011.

[34] Hideki Hayakawa. Photometric stereo under a light source with arbitrary motion. *JOSA A*, 11(11):3079–3089, 1994.

[35] Alan Yuille and Daniel Snow. Shape and albedo from multiple images using integrability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 158–164. IEEE, 1997.

[36] Neil G Alldrin, Satya P Mallick, and David J Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7. IEEE, 2007.

[37] Thoma Papadhimitri and Paolo Favaro. A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. *International Journal of Computer Vision*, 107(2):139–154, 2014.

[38] Thoma Papadhimitri and Paolo Favaro. A new perspective on uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1474–1481, 2013.

[39] Yvain Quéau, François Lauze, and Jean-Denis Durou. Solving uncalibrated photometric stereo using total variation. *Journal of Mathematical Imaging and Vision*, 52(1):87–107, 2015.

[40] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999.

[41] Jonathan T Barron and Jitendra Malik. High-frequency shape and albedo from shading using natural image statistics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2521–2528. IEEE, 2011.

[42] Lap-Fai Yu, Sai-Kit Yeung, Yu-Wing Tai, and Stephen Lin. Shading-based shape refinement of rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1415–1422, 2013.

[43] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17–24, 2013.

[44] Avishek Chatterjee and Venu Madhav Govindu. Photometric refinement of depth maps for multi-albedo objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 933–941, 2015.

[45] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 969–976. IEEE, 2011.

[46] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000.

[47] Qing Zhang, Mao Ye, Ruigang Yang, Yasuyuki Matsushita, Bennett Wilburn, and Huimin Yu. Edge-preserving photometric stereo via depth fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2479. IEEE, 2012.

[48] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 839–846. IEEE, 1998.

[49] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

[50] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[51] Wallace Casaca, Luis Gustavo Nonato, and Gabriel Taubin. Laplacian coordinates for seeded image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 384–391, 2014.

[52] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[53] Markus Unger, Thomas Pock, Manuel Werlberger, and Horst Bischof. A convex approach for variational super-resolution. In *Joint Pattern Recognition Symposium*, pages 313–322. Springer, 2010.

[54] Paolo Favaro and Thoma Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–828. IEEE, 2012.

[55] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[56] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. FlexISP: A flexible camera image processing framework. *ACM Transactions on Graphics (TOG)*, 33(6):231, 2014.

[57] Tim Meinhardt, Michael Möller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. *arXiv preprint arXiv:1704.03488*, 2017.